Imperial College London

Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions

Rachel Phillips

Thesis Submitted in Fulfilment for the Degree of Doctor of Philosophy (PhD)

School of Public Health, Faculty of Medicine, Imperial College London

FUNDED BY NIHR | National Institute for Health Research

Acknowledgements

This work would not have been possible without the unwavering support and guidance of my supervisors Dr Victoria Cornelius and Dr Odile Sauzet. A particular thanks to Victoria who believed in me from the beginning and encouraged me to pursue both a PhD and a National Institute for Health Research (NIHR) fellowship.

I would like to thank the NIHR for providing funding to pursue this PhD as part of a Doctoral Research Fellowship. In addition, to King's College London Research Design Service for providing invaluable feedback and support when I was applying for this award, with special thanks to Fiona Reid and Dr Peter Lovell for their unwavering support.

My thanks goes to a number of collaborators who have supported many aspects of this project. Including Dr Suzie Cro, Lorna Hazell, Anca Chris Ster, Dr. Daniela Junquiera, Professor Stephen Julious, Dr Tianjing Li, Dr Riaz Qureshi, Professor Catherine Hewitt, Louise Williams and the wider UKCRC CTU statistics operations group.

Finally to my amazing colleagues at the Imperial Clinical Trials Unit who offered advice, support and most importantly friendship throughout.

Dedication

To my parents without whom none of this would have been possible, I will be forever grateful for your belief in me and your constant words of wisdom - "*cool head, sharp axe, steady hand*".

Statement of Originality

The work presented in this thesis is independent research funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (Award Number: DRF-2017-10-131). The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

The author declares that the work contained within this thesis is their own original work. Support of supervisors and collaborators is recognised in the Acknowledgement section, at the beginning of each relevant chapter and cited in the reference list as appropriate.

Copyright Declaration

Relevant peer reviewed publications (first author)

Phillips R., Hazell L., Sauzet O., et al. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open 2019; 9: e024537. 2019/03/04. DOI: 10.1136/bmjopen-2018-024537.

Phillips R., Sauzet O. and Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. BMC Medical Research Methodology 2020; 20: 288. DOI: 10.1186/s12874-020-01167-9.

Phillips R. and Cornelius V. Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry. BMJ Open 2020; 10: e036875. DOI: 10.1136/bmjopen-2020-036875.

Phillips R., Cro S., Wheeler G. et al. Recommendations for visualising harms in Randomised Controlled Trial publications: a consensus. BMJ (in submission)

Relevant peer reviewed publications (co-author)

Cornelius V., Cro S. and Phillips R. Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials. Trials 2020; 21: 1028. DOI: 10.1186/s13063-020-04903-0.

Chis Ster A., Phillips R., Sauzet O., et al. Improving analysis practice of continuous adverse event outcomes in randomised controlled trials - a distributional approach. Trials 2021; 22: 419. DOI: 10.1186/s13063-021-05343-0.

Junqueira D., Phillips R., Zorzela L., et al. Commentary: Time to improve the reporting of harms in randomized controlled trials. Journal of Clinical Epidemiology 2021; 136:216-22. DOI: https://doi.org/10.1016/j.jclinepi.2021.04.020

Cornelius V. and Phillips R. Improving the analysis of adverse event data in randomised controlled. Journal of Clinical Epidemiology (under review)

Qureshi R., Chen X., Görg C., Mayo-Wilson E., Dickinson S., Golzarri-Arroyo L., Hong H., Phillips R., Cornelius V., McAdams DeMarco M., Guallar E., Li T. Comparing the value of data visualization methods for communicating harms in clinical trials. Epidemiological Reviews (under review)

Invited oral conference or meeting presentations

PRIMENT seminar, University College London. Talk title: Recommendations for visualising the drug harm profile in RCTs: a consensus. R Phillips. October 2021

MRC-NIHR Trials Methodology Research Partnership, Outcomes Working Group. Talk title: Gauging interest in progressing research in the area of adverse events in RCTs. R Phillips & V Cornelius. May 2021.

OXSTAT seminar, University of Oxford. Talk title: Data visualisations for adverse events in RCTs. R Phillips & V Cornelius. May 2021

US Agency for Healthcare Research and Quality Methods Symposium. Talk title: Harms data analysis and visualisations. R Phillips. March 2021.

French Statistical Society Conference on behalf of the Royal Statistical Society Medical Section. Talk title: Data visualisations of adverse events in randomised controlled trials, V Cornelius, S Cro & R Phillips. May 2020 - event postponed.

Medical Statistics Early Career Research workshop hosted by Plymouth University. Talk title: An evaluation and application of statistical methods designed to analyse adverse event data in RCTs, R Phillips. May 2020 - event postponed.

Unit of Medical Statistics, King's College London seminar. Talk title: An evaluation and application of statistical methods designed to analyse adverse event data in RCTs – a methodological review. R Phillips. January 2020.

BMJ editors meeting. Talk title: The use of visual analytics for clinical trial safety outcomes, R Phillips. December 2020.

Hosted the UKCRC Statistics Operation Group biannual meeting themed around adverse event analysis in RCTs. V Cornelius, S Cro & R Phillips. November 2019

Contributed oral conference or meeting presentations

NIHR Statistics group annual conference. Talk title: Recommendations for visualising the drug harm profile in Randomised Controlled Trials: a consensus. R Phillips, V Cornelius, S Cro, G Wheeler. June 2021.

Society for Clinical Trials annual conference. Talk title: Recommendations for visualising the drug harm profile in RCTs: a consensus. R Phillips. May 2021

NIHR statistics group annual conference. Talk title: The use of visual analytics for harm outcomes in pharmacological trials. R Phillips, A Chis Ster, V Cornelius. June 2020

5th International Clinical Trials Methodology Conference (ICTMC). Talk title: An evaluation and application of statistical methods designed to analyse AE data in RCTs, a methodological review. R Phillips, O Sauzet, V Cornelius. October 2019

5th International Clinical Trials Methodology Conference (ICTMC). Talk title: The use of visual analytics for clinical trial safety outcomes, a methodological review. R Phillips, O Sauzet, V Cornelius. October 2019

PSI annual conference. Talk title: Statistical methods available to analyses AEs in RCTs are not being used. Can you help us understand why? R Phillips, O Sauzet, V Cornelius. June 2019

PSI annual conference. Talk title: Lessons to learn from the reporting of adverse events (AEs) in randomised controlled trials: a systematic review of published reports in four high impact journals. R Phillips, L Hazell, O Sauzet, V Cornelius. June 2018

4th International Clinical Trials Methodology Conference (ICTMC) and the 38th Annual Meeting of the Society of Clinical Trials. Talk title: Lessons to learn from the reporting of adverse events in RCTs published in four high impact journals. R Phillips & V Cornelius. May 2017

Conference poster presentations

Royal Statistical Society conference. Poster title: Lessons to learn from the reporting of adverse events in randomised controlled trials: a systematic review of published reports in four high impact journals. R Phillips, L Hazell, O Sauzet, V Cornelius. September 2018

5th International Clinical Trials Methodology Conference (ICTMC). Poster title: Opportunities and experiences of accessing pharmaceutical individual patient data for statistical research. R Phillips, O Sauzet, V Cornelius. October 2019

Other relevant research contributions

Phillips R. and Cro S. AEFDR: Stata module to perform false discovery rate p-value adjustment for adverse event data. Statistical Software Components S458733. Boston College Department of Economics, 2020.

Phillips R. and Cro S. AEDOT: Stata module to produce dot plot for adverse event data. Statistical Software Components S458735. Boston College Department of Economics, 2020.

Phillips R. and Cro S. AEVOLCANO: Stata module to produce volcano plot for adverse event data. Statistical Software Components S458736. Boston College Department of Economics, 2020.

Phillips R., Cornelius V. and Sauzet O. An overview of statistical methods developed to analyse adverse events in clinical trials: protocol for a methodological review, https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=97442 (2018)

Abstract

Introduction

Randomised controlled trials (RCTs) provide data to help establish the harm-profile of drugs but evidence suggests that this data is underutilised and analysis practices are suboptimal.

Aims

To develop and assess methods for the analysis and presentation of harm outcomes in phase II/III drug trials that can facilitate the detection of adverse drug reactions (ADRs) and enable communication of informative harm-profiles.

Methods

A systematic review looked at current practice for collection, analysis and reporting of harm outcomes and a scoping review to identify statistical methods proposed for their analysis was undertaken. A survey of clinical trial statisticians measured awareness of methods for the analysis of harm outcomes, barriers to their use and opinions on solutions to improve practice. Alternative strategies for analysis and presentation of harm outcomes were explored.

Results

The review of current practice confirmed that data on harm outcomes is not being fully utilised, providing evidence of inappropriate and inconsistent practices. The scoping review revealed a broad range of methods for the analysis of both prespecified and emerging harms. The survey confirmed sub-optimal practices and while there was a moderate level of awareness of alternative approaches, use was limited. Guidance and training on more appropriate methods was unanimously supported. Recommendation were devised via consensus to encourage trialists to use visualisations for analysing and reporting harm outcomes. Of the evaluated methods for the analysis of emerging harms none were appropriate in trials ≤5000 participants with some utility in specific scenarios, recommendations for use are provided.

Conclusion

Clinical trial statisticians agree that there is a need to improve how we analyse and report harm outcomes in RCTs. Efforts to date have focused on prespecified harm outcomes, with little thought given to emerging harms. Several solutions for immediate adoption are proposed but there remains the need for an easy to implement, objective, signal detection approach. Guidelines for best analysis practice that are endorsed by key stakeholders would also enable a more coherent and consistent path for change.

Table of contents

1. Introduction 22

1.1 Drug development pathway 22

1.2 Terminology and definitions (table 1.2) 26

1.3 Harm outcomes in phase II/III RCTs 28

1.4 State of play for harm outcomes in RCTs 29

1.5 Outline of the scope of this thesis 31

1.6 Challenges of analysing harm outcomes in RCTs 33

1.7 Motivation 38

1.8 Aims 39

1.9 Thesis outline 40

1.10 Summary 40

2. Current analysis and reporting of emerging harm outcomes in published RCTs 41

2.1 Introduction 41

2.2 Aims 42

2.3 Methods 42

2.3.1 Eligibility criteria 42

2.3.2 Search strategy and data extraction 43

2.3.3 Data analysis 44

2.4 Results 46

2.4.1 Data extraction 46

2.4.2 Study characteristics 47

2.4.3 Collection and assessment methods for emerging harms (constructs i and ii of table 2.1 and items 8-11 appendix A2.1) 47

2.4.4 Prespecified analysis for emerging harms (construct iii of table 2.1 and items 12-14 appendix A2.1) 49

2.4.5 Reported results for emerging harms (constructs iv and v of table 2.1 and items 15-16, 18-21, 28-32 and 34 of appendix A2.1) 53

2.4.6 Analysis of emerging harm outcomes (construct vi of table 2.1 and items 17, 22-27 and 33 of table A2.1) 57

2.4.7 Results summarised by funding source 63

2.5 Discussion 71

2.5.1 Summary and implications of this review 72

2.5.2 Differences between public and industry funded studies 78

2.5.3 Limitations of this study 79

2.5.4 Changes for immediate adoption and future research 80

2.5.5 Conclusions 80

3. Statistical methods for the analysis of harm outcomes in RCTs: a scoping review 82

3.1 Introduction 82
3.2 Aims 83
3.3 Methods 84
3.3.1 Search strategy 84
3.3.2 Selection criteria 85
3.3.3 Data extraction 86
3.3.4 Analysis 86
3.4 Results 86
3.4.1 Article selection 86
3.4.2 Characteristics of articles 87
3.4.3 Taxonomy of statistical methods for the analysis of harm outcomes 89
3.4.4 Summary of methods by taxonomy 95
3.4.5 Software development for selected visualisations 114
3.5 Discussion 116
3.5.1 Summary 116
3.5.2 Recommendations for immediate adoption 120
3.5.3 Limitations 121
3.5.4 Areas to explore further within this thesis 121
3.5.5 Conclusions 122
4. Understanding current practice, identifying barriers and exploring priorities for the analysis of harm outcomes in RCTs: a survey of academic and industry statisticians 124
4.1 Introduction 124
4.2 Aims 125
4.3 Methods 125
4.3.1 Study design 125
4.3.2 Sample size 126
4.3.3 Development of content and structure 126
4.3.4 Sampling and recruitment 128
4.3.5 Ethics and consent 130
4.3.6 Analysis 130
4.4 Results 131
4.4.1 Participant flow 131
4.4.2 Participant characteristics 133
4.4.3 Results 135
4.5 Discussion 159
4.5.1 Summary of findings 159
4.5.2 How does self-reported practice compare to that reported in the literature? 160

4.5.3 Priorities for future work as highlighted by research participants 162
4.5.4 Strengths and limitations 165
4.5.5 Plans for future work 165
4.5.6 Conclusions 166
5. Recommendations for visualising harms in RCT publications: a national consensus 167
5.1 Introduction 167
5.2 Aims 168
5.3 Rationale for consensus approach 169
5.4 Methods 169
5.4.1 Study design 169
5.4.2 Sample size 170
5.4.3 Sampling and recruitment 171
5.4.4 Participant eligibility 171
5.4.5 Meeting overview 172
5.4.6 Analysis 179
5.5 Consensus results 179
5.5.1 Participant characteristics 180
5.5.2 Multiple binary outcomes 180
5.5.3 Single binary outcomes 184
5.5.4 Multiple time-to-event outcomes 186
5.5.5 Single time-to-event outcomes 187
5.5.6 Multiple continuous outcomes 190
5.5.7 Single continuous outcomes 191
5.6 Final recommendations 194
5.6.1 Multiple binary outcome 196
5.6.2 Single binary outcomes 200
5.6.3 Multiple time-to-event outcomes 202
5.6.4 Single time-to-event outcomes 203
5.6.5 Multiple continuous outcomes 209
5.6.6 Single continuous outcomes 211
5.6.7 Areas for further development 219
5.7 Discussion 219
5.7.1 Summary 219
5.7.2 Application of recommendations in practice 220
5.7.3 Adoption and endorsement of recommendations 221
5.7.4 Strengths and limitations 222
5.7.5 Future work 224

5.7.6 Conclusion 226

6. Utilising time-to-event methodology to detect signals for adverse drug reactions .. 227

6.1 Introduction 227

6.1.1 What is time-to-event analysis? 228

6.1.2 Why are time-to-event methods potentially useful in the context of analysis of harms? 235

6.1.3 Prior use of time-to-event methods to raise signals of harm in exposure only cohort studies 236

6.1.4 Time-to-event methods in the RCT setting for the analysis of harms 237

6.2 Aims 237

6.3 Development of a novel approach to detect signals for potential ADRs 238

6.4 Existing methods that could be used to detect signals for potential ADRs 242

6.5 Software for implementation of tests to detect signals for potential ADRs 251

6.6 Simulations to assess the performance of the described tests to detect signals for ADRs in RCTs 254

6.6.1 Methods to generate the datasets - data generating mechanism (DGM) 255

6.6.2 Scenarios to be investigated - simulated scenarios 256

6.6.3 Simulation procedures - computational and coding considerations 259

6.6.4 Criteria to evaluate the performance of the statistical methods across scenarios 260

6.6.5 Number of simulations 262

6.6.6 Analysis 262

7. Utilising time-to-event methodology to detect signals for ADRs: simulation results 266

7.1 Overall results across simulated trial scenarios 266

7.2 Performance of the novel Weibull methods and the modified Cox proportional hazard model 272

7.3 Developing a signal detection strategy for screening emerging harm outcomes to detect ADRs based on simulation results 277

7.3.1 Signal detection strategy specifying the size of effect to detect (figure 7.3 & table 7.4) 277

7.3.2 Signal detection strategy specifying a period of concern for detection (figure 7.4 & table 7.5) 283

7.3.3 Signal detection strategy utilising prior knowledge on background event rates (figure 7.5 & tables 7.6 and 7.7) 289

7.4 Summary and recommendations 297

7.4.1 Comparisons to existing work 302

7.4.2 Limitations and future work 305

7.4.3 Conclusions 310

8. Discussion and recommendations 311

8.1 Summary 311

8.2 Main findings relating to current practice 312
8.2.1 Summary 312
8.2.2 What could clinical trial statisticians be doing? 313
8.2.3 Feedback from clinical trial statisticians 313
8.2.4 Differences in analysis practice between academia and industry 314
8.3 Recommendations 315
8.3.1 Changes for immediate adoption 315
8.3.2 Incorporation of visualisations 320
8.3.3 Signal detection methods 321
8.4 How does this work compare with recent research in the field? 322
8.5 Limitations of the work undertaken in this thesis 324
8.6 Limitation of RCTs for the analysis of harm outcomes 326
8.7 Strengths of this thesis 326
8.8 Ongoing collaborative projects 328
8.9 Future research 328
8.10 Overall conclusions 329
Glossary 332
References 336
Appendices 352

List of Figures

Figure 1.1 Drug development pathway 24
Figure 1.2: Example of the five levels of the MedDRA hierarchy [69] 37
Figure 3.1: Flow diagram describing the assessment of sources of evidence 88
Figure 3.2: Taxonomy of methods for the analysis of harm outcomes 90
Figure 3.3: Volcano and dot plot for emerging harms experienced by at least three participants in either treatment group from summary results presented in Whone et al.[178] 115
Figure 4.1: Flow diagram of participation in the online survey 132
Figure 4.2: Participant characteristics by employment sector and overall 134
Figure 4.3: Visual summary of analysis practices of survey respondents by employment sector 137
Figure 4.4: Influences on analysis performed on emerging harm outcomes by employment sector (always and often categories combined) 144
Figure 4.5: Barriers when analysing emerging harm outcomes by employment sector (strongly agree and always agree responses combined) 145
Figure 4.6: Opinions about analysis of emerging harm outcomes by employment sector (agreed and strongly agreed categories combined) 146
Figure 4.7: Concerns about current analysis practice for emerging harm outcomes by employment sector (moderately to extremely concerned categories combined) 150
Figure 4.8: Solutions to support a change in analysis practices for emerging harm outcomes by employment sector (strongly agree and agree categories combined) 151
Figure 5.1: Thumbnails of considered plots for multiple binary outcomes in order of preference 182
Figure 5.2: Summaries of overall scores and rankings for multiple binary outcome plots .. 183
Figure 5.3: Thumbnails of considered plots for single binary outcomes 185
Figure 5.4: Thumbnails of considered plots for multiple time-to-event outcomes in order of preference 186
Figure 5.5: Summaries of overall scores and rankings for multiple time-to-event outcome plots 187
Figure 5.6: Thumbnails of considered plots for single time-to-event outcomes in order of preference 188
Figure 5.7: Summaries of overall scores and rankings for single time-to-event outcome plots 189
Figure 5.8: Thumbnails of considered plots for multiple continuous outcomes in order of preference 190
Figure 5.9: Thumbnails of considered plots for single continuous outcomes in order of preference 192
Figure 5.10: Summaries of overall scores and rankings for single continuous outcomes .. 193
Figure 5.11: Dot plot of events - data taken from the two-arm example dataset with 1:1 allocation ratio 197
Figure 5.12: Horizontal stacked bar chart of events by maximum severity – data taken from the two-arm example dataset with 1:1 allocation ratio 199
Figure 5.13: Bar chart of event counts – data taken from the two-arm example dataset with 1:1 allocation ratio 201
Figure 5.14: Bar chart of event counts – data taken from the three-arm Mepolizumab dataset with 1:1 allocation ratio 201
Figure 5.15: Kaplan-Meier plot for an event of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio 204
Figure 5.16: Mean cumulative function plot for all events – data taken from the two-arm Paroxetine dataset with 1:1 allocation ratio 206
Figure 5.17: Event-free ratio plot for an event of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio 208
Figure 5.18: Scatterplot matrix for continuous harm outcomes – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio 210

Figure 5.19: Line graph of a summary statistic over time for a continuous harm outcome of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio..... 212
Figure 5.20: Violin plot summarising the distribution of a continuous harm outcome of interest over time - data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio..... 214
Figure 5.21: Kernel density plot for a continuous harm outcome of interest - data taken from the two-arm Paroxetine study with 2:1 allocation ratio..... 216
Figure 5.22: Decision tree to help researchers decide which plot(s) to use to visualise data on harm outcomes 218
Figure 6.1: Kaplan-Meier plots at month 1, 3, 6 and 11 (relative to a 12 month trial) for the scenario where n=2000, AE background rate = 10% and the ADR increase= 100% 259
Figure 7.1: Power of each test by sample size - summarised over varying AE background rates, time of increase and increases in background rate due to ADRs..... 271
Figure 7.2: False positive rate for each test by sample size – summarised over varying AE background rates, time of increase and increases in background rate due to ADRs 271
Figure 7.3: Power of each test by sample size and percentage increase in background event rate due to ADRs - summarised over background event rates and time of increase..... 282
Figure 7.4: Power of each test by sample size and time of increase for ADR - summarised over varying AE background rates and increases in background rate due to ADRs 288
Figure 7.5: Power of each test by sample size and AE background rates - summarised over varying times of increase and increased background rates due to ADRs 294
Figure 7.6: Design considerations and recommendations when developing a signal detection strategy for the analysis of emerging harm outcomes 301

List of Tables

Table 1.1: Clinical trial phases in the drug development pathway and the advantages and limitations of each[10, 21, 24, 26, 27] 25
Table 1.2: Key terms and definitions used throughout this thesis 28
Table 2.1: Constructs for extraction with rationale 45
Table 2.2: Characteristics of included studies 48
Table 2.3: Examples of good reporting practice in reviewed articles 49
Table 2.4: Collection, assessment and analysis methods reported in included articles 51
Table 2.5: Prespecified stopping criteria for harm 52
Table 2.6: Reporting practices across included articles 54
Table 2.7: Selection criteria categories for choice of events included in articles 55
Table 2.8: Example of multi-faceted rules used to select events to report in journal articles 56
Table 2.9: Summary of results presented and analysis undertaken 58
Table 2.10: Detailed population summaries used for analysis 60
Table 2.11: Collection, assessment and analysis methods reported in included articles by funding source 65
Table 2.12: Reporting practices across included articles by funding source 67
Table 2.13: Summary of analysis practices by funding source 68
Table 2.14: Key components to include to improve the reporting of emerging harm outcomes in clinical trial publications 81
Table 3.1: Taxonomy of methods for the analysis of harm outcomes 89
Table 3.2: Summary level classifications of identified articles and methods 91
Table 3.3: Detailed article classifications 92
Table 3.4: Summary of visual approaches to summarise harm outcomes in phase II/III RCTs 97
Table 3.5: Recommendations for analysis of harm outcomes 120
Table 4.1: Participant characteristics by employment sector and overall 135
Table 4.2: Information on emerging harms typically presented by employment sector and overall 138
Table 4.3: Methods participants mentioned they were aware of specifically for the analysis of harm outcomes 139
Table 4.4: Participants' use of specialist methods for analysis of emerging harm outcomes 140
Table 4.5: Reasons specialist methods are not used (by participants who were aware of such methods) 141
Table 4.6: Classification of participants' comments on the reasons for a lack of use of specialist methods for the analysis of emerging harm outcomes 142
Table 4.7: Influences the analysis performed by employment sector and overall 147
Table 4.8: Barriers when analysing emerging harm outcomes by employment sector and overall 148
Table 4.9: Opinions regarding analysis of emerging harms by employment sector and overall 149
Table 4.10: Concerns regarding the analysis of emerging harm outcomes by employment sector and overall 152
Table 4.11: Solutions to support a change in the analysis of emerging harms by employment sector and overall 153
Table 4.12: Classification of participants' comments on solutions to support change in analysis practices for emerging harm outcomes 154
Table 4.13: Classification of participants' general comments raised regarding analysis practices for emerging harm outcomes 157
Table 5.1: Draft framework for assessing the properties of graphical displays 174
Table 5.2: Framework for assessing the properties of graphical displays 175
Table 5.3: Visualisations for summarising the harm profile 217
Table 5.4: Visualisations to summarise individual event(s) of interest* 217

Table 6.1: Examples of common ADRs and time at which they typically occur relative to exposure 235
Table 6.2: Summaries of the methods considered as candidate signal detection tests to identify time-dependent treatment effects indicative of ADRs 252
Table 6.3: Scenarios to be simulated and data generating mechanisms (DGM) used to create them 256
Table 6.4: Study characteristics of simulated scenarios 258
Table 6.5: Summary of simulated scenarios where signals truly exist (n = 270) 264
Table 6.6: Summary of simulated scenarios without signals (n= 18) 265
Table 7.1: Power of each test to detect a signal for an ADR by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial) 269
Table 7.2: False positive rate for each test by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial) 270
Table 7.3: Sample sizes required for the proposed novel signal detection tests to achieve ≥ 80% power by simulated scenarios 276
Table 7.4: Power of each test by sample size and increases in background event rates due to ADRs of 25%, 50% & 100% over background rates of 1%, 5% & 10% at day 1, month 1, 3, 6 & 11 280
Table 7.5: Power of each test by sample size and time of increase over: background rates of 1%, 5% & 10% & increases in background rates due to ADRs of 25%, 50% & 100% 286
Table 7.6: Power of each test by sample size & AE background rates over: increases in background rates due to ADRs of 25%, 50% & 100%, at day 1, month 1, 3, 6 & 11 292
Table 7.7: False positives of each test by sample size & AE background rates over: increases in background rates due to ADRs of 25%, 50% & 100%, at day 1, month 1, 3, 6 & 11 295
Table 7.8: Recommendations when undertaking signal detection analysis on emerging harms to detect time dependent ADRs 300

List of abbreviations

AE	adverse event
ADR	adverse drug reaction
ANCOVA	analysis of covariance
ASA	American Statistical Association
BMJ	British Medical Journal
CONSORT	Consolidated Standards of Reporting Trials
CIOMS	Council for International Organizations of Medical Sciences
CRO	clinical research organisation
CTCAE	Common Terminology Criteria for Adverse Events
CTIMP	Clinical Trials of Investigational Medicinal Product
CTU	clinical trial unit
DGM	data generating mechanism
DMC	Data Monitoring Committee
EMA	European Medicines Agency
FDA	Food and Drug Administration
FDR	false discovery rate
GSK	GlaxoSmithKline
HRA	Health Research Authority
ICH	The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICTMC	International Clinical Trials Methodology Conference
IF	impact factor
IQR	inter-quartile range
IRR	incidence rate ratio
ITT	intention-to-treat
JAMA	Journal of the American Medical Association
JRCO	Joint Research Compliance Office
MCF	mean cumulative function
MedDRA	Medical Dictionary for Regulatory Activities
max	maximum
MCID	minimum clinically important difference
MHRA	Medicines and Healthcare products Regulatory Agency
min	minimum
MRC	Medical Research Council
mVWLR	modified versatile weighted log-rank test
NEJM	New England Journal of Medicine
NIHR	National Institute for Health Research
OR	odds ratio
PPI	patient and public involvement
PSI	Statisticians in the Pharmaceutical Industry
RCT	randomised controlled trial
RMST	restricted mean survival time
RR	risk ratio
SAE	serious adverse event
SD	standard deviation
SmPC	summary of product characteristics
SPERT	Safety Planning, Evaluation and Reporting Team
TMRP	Trials Methodology Research Partnership
UKCRC	United Kingdom Clinical Research Collaboration

WC	weighted combined
WHO	World Health Organisation
WHO-ART	World Health Organisation Adverse Reaction Terminology

1. Introduction

1.1 Drug development pathway

Development of a novel drug starts with preclinical studies in the laboratory and in animal models. The first human testing takes place in phase I studies. These aim to identify a dose range that has acceptable toxicity, and help to understand the properties of the drug such as how it metabolises in the human body i.e. the pharmacokinetics. Phase I studies are typically single arm and undertaken in a small number of healthy volunteers (approximately 20-50).[1, 2] They are designed to ascertain a prescribing dose typically based on the 'maximum tolerated dose' with some also investigating doses based on efficacy outcomes and help identify acute and immediate harms.[3-5] The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use - Statistical Principles for Clinical Trials guidelines (ICH E9) remarks that early phase evaluations are "*only sensitive to frank expressions of toxicity*".[6] Drugs that progress from phase I continue into phase II and III trials that focus on establishing the efficacy and harm-profile of the drug. Phase II/III trials offer the opportunity to understand the wider profile of potential harm as larger sample sizes enable a more comprehensive characterisation, and they include a control group that provides opportunity to establish a causal relationship. As described in ICH E9 "*later phase controlled trials represent an important means of exploring in an unbiased manner any new potential adverse effects, even if such trials generally lack power in this respect*".[6] All trials regardless of phase, provide high quality, prospectively collected data and are considered by some to "*provide the most interpretable evaluation of safety*".[7] Once effective treatments are approved they continue to undergo scrutiny in post-marketing research, which include phase IV clinical trials and post-marketing surveillance studies, which may include monitoring through spontaneous reporting systems such as the United Kingdom's Yellow Card Scheme and Europe's EudraVigilance system.[8-10] Post-marketing studies can include both randomised controlled trials (RCTs) and observational studies. They aim to evaluate the real-world effectiveness of interventions, further defining the harm profile of a drug in more

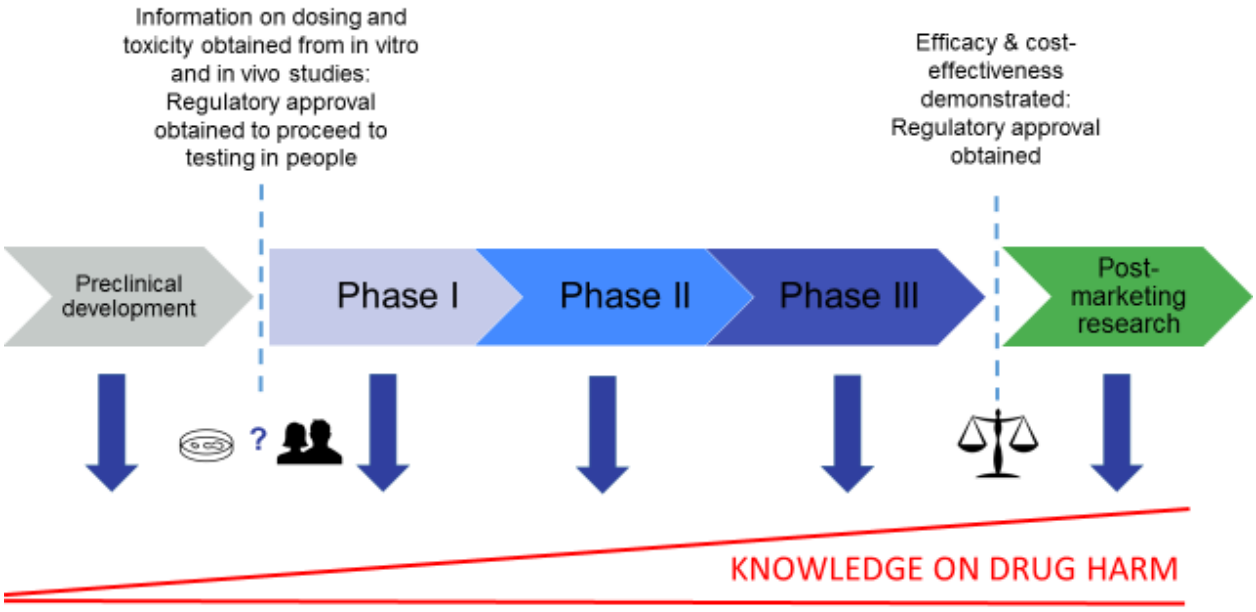
heterogeneous samples and information collected in this phase can help identify rarer harms, harms with long latency and any adverse impacts of long-term use.[11, 12] Throughout the research pathway, information on the harm profile of the drug is gathered and this continues beyond the research arena once drugs are approved and are prescribed in practice (figure 1.1).[13] Each stage makes its own contribution to identifying harm and no one stage is sufficient alone but cumulative evidence gathered in each allows establishment of the harm profile (table 1.1).

Early phase studies are pivotal in establishing dose ranges with acceptable toxicity to be recommended for further testing in later phase studies. Therefore, such studies are carefully designed and analysed with this in mind, and much progress has been made toward implementation of sophisticated model based designs over more simplistic algorithmic approaches.[5, 14, 15] Once the drug is approved for use and marketed there are well-accepted signal detection statistics used in the post-marketing setting to identify suspected adverse drug reactions (ADRs). These include the proportional reporting ratio, reporting odds ratio, information component and empirical Bayes geometric mean.[16] While industry have resources to undertake extensive pharmacovigilance* and signal detection activities, independent centres such as the Uppsala Monitoring Centre and the Drug Safety Research Unit have been established to undertake independent pharmacovigilance activities, and regulatory bodies such as the Medicines and Healthcare products Regulatory Agency (MHRA) and the Food and Drug Administration (FDA) require evidence submissions for review and approval before new products are released to market.[16, 18, 19] Despite the careful monitoring and sophisticated statistical methods used, limitations remain and identification of harm can take many years after approvals.[20, 21] There is a question as to whether the harm identified in post-marketing settings could

* "*Pharmacovigilance is the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem*." 17. European Medicines Agency. Pharmacovigilance: Overview, https://www.ema.europa.eu/en/human-regulatory/overview/pharmacovigilance-overview (accessed 28/10/2020).

have been identified earlier and whether the information accumulated in trials can play a role in earlier identification. As Stephen Senn describes in Statistical Issues in Drug Development, "*If there are problems with a drug, then the sooner they are discovered the better*".[22] Information from the phase II/III setting could provide additional valuable information regarding potential drug-event relationships feeding into post-marketing monitoring. However, there is evidence that the analysis practices for harms in the intermediate, phase II/III stage have been neglected with no "*universally acceptable gold standard*" for analysis, resulting in this high quality, prospectively collected data that allows for a causal assessment being underutilised.[23-25]

Figure 1.1 Drug development pathway



Summary of the drug development pathway showing how information on harm is collected across phases allowing the harm profile to emerge.

Table 1.1: Clinical trial phases in the drug development pathway and the advantages and limitations of each[10, 21, 24, 26, 27]

Phase	Description	(Typical) Size and design	Advantages to assess harms	Limitations to assess harms
I	First in-human testing, typically undertaken in healthy volunteers. Aim to determine dose range with acceptable toxicity and gain an understanding of drug properties in the human body	20-50, single arm studies	Identifies dose range with acceptable toxicity Identifies acute and immediate harms Model based approaches for analysis becoming more established and easier to implement Prospectively collected data Closely monitored Well specified indication and measured drug use	No comparison group Small samples
II	To establish preliminary efficacy and short-term harms	30-300, controlled studies	Comparator group allows an assessment of causality Systematic collection High quality, detailed data Prospectively collected data Well specified indication and measured drug use Design minimises potential for confounding and bias	Exclusions can lead to a lack of generalisability Limited follow-up can miss events with long latency A lot of complex data collected Analysis under a traditional hypothesis-testing framework will typically be under powered
III	To establish or confirm definitive efficacy and/or effectiveness plus additional information on the harm profile	300-5000, controlled studies		
IV	Post-marketing studies in normal clinical use to optimise use and further establish the harm profile and identify rare events	> 300, observational studies and surveillance databases	Large samples can identify rare and/or unexpected events Inclusive/heterogeneous samples can identify harms in new populations not previously exposed and drug-drug interactions leading to harm Can identify events with long latency and impacts of long-term use Established sophisticated statistical methods for analysis	Often no control group Denominator typically unknown Suffer from under-reporting and selective reporting Sometimes data are retrospectively collected which can suffer from a variety of biases Difficult to measure accurate drug use Large potential for confounding

1.2 Terminology and definitions (table 1.2)

In clinical trials, the terms used to refer to harm outcomes are numerous and provide insight into one of the many challenges faced by trialists and key stakeholders (such as prescribers, researchers, patients and regulators) when collecting, analysing, reporting and interpreting such data. Common terms used include safety, adverse events, toxicity, risk, and harms. The Consolidated Standards of Reporting Trials (CONSORT) statement extension to harm outcomes aimed to promote use of standard terminology and encouraged authors to use the term harm instead of safety, which they felt could be misleadingly reassuring.[28] Recent discussions amongst the CONSORT harm working-group has reaffirmed their stance on this topic. Whilst discussions touched upon the emergence of the use of 'risks' and the advantages this would offer by aligning it with the risk-benefit literature, it was deemed to be inappropriate due to its statistical interpretation. Consistency in terminology is important as it provides clarity to audiences and allows concepts to be more readily reinforced, easing education of trialists and helping them to understand where guides, methods etc. are relevant, it also simplifies systematic searches of the literature that aim to synthesise existing research on a common area. Unfortunately, use of the term harm is not universally adopted across the trials community, with industry and regulators continuing to refer to safety outcomes. Whilst the debate continues, given the academic viewpoint from which this thesis is presented and in line with CONSORT, I will use the term harm, referring to harm outcomes when referring to individual events and to the harm profile when referring to the summary or burden of the cumulative effect of all harm outcomes. Use of the term 'harms' reflects the aim to establish any harmful effects of interventions and not to establish that they are safe.[28] Trials that aim to establish that a drug is safe are defined in this thesis as those looking for the absence of harm, such trials are not the focus of this thesis, and where necessary will be referred to as safety studies. Use of the term adverse event (AE) is used interchangeably in the literature to refer to harm outcomes but reference to AEs in this thesis will be a subset of harm outcomes collected in clinical trials that contribute to the harm profile. AEs are defined as per the World Health Organisation (WHO) definition as "*any untoward*

medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment".[29, 30] The closely related term, adverse (drug) reaction (A(D)R) is used to indicate that a causal relationship between the intervention and event is "*at least a reasonable possibility*". Adverse reactions are a subgroup of adverse events. The term 'signal' will be used to refer to the information that raises or 'flags' the possibility of this causal relationship.[30] Signals can be used to indicate that closer examination of an outcome is needed; this might involve closer examination of the event in ongoing studies or inform outcomes to prespecify in future studies including subsequent RCTs, systematic reviews or post-marketing research.

Harm outcomes can include prespecified events listed in advance as outcomes of interest to follow-up. These may be events that are already known or suspected to be associated to the intervention, or events that are followed-up for reasons of interest. In addition, information on non-prespecified harms will also be reported and collected during a trial and these will be referred to as emerging events. Key terms used throughout this thesis are summarised in table 1.2 and a full glossary of terms is provided in the supplementary material.

Table 1.2: Key terms and definitions used throughout this thesis

Term	Definition
Adverse drug reaction (ADR)	Harm outcomes where a causal relationship between the intervention and event is "*at least a reasonable possibility*".[30]
Adverse event (AE)	Subset of harm outcomes that includes "*any untoward medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment*".[29, 30]
Emerging	Non-prespecified events that are reported and collected during the trial and may be unexpected. Includes AEs, and laboratory and vital sign data indicative of harm.
Harm outcomes	Individual events encompassing emerging events and prespecified events of interest.
Harm profile	The summary or burden of the cumulative effect of all harm outcomes.
Pharmacovigilance	The science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem
Prespecified	Individual events that are listed in advance as harm outcomes of interest to follow. They may be known or suspected to be associated to the intervention, or followed for reasons of interest.
Safety studies	Trials that aim to establish the absence of harm.
Signal	Information that raises the possibility of a causal relationship between the drug and event.

1.3 Harm outcomes in phase II/III RCTs

Phase II/III trials usually take the form of RCTs and are typically designed to address questions of efficacy (does it work under ideal circumstances) or effectiveness (does it work under 'real-world' circumstances) of either new drug interventions or existing interventions for a new indication (a process known as repurposing). The primary research question is most often to determine the efficacy or effectiveness of an intervention, which typically include questions about benefits; they are rarely designed to determine whether a treatment is harmful with such outcomes typically considered as either pre-specified secondary outcomes or non-specific emerging events (definitions provided in table 1.2).[12, 31] As such "*much of the statistical theory of planning clinical trials has to do with investigating efficacy rather than safety*".[22]

As illustrated in figure 1.1, phase II/III RCTs are just one stage in the clinical evidence pathway. Once a drug reaches this stage there will already exist some knowledge about the harm profile. Studies in

the design stage can use this existing knowledge to prespecify harm outcomes of interest for monitoring and analysis. However, there are still many unknowns and additional data on emerging events are captured via unsolicited and spontaneous reports of AEs, laboratory data (including blood tests and culture data), and vital signs and other physical findings (e.g. pulse rate, temperature, blood pressure and electrocardiograms) to help identify signals for potential ADRs and build a more informative and comprehensive harm profile.[29, 30] Crowe et al. also recognised this distinction in event specification, categorising events into tiers, where prespecified events are classed as tier one events and events without a prespecified hypothesis are classed as either tier two or tier three dependent on frequency (common events are referred to as tier two and rare events equate to tier three).[7] However, use of this classification system seems to be limited and therefore I will use the more intuitive prespecified and emerging terminology throughout this thesis.

1.4 State of play for harm outcomes in RCTs

RCTs are considered the 'gold standard' in the evidence pathway and as such, good practice methods to evaluate, analyse and report efficacy (used here and onwards to be synonymous with effectiveness unless explicitly stated otherwise) outcomes within RCTs are well established. These include advanced analysis techniques that can account for, for example, missing data or post-randomisation effects, which have been developed and adopted to meet the rapidly advancing field of clinical trials.[32, 33]

In addition, the last 15 years has seen increasing emphasis on developing harm profiles of drugs, especially from the pharmaceutical sector. Working groups have developed guidance on the reporting of harm data from RCTs in journal articles. This includes: the harms extension to the CONSORT statement, which provided a 10-point checklist of items to include when reporting harms

in RCTs; the joint pharmaceutical/journal editor collaboration guidance on reporting of harm outcomes in journal articles, which was proposed to complement the CONSORT harms checklist to improve reporting of harm; and the extension of PRISMA for harm outcomes reported in systematic reviews, which proposed an additional four items to the original statement to improve harms reporting in reviews.[28, 34, 35] Regulators including the European Commission and the FDA and other bodies such as the ICH (a joint regulator and industry initiative that aims to harmonise drug development) and the Council for International Organizations of Medical Sciences (CIOMS) (an international, non-governmental, non-profit organization whose aim is to advance public health through guidance), have also issued detailed guidance on the collection and presentation of harms related data arising in clinical trials.[36-40] A recent initiative from the American Statistical Association (ASA) Biopharmaceutical Section safety working group set out their intentions in coming years to "*contribute ideas for process, tools, methods and applications for the evaluation of drug safety*", indicating a move towards addressing analysis.[41] An initial paper from this group set out the challenges researchers face when analysing harm outcomes and suggest visualisations are "*key to effective communication*", proposing some possible graphics for harm outcomes and highlighted the importance of utilising information on time, both themes that will be explored further in this thesis.[42] In addition, the pharmaceutical industry standard from the Safety Planning, Evaluation and Reporting Team (SPERT), provided recommendations based on a program safety analysis plan to be implemented across the lifecycle of a drug development program. This is based on standardised collection and recommended analytical approaches proposed under the three-tier classification system for events proposed by Crowe et al.[5] They too highlighted graphics as a potentially useful tool for analysis of harm outcomes and the importance of incorporating information on time. They recommended more advanced analytical approaches for prespecified events (or as they refer to them, tier one events) but recommendations for the analysis of emerging events were limited to providing point estimates such as risk differences or odds ratios with confidence interval and/or p-values. This in line with the 2011 book, Stephens' Detection and Evaluation of Adverse Drug

Reactions, which dedicated a chapter to methods for the analysis and presentation of data on harm outcomes obtained in clinical trials.[43] The authors propose using point estimates such as the risk difference, risk ratio and odds ratio to compare frequencies between treatment groups with corresponding confidence intervals and suggests using the chi-squared or Fishers' exact test to test for statistically significant differences. In addition, they suggest plotting Kaplan-Meier plots to incorporate time-to-event data and testing for differences with the log-rank test, Cox models or parametric regression models such as a Poisson regression model.

Whilst recommendations and guidelines call for better practice in collection and reporting, they are limited in their recommendations for statistical analysis practices, focusing on prespecified events, neglecting the analysis of emerging events. The progress seen for the analysis of efficacy outcomes has not been matched for the analysis of harm outcomes in published reports of RCTs.[41, 44] To date, the SPERT working group has given this the most consideration, suggesting a move toward better approaches for prespecified events but analysis of emerging events seems to remain a neglected area for review and development and there is a clear gap in guidance for analysis of emerging harm outcomes.[7, 45] How true this statement is and the reasons why this might be will be explored in this thesis.

1.5 Outline of the scope of this thesis

This thesis will examine analysis practice in trials of pharmacological interventions. This is one area of interventional trials and fits within the class of clinical trials of investigational medicinal products (CTIMPs) as defined by the MHRA in the Medicines for Human Use (Clinical Trials) Regulations 2004 (SI 1031).[46] There are other trial types that examine effects of psychological, behavioural, surgical, lifestyle or educational interventions. These interventions are often comprised of multiple

interacting components and are referred to as complex interventions but can also be simple interventions, and are classed as non-CTIMPs.[47] While there are many similarities in terms of design and analysis between CTIMPs and non-CTIMPs, non-CTIMPs, especially those of complex interventions present their own unique challenges for the analysis of harm outcomes. It has been argued that non-CTIMPs, specifically trials of psychological, behavioural and lifestyle interventions require their own guidelines on how harm outcomes should be identified, monitored and reported.[48, 49] At present, non-CTIMPs require less stringent reporting of harmful effects to the Health Research Authority (HRA) which only require reports of related and unexpected serious adverse events.[50] The HRA is the central body in the UK that is responsible for the regulation and approval of different aspects of health and social care research. Hence, such trials are not examined in this piece of research but there will be elements of this thesis that are applicable for all clinical trials.

The emerging data during phase II/III trials will typically undergo periodic review by a data monitoring committee (DMC) comprised of independent experts who review accumulating data to assess study progress (e.g. recruitment rates), trial conduct (e.g. protocol deviations), harm outcomes and potentially important efficacy outcomes in interim analyses.[51] In recent years, there has been some effort toward improving the reporting of harm outcomes in DMC reports including published template reports and recommendations on graphical displays to account for differential follow-up at interim analysis.[52-54] Whilst I include some discussion on the methods proposed for the monitoring of harm outcomes in ongoing studies, the focus of this thesis will be on the analysis and reporting of harm outcomes for the final analysis reported in publications.

1.6 Challenges of analysing harm outcomes in RCTs

1.6.1 Numerous and undefined harm outcomes

Analysing harm outcomes in RCTs presents several challenges, which could help explain a lack of progress in analysis practices.[24, 42] Unlike efficacy outcomes which are well defined and restricted in number at the planning stage of a RCT, numerous, undefined harms are collected in RCTs. We can define a number of prespecified harm outcomes as secondary outcomes but many true harms are often unknown and/or unexpected at this stage and hence undefined.[6, 43] Thus the range of possible events is large. Furthermore, there are a mix of different outcome types e.g. binary, count, time-to-event and continuous, and collection requires additional information to be obtained on factors such as seriousness which measures the "*extent to which the reaction can or does cause harm*" and severity which measures "*the extent to which the reaction develops*", which are both important but distinct concepts i.e. a harm may be severe but not serious.[55] Plus we are interested in the timing and duration, number of occurrences, and outcome, which for efficacy outcomes would have all been predefined.[45] In addition to all of this, events often occur at very low rates and therefore true ADRs can be difficult to detect.

Careful consideration is needed on how best to communicate and present vast amounts of complex information on harm, ensuring it is fairly balanced with the evidence on efficacy. All treatments come with some risk of harm, what is an acceptable level varies between diseases and patients, and how this is balanced with benefits is not straightforward.[43] It also needs to be presented in a manner not to overwhelm readers and to ensure important data are not inadvertently omitted, which could lead to missed signals. Alternative solutions to present this data will be explored in this thesis. In addition, there is evidence that the methods by which to select events to present in journal articles is lacking and may need more careful consideration.[56, 57] Whilst this will not be directly addressed in this thesis, analysis approaches explored might help to inform more objective selection strategies in

future work. It is also important to consider the limited space available in journal articles. Previous research suggests that results on harm outcomes are not given sufficient priority, with one study reporting, "*over two-thirds of published articles were found to dedicate more space to authors' affiliations than descriptions of harm*".[58, 59] So the perceived importance (or lack) of including information on harms in journal articles may also need to be addressed.

1.6.2 Statistical considerations

From a statistical perspective, under a traditional hypothesis-testing framework consideration to type-I (false-positive) errors and type-II (false-negative) errors is crucial, especially when considering how to analyse emerging events i.e. those that have not been prespecified. RCTs are typically designed to test the efficacy of an intervention and are not powered to detect differences in harm outcomes such as detecting differences in proportions of events, which could be indicative of an ADR.[6] Added to this is that relatively small effect sizes are often important and of interest. As a trial is not powered to detect ADRs, there is a possibility that any statistical testing under a hypothesis-testing framework may result in the drug being deemed safe or a trial not being stopped early enough, resulting in more patients than necessary suffering an ADR. In addition, the vast number of potential emerging events can lead to issues of multiplicity.[44, 60] For example, it would not be unheard of for the number of emerging harms to exceed the number of participants in a clinical trial and a large number of statistical comparisons could lead to a false signal of an ADR i.e. a chance imbalance, resulting in the drug being deemed unsafe or a trial being incorrectly halted early. That said any adjustment for multiplicity is likely to make a "*finding untenable*" and therefore the value of adopting traditional sequential monitoring methods used for efficacy outcomes might be limited for monitoring harms.[61] However, researchers have highlighted that there is still a "*need for inferential statistics"* in this setting and the lack of a prespecified hypothesis does not negate the need for formal analysis.[22, 25] Consideration to alternatives to the traditional hypothesis-testing framework,

using evidence to identify events for closer inspection rather than drawing definitive conclusions have been called for and will be discussed in this thesis.[25]

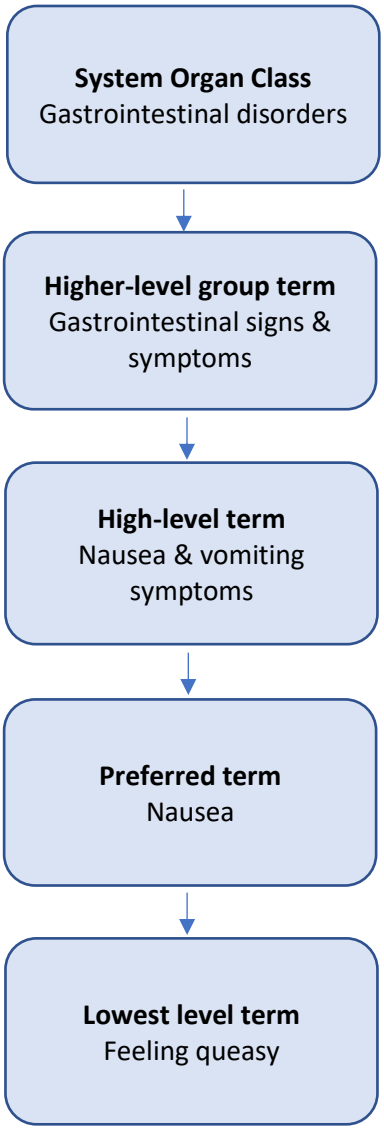
It is also important to consider the impact of differential follow-up and/or exposure times, the time events occur and dependencies between events, and analysis should account for these factors where necessary.[62] For example, if a disproportionate number of participants withdraw from the control group due to a lack of efficacy it may appear that there are more events in the intervention group simply through a greater opportunity for participants receiving the intervention to experience and report such events. Failing to account for such differential follow-up will lead to biased results. Alternatively if intervention participants are withdrawing due to intolerability (i.e. harm), limiting their follow-up and opportunity to experience recurrent or more severe events it could make treatments looker less harmful than they are.[54] Recent attention has been given to this under the emerging estimand framework. Unkel et al. discussed statistical methods suitable to analyse harm outcomes in the presence of varying follow-up times.[62] They also sought to address the contested topic of appropriate analysis populations for analysis of harm outcomes. CONSORT harms recommend that analysis of harm data should be performed on the intention-to-treat (ITT) population to maintain the random assignment.[28] However, both the CIOMS working group VI and the SPERT initiative consider a more appropriate population for analysis of harms to be those that receive at least some quantity of the intervention.[7, 40] Unkel et al. reframe this into the estimand framework and encourage researchers to instead think about their research question and what they wish to estimate when selecting the analysis population for a specific harm outcome of interest. There is no consideration given to emerging events i.e. those that are not prespecified, which highlights an often unmet consideration in the analysis of harm outcomes – the unspecified research question being asked when analysing emerging events. Methods exploring differential follow-up for emerging events will be investigated in this thesis.

1.6.3 Collection of data on harm outcomes in RCTs

Emerging harm outcomes can include reports of AEs that are collected either via unprompted participant reports, which is referred to as spontaneous collection, or via prespecified checklists or asking non-leading questions, which is referred to as active collection of data. Harms are also collected from routine surveillance procedures such as clinical and biological tests that participants undergo at regular trial visits. These procedures to collect harms and the coding practices to standardised events can also produce challenges. Whilst such issues are not the focus of this thesis, it is important to remain mindful of the implications of different practices undertaken across trials when making conclusions about the harm profile of an intervention and these will be discussed briefly in this thesis. For example, medical dictionaries can be used to standardise events bringing order to the body of emerging events, the most commonly used being the Medical Dictionary for Regulatory Activities (MedDRA).[63] Medical dictionaries code events according to hierarchies, within MedDRA, at the lower end, events are classed into preferred terms, which describe 'unique medical concepts' (below this are lower level terms which might consist of alternative spellings for the preferred term or even more specific descriptions of the preferred term) such as 'nausea'. Each preferred term is classified into one unique system organ class or body system (via high level terms and high level group terms) such as 'gastrointestinal disorders' (figure 1.2 gives a detailed example).[64, 65] Alternative dictionaries are available such as the Common Terminology Criteria for Adverse Events (CTCAE) or the World Health Organisation Adverse Reaction Terminology (WHO-ART), however since version 4.0 of the CTCAE, released in 2009, all terms are MedDRA terms and since 2015 the WHO-ART has no longer been maintained with trialists and pharmacovigilance teams directed to MedDRA.[66, 67] It is believed that going forward use of a single universally accepted dictionary will help improve standardisation across trials, however it is still important to bear in mind that the level of coding used within a dictionary for analysis and reporting is likely to have

consequences on signals detected. For example, if terms used to classify events are too broad then signals at the preferred term level may be masked since it relies on the assumption that drugs act on the 'system' level and potentially misses important events that occur in isolation. Paradoxically if classification is too specific, splitting events into too many sub-categories, then potential signals for an adverse reaction may also be missed because of low numbers of events within sub-categories.[63] In addition, methods that utilize such coding systems rely on events being correctly and consistently classified.[68]

Figure 1.2: Example of the five levels of the MedDRA hierarchy [69]



1.7 Motivation

Whilst guidance has been proposed to improve the reporting of harm outcomes, recommendations for analysis are limited and there is a suggestion that analysis practices are frequently inadequate and what is reported is inconsistent.[24, 41] Based on a review of the oncological literature, Drago et al. concluded that opportunities to "*quantify AE profiles"* in phase III trials are being missed and improvements in analysis are needed to allow them to "*better fulfil their potential*".[25] Crowe et al. highlighted the value of harm profiles from completed clinical trials, stating that they could be used to identify events of interest which could be promoted to prespecified events of interest for monitoring in future trials (where such events have a prespecified hypothesis to be tested).[7] This information could also be used to inform post-marketing monitoring activities, where there is evidence of considerable delays between a drug being released to market, evidence of adverse reactions emerging and subsequent necessary drug withdrawals.[20] In addition, in many trial scenarios there is a potential for large amounts of data to be collected and presenting everything has the potential to overwhelm audiences and makes interpretation difficult, whilst omitting information could result in missing important signals of harm. It is not clear if trialists know how to analyse and present more informative summaries of harm outcomes. Ensuring this robust, controlled data on harms is fully utilised, could enable a more informative harm-profile to be presented in a more efficient manner; and could help detect signals of harm for future monitoring in subsequent clinical trials or post-marketing studies, which could potentially lead to improved patient care.

Another important aspect is when regulators such as the European Medicines Agency (EMA) or FDA assess the benefit-risk profile to make drug-licensing decisions. These are complex decisions, as many different aspects need to be taken into account. Factors taken into consideration and weight given to each factor will change from one disease area to another; for example, patients with more serious diseases may be more willing to accept a higher burden of harmful effects than those with

more mild diseases or prognoses. Undertaking the best analysis and presenting more informative harm profiles that could be used in benefit-risk assessments could lead to more accurate benefit-risk profiles and decisions from regulators.[70-72]

1.8 Aims

The overarching aim of the research presented in this thesis is to develop and assess new methods for the analysis and presentation of harm outcomes in phase II and III pharmacology trials that can facilitate the detection of ADRs and enable communication of informative harm profiles. This will be achieved through exploration of existing statistical methods and development of new tests to raise signals of potential ADRs for use in RCTs. The value of the tests as statistical tools to inform the planning of future studies will be explored through simulations. Alternative approaches to the presentation of this information will also be explored. To achieve this the following specific aims will be investigated:

1) Ascertain current practice for the collection, reporting and analysis of harm outcomes in phase II/III pharmacological RCTs in journal articles.
2) Identify and examine statistical methods that have been developed to analyse harm outcomes in controlled clinical trials.
3) Ascertain current practice for analysis of harm outcomes from the statistical community and identify their priorities and solutions to improve analysis practice.
4) Explore approaches to summarise and present complex harm data from RCTs including visualisations.
5) Develop and compare statistical tests to detect signals of ADRs in RCTs.
6) Assess the utility of the signal detection tests developed in (5) via simulations. Comparing these results with visualisations identified in (4) and current standard practice.

1.9 Thesis outline

In chapter two, current practices for collection, analysis and reporting of harm outcomes will be examined through a systematic review of high impact general medical journals, as well as identification of areas of good practice and any areas for improvement. In chapter three, statistical methods that have been proposed for analysis of both prespecified and emerging harms in RCTs are identified through a scoping review and are summarised and discussed. In chapter four, a survey of practicing clinical trial statisticians from both academia and the pharmaceutical industry is undertaken to measure awareness and opinions of alternative model based methods for analysis of harm outcomes and to gain an understanding of any perceived barriers to implementation of these methods, as well as seek opinions on potential facilitators to improve practice. Chapters two to four will establish the current state of play for the analysis of harms outcomes. Using the results presented in these initial chapters, alternative strategies for analysis of harm outcomes will be explored in the rest of this thesis. In chapter five visual approaches for analysis and reporting of harm outcomes are explored, a consensus to support researchers in their choice of visualisations for RCT publications is sought and recommendations are developed. In chapters six to seven development of signal detection tests to identify emerging ADRs are explored and presented. In chapter eight the results of this thesis are discussed, recommendations on analysis strategies for harms outcomes collected in RCTs are provided and an outline on future work is given.

1.10 Summary

RCTs provide an abundance of data to help establish the harm-profile of drugs but there is evidence to suggest that this data is underutilised and suboptimal analysis practices are common. In this thesis, I seek to understand current practice and develop potential solutions, including visualisations and signal detection tests with an aim to provide tools to improve practice.

2. Current analysis and reporting of emerging harm outcomes in published RCTs

2.1 Introduction

Previous studies have examined the methods for collection and presentation of emerging harm outcomes in RCTs, and highlighted the inadequacies in reporting practices in journal articles.[56, 58, 59, 73-79] In 2004 the CONSORT Group produced an extension to their guidelines for reporting trial results in journal articles to cover the reporting of harm outcomes, however implementation of these guidelines has been shown to be poor.[28, 58, 75, 78-80] In late 2016 a joint pharmaceutical/journal collaboration published practical guidance and examples on what should be reported in journal articles and how it should be displayed to ensure transparency and aid clinical interpretation for harm outcomes. The article advocated that certain events are "*always clinically relevant*" such as deaths and events leading to treatment discontinuation and should always be reported, plus events of interest that should be prespecified. In addition, they promoted the use of "*clinical judgement"* in selecting events to report, harnessing clinician experience rather than mandating what should be reported.[34] Whilst both guidelines focus on reporting practices, they were limited in recommendations for endorsing good statistical analysis. The 2009 pharmaceutical industry standard from SPERT goes some way to addressing analysis practices commenting on general analytical considerations, for example, what they consider an appropriate analysis population for harm outcomes, where graphics can add benefit and incorporating 'time at risk' as an important factor for consideration.[7] However, opinions in the research community and personal experience indicates that simple practices prevail, often with heterogeneous approaches in analysis performed on harm outcomes, specifically for emerging events, in the primary publication of RCTs.[7, 41] In this chapter I aim to formally investigate this, examining what current 'best' practice looks like for both reporting and analysis of emerging harms in the primary publication of RCTs, one of the main sources of dissemination of clinical trial results.

Some of the work presented in this chapter has been published and presented at several international conferences.[45, 81] This chapter acknowledges the support and contribution of Lorna Hazel of the Drug Safety Research Unit and University of Portsmouth who assisted with data extraction and critical revision of the published manuscript along with my PhD supervisors.

2.2 Aims

In order to ascertain best practice for collection, analysis and reporting of emerging harm outcomes in pharmacological clinical trials I undertook a systematic review of journal articles of RCTs where the primary aim was to determine efficacy. The specific objectives were to:

1) Summarise contemporary practice as exemplified in articles of RCTs published in four high-ranking general medical journals determined by impact factor.
2) To identify and highlight examples or areas of good practice.
3) To highlight any areas for improvement.

2.3 Methods

2.3.1 Eligibility criteria

Articles published in the top four general medical journals as ranked by impact factor (IF) were eligible for inclusion. This included the New England Journal of Medicine (NEJM, IF 72.41), the Lancet (IF 47.83), the Journal of the American Medical Association (JAMA, IF 44.41) and the British Medical Journal (BMJ, IF 20.79).[82] Impact factors were taken for the period from which the articles were drawn (2016). These journals each publish results of clinical trials of drug interventions and were chosen as a sample of best practice. Impact factors were selected as a proxy measure of best practice as it was expected that practice in these journals would be of the highest standard as they are highly competitive to publish in and attract large, high quality trials with a rigorous editorial process including specialist statistical and methodological peer review. Articles were eligible if they reported the results of a phase II-III RCT of an investigational medical product (IMP), where the

primary outcome was efficacy of the intervention. There were no restrictions according to number of treatment arms and both parallel and cluster RCTs were included. Crossover RCTs and RCTs with adaptive randomisation were excluded as they provide additional analysis challenges, for example, crossover RCTs raise separate issues around the appropriate analysis population for harm outcomes. Observational studies, case reports, editorials and letters were also excluded. RCTs where the intervention did not contain a drug product i.e. not classified as a CTIMP were also excluded. As the study aimed to assess how authors report and analyse harm outcomes in efficacy trials, trials that were designed and powered to demonstrate absence of harm (safety studies) were excluded as such studies would follow standard analysis practices undertaken for primary and secondary outcomes in the efficacy setting, following established reporting guidelines to comply with ICH E9 and CONSORT recommendations.[6, 83]

2.3.2 Search strategy and data extraction

A manual search of the electronic contents table of the journals for articles of original RCTs results published between September 2015 and September 2016, inclusive was conducted. Articles identified in the search were screened for inclusion based on titles and abstracts by one reviewer (RP). Full text of eligible studies were retrieved and allocated to three reviewers who undertook full text review and data extraction. Reviewers included one supervisor (VC), a collaborator (Lorna Hazell (LH)) and myself (RP). Supplementary material was retrieved, reviewed and relevant material extracted if readers were referred to them from the main article for further results. Appendix A2.1 lists all data items captured with the guidance notes given to reviewers to aid extraction. Single data extraction was performed by reviewers, with 10% independent check of a randomly sampled subset of articles to verify quality. Where specific items were flagged for poor agreement these were re-extracted. Any queries during data extraction were shared and disagreements between reviewers were resolved through discussion. Details of these discrepancies are provided in the results.

The items to be extracted were based on an earlier review by Cornelius et al., the CONSORT harm extension and additional items identified in the design stage (via discussion with supervisors) to capture more specific information on reported analysis practices.[28, 56] Information on the following areas was extracted with the rationale for each provided in table 2.1:

i) How data was collected (mode of collection, timing) and defined (coding) during the study.

ii) Assessment practices to ascertain severity of the event or relatedness to the intervention (attribution).

iii) Planned analysis including final and interim monitoring plans and analysis populations.

iv) How events were selected for inclusion in the journal article.

v) How summary event information was presented in the journal article.

vi) What analysis was undertaken?

2.3.3 Data analysis

Descriptive statistics were undertaken and included the proportion of trials reporting each item, 3-4 and 8-34 (e.g. funding source, collection method, selection criteria) and summary statistics (median and ranges) for items 5-7 (e.g. number of centres, number randomised and study duration) in appendix A2.1. Additional summaries are presented stratified by funding source. All analyses were performed in Stata version 15 or later.[84]

Table 2.1: Constructs for extraction with rationale

	Constructs	Rationale and importance of each construct
i.	How data was **collected** (mode of collection, timing) **and defined** (coding) during the study.	Variation in the collection and definition of events could explain differences in the incidence of observed events.[85, 86] For example, collecting AEs by actively asking participants about an event of interest in one trial whilst relying on patient report in another will lead to differing incidence rates between trials. Hindering between trial comparisons and any future systematic reviews and meta-analysis of harm outcomes. Differing visit schedules within trial treatment groups can also have important implications for within trial treatment comparisons, for example with increased visits in one treatment group providing more opportunity for events to be reported. Medical dictionaries are often used to standardise reported events, helping both within and between trial comparisons, however consideration to the level of coding used within a dictionary is important, as it is likely to impact detection of ADRs. For example, if classification is too broad then signals may be masked since it relies on the assumption that drugs act on the 'system' level and potentially misses important events that occur in isolation. Paradoxically if classification is too specific, splitting events into too many sub-categories, then potential ADRs will be missed because of low numbers of events occurring within sub-categories.
ii.	Assessment practices to ascertain **severity** of the event or **relatedness** to the intervention (attribution).	Attribution of causality for each event is a requirement of the ICH E8 general considerations for clinical studies guidelines. How this is done and by whom has important implications. For example, attribution by an unblinded assessor lacks objectivity and can allow bias (even if subconscious) to enter into the decision, which can lead to variations within and between trials, which can have important implications on the identification of true ADRs and subsequent risk-benefit assessments. Subjective assessments for every reported event can also become logistically burdensome.
iii.	**Planned analysis** including final and interim monitoring plans and analysis populations.	Whilst formal hypothesis tests for emerging events (not prespecified outcomes) are typically not appropriate, transparency is a core value of clinical trials and therefore it is good practice (as one would for efficacy outcomes) to provide details of the planned analysis, in which population it will be conducted and details of how conclusions will be drawn from such analyses.[28, 87] This reveals what was planned and what was undertaken post-hoc, allowing for appropriate interpretation of the results presented.
iv.	How events were **selected for inclusion** in the journal article.	Due to the space constraints in journal articles, it is not always feasible or helpful to report all events experienced by participants. Journal articles often only report a subset of events and how these are selected for inclusion has important implications for the evaluation of the harm profile. Arbitrary selection criteria can lead to inconsistencies in what is presented across trials for the same disease and/or drug. This prevents an accurate overview of the events experienced and invalidates any potential systematic review and meta-analysis of events.[56, 57]
v	How summary event **information was presented** in the journal article.	With a lack of consensus on measuring impact of harms, a potential proxy is the proportion of events that cause participants to discontinue treatment or withdraw from a study. Likewise, severity ratings can be central to risk-benefit decisions made by both patients and prescribers; and number of events experienced provides further insight into the impact, with repeated events potentially having far wider clinical implications than a single event. Failure to report such information could conceal important implications of interventions that are important to inform both patients' and clinicians' treatment decisions.
vi	What **analysis** was undertaken and on what population?	Analysis of the multifaceted data collected on harm outcomes requires careful consideration and there are many statistical challenges to consider. For example, underpowered statistical testing under the hypothesis-testing framework we adopt for primary and secondary efficacy outcomes can lead to misleading conclusions when analysing emerging harms e.g. failure to find a statistically significant result from underpowered hypothesis tests leading authors to conclude that the intervention is 'safe'.[25, 88]

2.4 Results

2.4.1 Data extraction

Each reviewer independently extracted data from a randomly selected subset of other reviewers' assigned studies. Results were compared to establish any discrepancies. Five-hundred and eighty-five items were independently extracted by a second reviewer for assessment and 95 discrepancies were identified. A small number of extracted items were flagged consistently for poor agreement. These included: study duration; the data collection method; timing of collection; how binary harm outcomes were summarised; whether continuous outcomes were dichotomised; and where continuous outcomes were left as continuous how they were analysed. Discussions amongst reviewers to redefine these questions and definitions provided clarity around what should be extracted and these items were re-extracted by a single reviewer.

Discrepancies related to study duration were due to a lack of clarity about the type of data to be extracted with one reviewer incorrectly extracting time of follow-up for the primary efficacy endpoint and not total study follow-up. Discrepancies relating to the collection method were due to a lack of clarity over the definitions of prompted and passive collection. The discrepancies between collection of laboratory values and clinical exams were because one reviewer only included this as happening if it explicitly said so, when it could often be discerned from tables of results. Likewise, table of results were also indicative if for example, laboratory values had been dichotomised but this was often missed. This also affected the information extracted on analysis methods for continuous outcomes. The discrepancies relating to timing of collection followed from the discrepancies in the preceding questions on collection methods. These discrepancies were easily resolved through discussions and amended as appropriate.

2.4.2 Study characteristics

One-hundred and eighty four articles reporting trial results were identified as eligible and included in the review (BMJ n=3; JAMA n=38, Lancet n=62; and NEJM n=81). Across included articles, 496,911 participants were randomised with a median of 556 participants per trial (range 30 to 205,513; interquartile range (IQR) 281 to 1704). The extreme upper range was due to a large vaccine trial that individually randomised over 200,000 patients, excluding this study, 291,398 participants were randomised and the median number of participants per trial was relatively unchanged at 554 but with a smaller range of between 30 and 16,590 participants (IQR 280 to 1645). The median number of participants per centre was 15 (range 1 to 15809; IQR 6 to 60). Again, these results were relatively unchanged after the removal of the large vaccine trial. The median trial follow-up was 52 weeks (range 48 hours to 10 years; IQR 24 to 104 weeks) and 93% were multi-centre trials. Fifty percent of studies had an active comparator group and over 50% of trials received some element of industry funding (table 2.2)

2.4.3 Collection and assessment methods for emerging harms (constructs i and ii of table 2.1 and items 8-11 appendix A2.1)

Variation in the means by which the occurrence of an event is elicited from participants and definitions of what constitutes an event could explain differences in the incidence of reported events between trials (table 2.1 construct i).[85, 86] Sixty-two percent of articles mentioned collection of information on emerging harms from participants but only 29% (n=53/184) of articles included specific information on how this was done e.g. prompted with questions about specific events, asked general questions about adverse effects, questionnaires, or diaries. In many articles, the reporting of the methods used to collect emerging harm outcomes were poor or absent. In table 2.3 examples 1 and 2 show how information about collection of emerging harms can be clearly incorporated into methods of published trials to comply with recommendation 4 of CONSORT harms ("*Clarify how harms-related information was collected (mode of data collection, timing, attribution methods,*

intensity of ascertainment, and harms-related monitoring and stopping rules, if pertinent").[28, 89, 90] Reports of proactive screening via clinical examinations (e.g. vital signs and blood pressure) or laboratory tests were infrequently explicitly reported, but it was clear from the results presented that participants had undergone these assessments (83% and 79% of studies reported clinical and laboratory results respectively).

Table 2.2: Characteristics of included studies†

Characteristic		N=184		
		Median	**(IQR)**	**min, max**
Sample size		556	(281, 1704)	30, 205513
Centres[a]		35	(12, 100)	1, 1368
Participants per centre[a]		15	(6, 60)	1, 15809
Trial duration (weeks)[b]		52	(24, 104)	0.3, 521
		n	**%**	
Journal	BMJ	3	1.6	
	JAMA	38	20.7	
	Lancet	62	33.7	
	NEJM	81	44.0	
Funding source[c]	Public	70	38.3	
	Industry	80	43.7	
	Both	33	18.0	
Centre	Single centre	12	7.0	
	Multi-centre	161	93.0	
Control	Placebo	95	51.6	
	Active	80	43.5	
	Both	8	4.4	
	Neither[d]	1	0.5	

Abbreviations: *IQR = Inter-quartile range; min = minimum; and max = maximum*
[a]11 articles did not specify the number of centres
[b]2 articles did not specify trial duration
[c] One trial failed to specify funding source
[d] One trial compared interventional drug to behavioural change intervention

The timing of data collection was often reported (91%, 48 out of 53 articles) in the articles that included specific details about the prompts used to collect events but the timing of clinical

† Reprinted with format modifications from *Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open 9(2): e024537) under a CC BY license:* https://creativecommons.org/licenses/by/4.0/

examinations and laboratory tests was less common, reported in 57% of articles (95 out of 166 articles with clinical examinations and/or laboratory results presented). Thirty-eight percent of articles used a dictionary to code events, with MedDRA being the most popular dictionary used. Assessment practices, i.e. who or how causality between the event and the intervention was assigned, was also poorly reported with less than 10% of articles reporting who was responsible for such assessments. Table 2.4 summarises collection and assessment practices identified in this review.

Table 2.3: Examples of good reporting practice in reviewed articles

Example	Study	Example practice	Example text
1	Litonjua et al.[89]	Description of collection method	"*Study staff met with pregnant women monthly to administer a brief health questionnaire, assess medication use, and monitor for complications (via the questionnaire and medical record review)… After delivery, children were monitored by telephone every 3 months and in-person annually for 3 years, during which time infants' health, respiratory symptoms, and medications were assessed*"
2	Miller et al.[90]	Description of collection method	"*Safety evaluations included physical examinations, assessment of vital signs, clinical laboratory tests, and reporting of adverse events at each study visit*"
3	Libman et al.[91]	Description of planned analysis	*"The proportions of participants experiencing any adverse event, any related adverse event, any gastrointestinal event, any event other than a gastrointestinal event, at least 1 severe hypoglycaemic event, and at least 1 diabetic ketoacidosis event in each treatment group were compared using the Fisher exact test. The number of adverse events, new adverse events, serious adverse events, and non-serious adverse events were compared between groups using a Wilcoxon rank sumtest."*
4	Gross et al.[92]	Description of planned analysis	"*Safety analyses and secondary efficacy analyses used binomial regression, analysis of covariance, or the marginal Cox proportional hazards model as appropriate*"
5	Marso et al.[93]	Description of planned analysis	"*We estimated the mean differences between the trial groups in the glycated hemoglobin level, weight, systolic and diastolic blood pressure, and pulse using a mixed model for repeated measurements, with adjustment for baseline covariates."*

2.4.4 Prespecified analysis for emerging harms (construct iii of table 2.1 and items 12-14 appendix A2.1)

Transparency is a core value of clinical trials as it reveals what analysis was planned and what analysis was undertaken post-hoc, allowing appropriate interpretation of the results presented.

Therefore it is good practice (as one would for efficacy outcomes) to provide details of the planned analysis.[28, 87] Table 2.4 summarises prespecified analysis practices identified in this review. Planned analysis for emerging harms was reported in less than a third of articles (31%) and just under half (45%) reported a prespecified analysis population for emerging harms, often referred to as a 'safety' population. Examples demonstrating clear incorporation of this information in the articles I reviewed are provided in table 2.3 examples 3-5. Whilst the method of analysis is clearly specified in these examples, they each lack details on analysis populations and how the results of such analyses will be interpreted, thus making results susceptible to bias. They each also fail to meet the full criteria laid out in recommendation 5 of the CONSORT harms checklist ("*Describe plans for presenting and analyzing information on harms (including coding, handling of recurrent events, specification of timing issues, handling of continuous measures, and any statistical analyses*").[28, 91, 92]

Whilst a quarter of trials reported planned interim analysis with stopping criteria, only five (2.7%) were designed to stop for a harmful event. This included: one trial where the rule was based on the primary efficacy outcome (survival) going the wrong way i.e. intervention resulting in increased mortality;[94] two trials where the rule was based on prespecified harm outcomes;[95, 96] one which looked at a range of outcomes to indicate harm, including the primary efficacy outcome going the wrong way, several prespecified harm outcomes and an increase in the rate of possible or probable ADRs;[97] and one trial where the rule was based on a comparison of event rates for each SAE.[98] Specific details reported on stopping criteria are provided in table 2.5. Example 1 (table 2.5) indicates a stopping rule for harm was used, but exemplifies a lack of transparency as authors provided no specific details.[95] In contrast examples 2-5 (table 2.5), utilised supplementary appendices to provide comprehensive details of the stopping rules used.[94, 96-98]

Table 2.4: Collection, assessment and analysis methods reported in included articles‡

Section	Component	Data item	N=184	
Collection			n	%
	How was information on emerging harms collected?	Collection mentioned	114	62.0
		Specific prompt for collection (N=114)	53	46.5
		No method of collection reported	70	38.0
	Timing of prompted collection specified (N=53)		48	90.6
	Did they undertake proactive screening?	Clinical examinations	153	83.2
		Laboratory tests	146	79.4
	Timing of active screening specified (N=166)		95	57.2
	Which, if any, dictionary was used to code data?			
		CTCAE	18	9.8
		CTCAE and MedDRA	1	0.5
		DAIDS	2	1.1
		ICD-10	1	0.5
		MedDRA	43	23.4
		Researcher defined	2	1.1
		Other	3	1.6
		No dictionary reported	114	62.0
Assessment				
	Who assigned attribution to study drug?	Blinded assessor	9	4.9
		Unblinded assessor	7	3.8
		Both	1	0.5
		Not specified	164	89.1
		Not applicable[a]	3	1.6
Analysis				
	Was any analysis for emerging harm outcomes specified in the methods section?			
		Yes	57	31.0
	Was a population for analysis of emerging harm outcomes specified?			
		Yes	82	44.6
	Was there a planned interim analysis with stopping criteria?			
		No	138	75.0
		Yes for efficacy	24	13.0
		Yes for efficacy & futility	11	6.0
		Yes for efficacy & safety	3	1.6
		Yes for efficacy, futility & safety	2	1.1
		Yes but no other details given	6	3.3

Abbreviations: *CTCAE = Common Terminology Criteria for Adverse Events; MedDRA = Medical Dictionary for Regulatory Activities; DAIDS = The Division of AIDS; and ICD-10 = International Classification of Diseases 10th revision.*

NOTE: Denominator (N) specified in item column if it differs from total sample

[a]3 articles made no reference to harm outcomes throughout the article

‡ Reprinted with format modifications from *Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open 9(2): e024537) under a CC BY license:* https://creativecommons.org/licenses/by/4.0/

Table 2.5: Prespecified stopping criteria for harm

Example	Study	Main article text	Appendix text
1	Myles et al.[95]	"O'Brien–Fleming stopping boundaries were used to assess efficacy, and *a less stringent boundary was used to assess harm*."	
2	Billings et al.[97]	"The data and safety monitoring board (DSMB) reviewed patient recruitment practices, safety reporting, and data quality after 30 patients completed the study; performed an interim analysis after 277 patients … had completed the study to assess *safety of the intervention*; and performed a second interim analysis after 546 patients … had completed the study to assess the safety, efficacy, and futility of the intervention. The DSMB made recommendations based on qualitative assessments of the safety, efficacy, and futility of the intervention…"	"*Suspend enrolment in any study arm … due to safety concerns based on study intervention. Safety concerns include*: • Increase in in-hospital all-cause mortality in subjects randomized to A or B such that the DSMB deems the increase is excessive compared to A or B. • Increased treatment toxicity in either treatment group deemed excessive. Toxicity is defined as moderate or severe myalgias. • Increased severity of adverse events deemed "Probably Related" or "Possibly Related" to study intervention in either treatment group. Itemized adverse event reports separated by treatment will be provided. • Increased AKI incidence in either treatment group deemed excessive. • Increased incidence of stroke or hemodialysis requirement in either group (secondary endpoints) deemed excessive."
3	Beardsley et al.[94]	"An independent data and safety monitoring committee oversaw trial safety and analyzed unblinded data after every 50 deaths, according to its charter ..."	"The Haybittle-Peto boundary, requiring p<0.001 at interim analysis to consider stopping for efficacy, will be used as guidance. *A level of significance of 1% will be used as a guide for stopping the trial early because of a detected harm of dexamethasone*. In addition, the DMEC will receive conditional power curves to assess whether it remains realistic that the trial will demonstrate superiority of dexamethasone conditional on the data accrued up to the point of the interim analysis. Importantly, the DMEC recommendations will not be based purely on statistical tables but will also use clinical judgment."
4	Kor et al.[98]	"In addition to statistical criteria for significance, the study included a priori "go-no-go" definitions for recommending continuation to phase 3 study ... Briefly, continuation to phase 3 would occur with a positive primary outcome finding along with an acceptable safety profile. An acceptable safety profile was defined as a serious adverse event profile for aspirin that was not statistically worse than placebo (95% CI for the relative risk of any serious adverse event covers the null value of relative risk = 1.0). The "no-go decision" was defined as early termination by the data and safety monitoring board for safety or unfavorable risk/benefit ratio. An indeterminate case in which there was a non–statistically significant effect but this effect was in a clinically meaningful direction was also defined."	Initiate Phase III Study: Demonstrated efficacy signal in addition to adequate safety profile Criteria: Early termination for benefit at interim analysis or p<0.08885 at final analysis (alpha=0.10 for study). *Serious adverse event profile of ASA not statistically worse than placebo (95% confidence interval for the relative risk of any SAE covers the null value of RR=1.0).* Further Development Potentially Required: Weak efficacy signal Criteria: Primary endpoint did not achieve a priori level of significance but there were at least a general consistency of secondary endpoints indicating propensity for efficacy with a larger sample size and/or more specific primary endpoint. Abandon Treatment Platform: Harm (in efficacy or safety endpoints) Criteria: Study terminated early per recommendation *by DSMB for safety and/or risk/benefit ratio concerns* (i.e., stop for futility, harm, unacceptable risk profile, etc.)
5	Nichol et al.[96]	We used a group sequential statistical approach to do two equally spaced pre-planned interim analyses (at 33% and 67% of total recruitment) to assess accumulated safety data (differential proportions of deep venous thrombosis and total mortality). This *approach was chosen to provide for early stopping for probable harm* or strong evidence of benefit. We applied the Haybittle-Peto criterion (|Zk|≥3) for early stopping at these analyses.	

2.4.5 *Reported results for emerging harms (constructs iv and v of table 2.1 and items 15-16, 18-21, 28-32 and 34 of appendix A2.1)*

It is not always feasible or informative to report every event experienced by participants but what is presented and how specific events are selected for inclusion has important implications for the evaluation of the harm profile. Only presenting overall summaries or using arbitrary selection criteria can lead to inconsistencies in what is presented across trials for the same disease and/or drug and prevents an accurate overview of the true harm profile.[56, 57] Five trials did not report any specific information on emerging events.[99-103] Two of these articles made the following vague statements "*there were no significant adverse events related to the procedure*" and "*no excess in mortality or major adverse events were found"*.[99, 103] Two articles only reported prespecified secondary harm outcomes, which included a trial on children with uncomplicated severe malnutrition that reported, "*No cases of severe allergy or anaphylaxis were identified. None of the clinical complications or deaths were reported to be related to the study drug.*" Despite not presenting results on emerging harms, the first sentence in this quote demonstrates elements of good practice by reporting on events despite none occurring. However, the value of the second sentence is diminished as it is unclear how this causality assessment was made, but the authors do report that "*all clinical and research staff members were unaware of treatment assignment"* therefore we know that these assessments were made blind to treatment allocation. [99] One trial failed to make any mention of harm outcomes. This trial looked at early versus late antiretroviral therapy in HIV-1 patients and only reported results for the primary outcome. Results of this trial were released early and the intervention was offered to the control arm due to promising interim efficacy results but the authors report that the trial continued as planned.[101]

Twenty-four (13%) trials only provided a summary of aggregated number of events or serious events in the main journal article providing no details on the actual events experienced. For example, "*Six serious adverse events occurred in the acetaminophen group and 12 in the ibuprofen group.*"[104] Ten

of which utilised supplementary material to provide specific details on the types of events. This left 8% of trials either not reporting any information on harm or only including a summary statement.

Table 2.6 summarises the reporting practices identified in this review.

Table 2.6: Reporting practices across included articles[§]

Component	Data item	N=184	
		n	%
What was reported in the manuscript?	Actual event terms	73	39.7
	Summaries of event type (e.g. AE, SAE)	24	13.0
	Both	80	43.5
	Neither	7	3.8
What was reported in the appendix?	Actual event terms	76	41.3
	Summaries of event type (e.g. AE, SAE)	7	3.8
	Both	22	12.0
	Neither	3	1.6
	Not applicable[a]	76	41.3
Did the report reference the CONSORT extension to harms?			
	No	184	100.0
	Yes	0	0.0

Abbreviations: *AE = Adverse Event; SAE = Serious Adverse Event; CI = Confidence Interval; and CONSORT = Consolidated Standards for Reporting Trials.*

[a] Make no reference to the appendix

Eleven percent of trials reported all the events they collected in the journal article. For the remaining studies, how events were 'selected' for inclusion was not consistent or always clear. For 3% of studies it was impossible to discern how the authors had selected the events they presented in the journal article; a further 3% only reported summaries, 6% did not present results in the journal article and 14% only presented prespecified events. The majority of studies (63%) used a rule-based approach to select events to report in the manuscript. The use of such rules will place more focus on common events than important events and can lead to inconsistencies in what is presented across trials for the same disease and/or drug. This prevents an accurate overview of the events experienced and invalidates any potential systematic review and meta-analysis of events.[56, 57] Table

[§] Reprinted with format modifications from *Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open 9(2): e024537) under a CC BY license: https://creativecommons.org/licenses/by/4.0/*

2.7 summarises selection criteria used; percentages are not independent as the majority of articles used several different criteria for selection. Twenty-eight percent of articles included events based on exceeding a frequency threshold e.g. events experienced in at least 10% of participants in any treatment group; 23% of articles included events if they were classified as serious; 9% of articles included events if they exceeded a severity threshold e.g. events of grade 3 or higher; 8% included events based on perceived relatedness to treatment; 3% included events that led to treatment discontinuation or interruption; and 12% used an ambiguous threshold e.g. the most common events. Example of combinations of rules are provided in table 2.8. Appendices A2.2 and A2.3 provides full details of selection criteria used across the journal article and supplementary material, respectively.

Table 2.7: Selection criteria categories for choice of events included in articles

Selection category	**n (%)**	**Examples of selection criteria used**
Frequency threshold	52 (28.3)	*"AEs that occurred in at least 10% of patients in any treatment group"*
		"AEs that occurred in two or more patients receiving treatment or placebo"
		"AEs reported by at least two patients"
Seriousness	43 (23.4)	*"Overall summaries for predefined harms events, AEs leading to hospitalisation, AEs leading to death, AEs leading to permanent study drug discontinuation, AEs leading to temporary study discontinuation, SUSARs"*
		"Serious allergic reactions"
		"SAEs that occurred in at least 0.5% of patients and treatment related AEs and SAEs"
Ambiguous threshold	22 (12.0)	*"Most common AEs"*
		"Most common SAEs"
		"Most common (no criteria specified) grade 3 or higher AEs"
Severity threshold	17 (9.2)	*"Grade 3 and 4 AEs"*
		"Grade 3 or 4 non-hamatological events"
		"Grade 3 or higher laboratory events"
Relatedness	15 (8.2)	*"Intervention related AEs and death"*
		"Treatment related SAEs"
		"Infusion related reactions in more than 5% of patients"
Treatment discontinuation or interruption	6 (3.3)	*"AEs leading to study drug discontinuation/interruption"*

Table 2.8: Example of multi-faceted rules used to select events to report in journal articles

Example	Study	Examples of combinations of rules
1	Burger et al.[105]	*"AEs that occurred in at least 15% of patients in either group and for which the frequency differed between treatment groups by at least 5% and grade 3 or higher or SAEs that occurred in at least 2% of patients in either treatment group"*
2	McInnes et al.[106]	*"AEs that occurred in more than 2% of pooled intervention groups until end of treatment or AEs with an incidence of at least 5 per 100 patient-years in the pooled intervention group until the end of study"*
3	Herbst et al.[107]	*"AEs related to treatment occurring in more than 10% of patients in any treatment group, AEs of special interest occurring in more than 2% of patients in the intervention group, and deaths related to study treatment"*

Useful proxies that are often used as an overall measure of the impact of harm include: the proportion of participants that discontinue treatment or withdraw from a study due to harm; severity ratings; and number of events experienced. Seventy-nine percent of trials reported the number of participants who withdrew from the trial. Whether the withdrawals were due to harm was reported in 35% of articles (51 of 146 articles) and of these 24% (12 of 51 articles) reported the actual events that caused withdrawals in line with recommendation 6 of the CONSORT harms checklist ("*Describe for each arm the participant withdrawals that are due to harms and the experience with the allocated treatment*").[28] Eighty-four percent of articles performed analysis and presented results on 'participants with at least 1 event' providing no information on recurrent events (table 2.9).

Twenty-eight percent of articles included information on the timing of events and 5% reported information on duration for at least one event (table 2.9). The trials that presented information on duration did so in a variety of ways, including individual participant listings of events with durations, incorporating the information into a table of events with summary statistics such as the mean duration of events or presenting it for a subgroup of events in the footnotes of tables.[108-110]

Only forty-one percent of articles reported information on the severity rating of events but the number of serious events were typically well-documented (73%). A further six articles (3%) explicitly stated that no serious events had occurred. However, in forty-four (24%) articles it was not possible to determine if no serious events had occurred or whether the authors had failed to report them in the article. Forty-two percent (57 of 134 reports) of articles reported whether the events had been classified as related to the intervention (table 2.9).

2.4.6 Analysis of emerging harm outcomes (construct vi of table 2.1 and items 17, 22-27 and 33 of table A2.1)

Analysis population

The ITT population is typically used for efficacy outcomes but in the context of harm outcomes, this population may be considered conservative, as it is likely to underestimate the risk of an event by inflating the denominator with participants who may have never received the study drug. However, any other population will not preserve the balance between treatment groups achieved by randomisation. Therefore, no one correct population for the analysis of harm outcomes is obvious. This review revealed that the most common analysis population used was "*participants that received at least one dose*" (41%), followed by 29% of trials that used "*all randomised*" participants and 9% that did not specify the analysis population (table 2.10). A further 19% reported variations of treated and/or randomised such as "*took a single dose and underwent AE/toxicity assessment*", "*patients who treatment was at least attempted on*", "*randomised and underwent AE/toxicity assessment*" or "*randomised and assessed for primary outcome*". Analysis populations used were many and varied and are summarised in table 2.10.

Table 2.9: Summary of results presented and analysis undertaken[**]

Component	Data item	N=184	
		n	%
Which population was the analysis of emerging harms performed on?	All randomised	54	29.4
	Those that took at least a single dose	75	40.8
	Other	35	19.0
	Not specified	17	9.2
	Not applicable[a]	3	1.6
Were drop-outs/withdrawals reported?	No	33	17.9
	Yes by treatment arm	144	78.3
	Yes overall	2	1.1
	Not applicable[b]	5	2.7
Were withdrawals due to harms reported? (n=146)	No	89	61.0
	Yes	51	34.9
	Not applicable[c]	6	4.1
Were specific events causing withdrawals reported? (n=51)	No	39	76.5
	Yes	12	23.5
How were binary emerging harm outcomes summarised by arm?	Not summarised[d]	6	3.3
	Number of people with an event	154	83.7
	Number of events	11	6.0
	Both	12	6.5
	Unclear	1	0.5
Were frequencies of events reported by arm?	No	5	2.7
	Yes for some	13	7.1
	Yes for all	160	87.0
	Not applicable[d]	6	3.3
Were percentages of events reported by arm?	No	18	9.8
	Yes for some	25	13.6
	Yes for all	135	73.4
	Not applicable[d]	6	3.3
Were between arm differences & 95% CIs of events reported?	No	140	76.1
	Yes for some	17	9.2
	Yes for all	21	11.4
	Not applicable[d]	6	3.3
Were statistical significance tests between arms on events reported?	No	92	50.0
	Yes for some	31	16.9
	Yes for all	55	29.9
	Not applicable[d]	6	3.3
Were continuous emerging harm outcomes dichotomised for summaries?	No	10	5.4
	Yes for some	30	16.3
	Yes for all	106	57.6
	Not applicable	38	20.7
What between arm analyses was performed on continuous outcomes (not dichotomised)? (N=40)			

[**] Reprinted with format modifications from *Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open 9(2): e024537) under a CC BY license:* https://creativecommons.org/licenses/by/4.0/

Differences in measures of central tendency estimated with 95% CI	No	24	60.0
	Yes for some	1	2.5
	Yes for all	15	37.5
Between arm hypothesis tests performed	No	12	30.0
	Yes for some	2	5.0
	Yes for all	26	65.0
Were any 'signal detection' approaches used?	No	184	100.0
	Yes	0	0.0
Were there any graphical presentations of events?	No	162	88.0
	Yes	22	12.0
Were summaries of severity rating of events reported?	No	103	56.0
	Yes for some	41	22.3
	Yes for all	35	19.0
	Not applicable[e]	5	2.7
Were the number of serious events reported?	No	44	23.9
	Yes overall	2	1.1
	Yes by treatment arm	132	71.7
	Not applicable[f]	6	3.3
For serious events was relatedness given? (N=134)	No	77	57.5
	Yes for some	18	13.4
	Yes for all	38	28.4
	Yes overall	1	0.8
Were there any events where information on duration was reported?	No	175	95.1
	Yes	9	4.9
Were there any events where information on the time of occurrence was reported?	No	132	71.7
	Yes	52	28.3
If any significance tests were performed on events was multiplicity accounted for?	No	81	44.0
	Yes	3	1.6
	Not applicable	100	54.4

Abbreviations: *SAE = Serious Adverse Event; CI = Confidence Interval; and CONSORT = Consolidated Standards for Reporting Trials.*

[a] 3 articles made no reference to AEs or harms data throughout the article

[b] 5 articles indicate no withdrawals

[c] 6 articles specify the number of withdrawals and reasons but none of the reasons are related to AEs

[d] This includes 3 reports with no data on harm outcomes (as per footnote [a]), 2 reports that provide generic statements regarding harmful events and 1 report that only reported continuous outcomes

[e] This includes 3 reports with no data on harm outcomes and 2 reports that provide generic statements regarding harm data (as per footnote [d])

[f] 6 papers specifically state that no serious events occurred

Table 2.10: Detailed population summaries used for analysis

Analysis population	n	%
Those that took at least a single dose	75	40.8
All randomised	54	29.4
Randomised and not withdrawn/ineligible	19	10.3
Not specified	17	9.2
Not applicable	3	1.6
Took a single dose and underwent AE/toxicity assessment	3	1.6
Active treatment groups	2	1.1
Completed treatment and assessed for primary outcome	2	1.1
Other	2	1.1
Patients who treatment was at least attempted on	1	0.5
Intention-to-treat population	1	0.5
Randomised and assessed for primary outcome	1	0.5
Randomised and attended at least one follow-up visit	1	0.5
Randomised and remained in follow-up	1	0.5
Randomised and underwent AE/toxicity assessment	1	0.5
Randomised, eligible and received at least a single dose	1	0.5

Analysis of binary outcomes

Binary outcomes were predominantly summarised using frequencies (94%) and percentages (87%) (table 2.9). Nine percent (n=16) of articles reported incidence rates (i.e. accounting for participant exposure time). Twenty-one percent of articles reported between group differences using risk differences (n=11, (6%)), risk ratios (n=12, (7%)), odds ratios (n=11, (6%)) or hazard ratios from the Cox proportional hazard models (n=4, (2%)), all with accompanying confidence intervals.

There are many challenges to consider when analysing harm outcomes in clinical trials and inappropriate statistical testing under a hypothesis-testing framework can lead to misleading conclusions. Forty-seven percent of articles undertook formal hypothesis-testing, reporting p-values from a variety of statistical tests and often drew inappropriate conclusions from such analyses. For example, one article concluded "*There were no between-group differences in the rate of patients with at least 1 adverse event (16.7% [14 patients] in the clopidogrel group vs 21.8% [19 patients] in*

the placebo group; difference, −5.2% [95% CI, −17% to 6.6%]; P = .44)." failing to acknowledge that the trial was not powered to detect such a difference.[111]

Forty-eight percent of those conducting hypothesis tests on binary outcomes used either the Fisher's exact test or chi-squared test (it was not always clear which) making the unlikely assumption that follow-up was complete across participants. Four articles presenting incidence rates undertook formal statistical comparisons, one using the Poisson exact test, one using Poisson regression models, one using negative binomial models and one not specifying how they made the comparisons. Two additional studies used the Poisson distribution to calculate confidence intervals within treatment groups. Whilst statistical testing under a hypothesis-testing framework is inappropriate in this setting, approaches that account for the likely differential follow-up across participants to estimate effects are preferred to the simple proportions. One article stated "*a Poisson regression model was used to analyse exposure-adjusted incidence rates*" but no such analysis was presented. Four studies compared the total number of events between arms using a variety of tests including the Wilcoxon rank sum test, an ordinal regression model and a Poisson regression model. Utilising information on repeated events rather than relying on reports of those experiencing at least one event should be encouraged. One study reported using a Mann-Whitney U test (synonymous with Wilcoxon rank sum test) for comparison of severity grades of events (ordinal outcomes). Twenty-six (30%) studies reported results from formal hypothesis tests but failed to provide details on which test(s) had been undertaken, again demonstrating a lack of transparent reporting for harm outcomes.

Analysis of continuous outcomes

It is well known that dichotomisation of continuous outcomes is bad practice, however fifty-eight percent of studies dichotomised continuous clinical and laboratory outcomes into normal and abnormal values.[112] Of the trials that retained clinical and laboratory outcomes as continuous data,

70% performed statistical significance testing (table 2.9). The statistical tests undertaken included the Student's t-test (n=4 (14%)) or the Wilcoxon rank sum test (n=3 (11%)) and two (7%) studies where it was unclear if the t-test or Wilcoxon rank sum test was used. Six studies (21%) used analysis of covariance (ANCOVA), four (14%) fitted linear mixed models (an example of how this is reported is provided in table 2.3, example 5), one (4%) study used a Poisson regression model and one (4%) used marginal models with generalised estimating equations optimising use of information on repeated measurements. A further seven (25%) studies did not specify which test was used to calculate the p-values presented.

Multiplicity corrections

No multiplicity corrections for the multiple statistical tests performed on continuous outcomes were made. Of the trials that performed statistical significance testing on emerging harm outcomes, only three (2%) made an adjustment for multiplicity of tests (all three on binary outcomes).[108, 113, 114] Two of which used a Bonferroni correction, adjusting for the number of pairwise comparisons between each of the treatment groups for each individual event rather than the total number of significance tests performed across outcomes and would therefore have still been affected by issues of multiplicity.

Graphical approaches

Graphics were used in 12% of articles to present data on harm outcomes (table 2.9). Forty-one percent (n=9/22) of the plots were for binary outcomes and included dot plots, bar charts and Kaplan-Meier plots; and fifty-nine percent (n=13/22) plotted continuous outcomes using line graphs and scatter plots.

2.4.7 Results summarised by funding source

Forty three percent of studies were funded by industry (43%), compared to 38% publicly funded and 18% receiving both industry and public funds (table 2.2). For one article specific details on funding source were not reported. Characteristics of included studies were broadly similar with marginally more single site studies conducted by publicly funded sources (10% versus 3% and a median of 13 (IQR 4, 29) centres per study versus 76 (IQR 35, 148) centres per study) and of shorter duration (median 39 weeks (IQR 22, 104) versus 52 weeks (IQR 26, 100)) (summarised in appendix A2.4 and A2.5).

The reporting of the collection methods for emerging events were broadly similar across funding sources (table 2.11). Notable differences included fewer publicly funded studies reporting that they undertook clinical and laboratory monitoring during the trial. Clinical monitoring examinations were reported by 77% of publicly funded studies versus 90% of industry funded studies and laboratory tests were reported by 73% of publicly funded studies versus 85% of industry funded studies.

Industry funded studies were more likely to provide information on how emerging harms would be analysed, with 39% reporting this in the methods section compared to 26% of publicly funded studies, and 73% reporting the planned analysis population for harm outcomes compared to 26% of publicly funded studies (table 2.11).

There were several notable differences across funding source concerning the results presented and the analyses performed (table 2.12 and 2.13). Publicly funded studies were more likely to report overall summaries of harm (23% compared to 6.3%). Industry funded trials typically used the population who had taken a single dose of the intervention as their analysis population (70% versus 16%), with publicly funded studies using all randomised (36%) or another study specific population (29%). Reports of withdrawals were similar, but industry funded studies were more likely to report

whether withdrawals were due to harm (57% versus 21%) and details of the specific events leading to withdrawals (28% versus 8%). More publicly funded studies reported between arm differences and 95% confidence intervals (30% versus 9%) and results of hypothesis tests (60% versus 26%). Industry funded studies more frequently dichotomised continuous outcomes (85% versus 61%), but when analysed as continuous outcomes, publicly funded studies were more likely to perform hypothesis tests (82% versus 59%). Industry funded studies were also more likely to present results on severity ratings (48% versus 31%), seriousness (85% versus 64%), relatedness (44% versus 29%), duration (8% versus 3%) and timing of events (36% and 17%), as well as being more likely to use graphical representations (19% versus 6%).

Table 2.11: Collection, assessment and analysis methods reported in included articles by funding source

Section	Component	Data item	Public (N=70)		Industry (N=80)		Both (N=33)	
Collection			n	%	n	%	n	%
	How was information on emerging harms collected?	Collection mentioned	42	60.0	50	62.5	21	63.6
		Specific prompt for collection (N=113)	25	35.7	20	25.0	7	21.2
		No method of collection reported	28	40.0	30	37.5	12	36.4
	Timing of prompted collection specified (N=53)		23	92.0	18	90.0	6	85.7
	Did they undertake proactive screening?	Clinical examinations	54	77.1	72	90.0	27	81.8
		Laboratory tests	51	72.9	68	85.0	27	81.8
	Timing of active screening specified (N=166)		40	66.7	35	46.1	20	66.7
	Which, if any, dictionary was used to code data?							
		CTCAE	4	5.7	8	10.0	6	18.2
		CTCAE and MedDRA	0	0.0	1	1.3	0	0.0
		DAIDS	2	2.9	0	0.0	0	0.0
		ICD-10	1	1.4	0	0.0	0	0.0
		MedDRA	7	10.0	29	36.3	7	21.2
		Researcher defined	0	0.0	2	2.5	0	0.0
		Other	1	1.4	1	1.3	1	3.0
		No dictionary reported	55	78.6	39	48.8	19	57.6
Assessment								
	Who assigned attribution to study drug?	Blinded assessor	2	2.9	6	7.5	1	3.0
		Unblinded assessor	2	2.9	2	2.5	3	9.1
		Both	0	0.0	1	1.3	0	0.0
		Not specified	64	91.4	70	87.5	29	87.9
		Not applicable[a]	2	2.9	1	1.3	0	0.0
Analysis								
	Was any analysis for emerging harm outcomes specified in the methods section?							
		Yes	18	25.7	31	38.8	8	24.2
	Was a population for analysis of emerging harm outcomes specified?							
		Yes	13	18.6	58	72.5	11	33.3
	Was there a planned interim analysis with stopping criteria?							
		No	49	70.0	63	78.8	25	75.8

Yes for efficacy	10	14.3	12	15.0	2	6.1
Yes for efficacy and futility	3	4.3	4	5.0	4	12.1
Yes for efficacy and safety	2	2.9	0	0.0	1	3.0
Yes for efficacy, futility and safety	2	2.9	0	0.0	0	0.0
Yes but no other details given	4	5.7	1	1.3	1	3.0

Abbreviations: *AE = Adverse event; CTCAE = Common Terminology Criteria for Adverse Events; MedDRA = Medical Dictionary for Regulatory Activities; DAIDS = The Division of AIDS; and ICD-10 = International Classification of Diseases 10th revision.*

NOTE: Denominator specified in item column if it differs from total sample

[a]3 reports made no reference to harm outcomes throughout the article

Table 2.12: Reporting practices across included articles by funding source

Component	Data item	Public (N=70)		Industry (N=80)		Both (N=33)	
		n	%	n	%	n	%
What was reported in the manuscript?							
	Actual AE terms	27	38.6	28	35.0	18	54.5
	Summaries of AE type	16	22.9	5	6.3	3	9.1
	Both	23	32.9	46	57.5	10	30.3
	Neither	4	5.7	1	1.3	2	6.1
What was reported in the appendix?							
	Actual AE terms	29	41.4	37	46.3	10	30.3
	Summaries of AE type	2	2.9	3	3.8	2	6.1
	Both	6	8.6	12	15.0	4	12.1
	Neither	1	1.4	2	2.5	0	0.0
	Not applicable[a]	32	45.7	26	32.5	17	51.5
Did the report reference the CONSORT extension to harms?							
	No	70	100.0	80	100.0	33	100.0
	Yes	0	0.0	0	0.0	0	0.0

Abbreviations: *AE = Adverse Event; SAE = Serious Adverse Event; CI = Confidence Interval; and CONSORT = Consolidated Standards for Reporting Trials.*

[a] Make no reference to the appendix

Table 2.13: Summary of analysis practices by funding source

Component	Data item	Public (N=70)		Industry (N=80)		Both (N=33)	
		n	%	n	%	n	%
Which population was the analysis of emerging harms performed on?	All randomised	25	35.7	14	17.5	14	42.4
	Those that took at least a single dose	11	15.7	56	70.0	8	24.2
	Other	20	28.6	4	5.0	11	33.3
	Not specified	12	17.1	5	6.3	0	0.0
	Not applicable[a]	2	2.9	1	1.3	0	0.0
Were drop-outs/withdrawals reported?	No	11	15.7	16	20.0	5	15.2
	Yes by treatment arm	56	80.0	63	78.8	25	75.8
	Yes overall	1	1.4	0	0.0	1	3.0
	Not applicable[b]	2	2.9	1	1.3	2	6.1
Were withdrawals due to harms reported? (N=146)	No	41	71.9	26	41.3	22	84.6
	Yes	12	21.1	36	57.1	3	11.5
	Not applicable[c]	4	7.0	1	1.6	1	3.9
Were specific events causing withdrawals reported? (N=51)							
	No	11	91.7	26	72.2	2	66.7
	Yes	1	8.3	10	27.8	1	33.3
How were binary emerging harm outcomes summarised by arm?	Not summarised[d]	4	5.7	1	1.3	1	3.0%
	Number of people with an event	58	82.9	67	83.8	28	84.8
	Number of events	5	7.1	3	3.8	3	9.1
	Both	2	2.9	9	11.3	1	3.0
	Unclear	1	1.4	0	0.0	0	0.0
Were frequencies of events reported by arm?	No	3	4.3	2	2.5	0	0.0
	Yes for some	5	7.1	5	6.3	3	9.1
	Yes for all	58	82.9	72	90.0	29	87.9
	Not applicable[d]	4	5.7	1	1.3	1	3.0
Were percentages of events reported by arm?	No	14	20.0	4	5.0	0	0.0
	Yes for some	11	15.7	9	11.3	5	15.2

	Yes for all	41	58.6	66	82.5	27	81.8
	Not applicable[d]	4	5.7	1	1.3	1	3.0
Were between arm differences & 95% CIs of events reported?	No	45	64.3	72	90.0	23	69.7
	Yes for some	7	10.0	5	6.3	5	15.2
	Yes for all	14	20.0	2	2.5	4	12.1
	Not applicable[d]	4	5.7	1	1.3	1	3.0
Were statistical significance tests between arms on events reported?							
	No	24	34.3	58	72.5	9	27.3
	Yes for some	15	21.4	9	11.3	7	21.2
	Yes for all	27	38.6	12	15.0	16	48.5
	Not applicable[d]	4	5.7	1	1.3	1	3.0
Were continuous emerging harm outcomes dichotomised for summaries?							
	No	6	8.6	2	2.5	2	6.1
	Yes for some	11	15.7	15	18.8	4	12.1
	Yes for all	32	45.7	53	66.3	21	63.6
	Not applicable	21	30.0	10	12.5	6	18.2
What between arm analyses was performed on continuous outcomes (not dichotomised)? (N=40)							
Differences in measures of central tendency estimated with 95% CI							
	No	8	47.1	10	58.8	6	100.0
	Yes for some	1	5.9	0	0.0	0	0.0
	Yes for all	8	47.1	7	41.2	0	0.0
Between arm hypothesis tests performed	No	3	17.7	7	41.2	2	33.3
	Yes for some	2	11.8	0	0.0	0	0.0
	Yes for all	12	70.6	10	58.8	4	66.7
Were any 'signal detection' approaches used?	No	70	100.0	80	100.0	33	100.0
	Yes	0	0.0	0	0.0	0	0.0
Were there any graphical presentations of events?	No	66	94.3	65	81.3	30	90.9
	Yes	4	5.7	15	18.8	3	9.1
Were summaries of severity rating of events reported?	No	44	62.9	41	51.3	17	51.5
	Yes for some	14	20.0	20	25.0	7	21.2

	Yes for all	8	11.4	18	22.5	9	27.3
	Not applicable[e]	4	5.7	1	1.3	0	0.0
Were number of serious events reported?	No	21	30.0	10	12.5	12	36.4
	Yes overall	1	1.4	1	1.3	0	0.0
	Yes by treatment arm	44	62.9	67	83.8	21	63.6
	Not applicable[f]	4	5.7	2	2.5	0	0.0
For serious events was relatedness given? (N=134)	No	31	68.9	37	54.4	9	42.9
	Yes for some	1	2.2	11	16.2	6	28.6
	Yes for all	13	28.9	19	27.9	6	28.6
	Yes overall	0	0.0	1	1.5	0	0.0
Were there any events where information on duration was reported?							
	No	68	97.1	74	92.5	32	97.0
	Yes	2	2.9	6	7.5	1	3.0
Were there any events where information on the time of occurrence was reported?							
	No	58	82.9	51	63.8	22	66.7
	Yes	12	17.1	29	36.3	11	33.3
If any significance tests were performed on events was multiplicity accounted for?							
	No	38	54.3	22	27.5	21	63.6
	Yes	2	2.9	1	1.3	0	0.0
	Not applicable	30	42.9	57	71.3	12	36.4

Abbreviations: *AE = Adverse Event; SAE = Serious Adverse Event; CI = Confidence Interval; and CONSORT = Consolidated Standards for Reporting Trials.*

[a] 3 reports made no reference to AEs or harms data throughout the article

[b] 5 reports indicate no withdrawals

[c] 6 reports specify the number of withdrawals and reasons but none of the reasons are related to AEs

[d] This includes 3 reports with no AE data (as per footnote [a]), 2 reports that provide generic statements regarding harm data and 1 report that only reported continuous outcomes

[e] This includes 3 reports with no AE data and 2 reports that provide generic statements regarding AE data (as per footnote [d])

[f] 6 papers specifically state that no serious adverse events occurred

2.5 Discussion

To ensure that an informative and comprehensive presentation of the harm profile is available to prescribers, researchers, patients and regulators, clear and consistent reporting of data on emerging harms from clinical trials is essential. One of the main sources of information for these key stakeholders are journal articles of trial results. Previous research has shown the quality of reporting in journal articles is insufficient.[56, 58, 59, 73-79] This review goes beyond the existing research to examine contemporary practice for analysis of harm outcomes in the top four general medical journals, as well as examining collection and reporting practices, with the aim of identifying any areas for improvement and highlighting any examples of good practice.

Historical inadequacies in the reporting of harm outcomes led to the development of the CONSORT extension to harm, which aimed to improve the reporting of harm outcomes in RCTs. [28] Of the ten CONSORT recommendations made, this review found that many were not well reported. This has been confirmed in a subsequent review that explicitly assessed the impact of CONSORT harms on reporting practices and concluded there had been minimal improvements since its publication.[80] These results suggest that the CONSORT recommendations are not being used to aid the reporting of harms data. Journals typically request that authors include a completed, standard CONSORT checklist when they submit an article reporting the results of a RCT but to my knowledge, no journals request the CONSORT harm extension be submitted alongside. A review of the instructions to authors for each of the BMJ, JAMA, Lancet and NEJM, found that the Lancet was the only journal that made specific reference to the harms extension. In the Lancet, authors reporting harms are asked to describe them according to the extension but no formal requirement to provide evidence with a submission that the ten items have been addressed is required. This finding is in line with the lack of endorsement identified in an earlier review by Shamseer et al.[115] The 2010 CONSORT statement which is widely adopted contains a single item related to harms, it states '*all important harms or unintended effects in each group*' should be reported.[83, 116] This vague, subjective

statement may explain why many items listed on the CONSORT extension for harm were reported by so few trials. This is a sentiment that the current harms CONSORT group seem to agree with, and are currently working towards an update that integrates harm outcomes into the main CONSORT statement. In the interim the adoption of CONSORT harms by journals may support better reporting and is a view supported by others who have suggested that journals should "*strengthen their requirements for safety reporting by specifically mentioning the extensions and the updates of the CONSORT guidelines in their ''Instruction to Authors'' as well as in their ''Instruction to Reviewers.*""[117] This also calls for greater responsibility from reviewers to request better reporting of harms and to use their review as an opportunity to raise awareness to the extended guidelines. Whilst some may argue that this is beyond the remit of a reviewer and should fall to journal editors, it highlights that there is scope for improvement, which within the current framework should be easy to implement.

2.5.1 Summary and implications of this review

In the following, I summarise the findings of this review, putting them in the context of other research in this area and highlight components that I believe are important to include in reports of clinical trial results. The key components for reporting are summarised in table 2.14.

The method by which data on emerging harms was collected was inadequately reported across articles, a result which is in line with two recent reviews that also examined collection methods.[58, 79] Collection practices have important implications for the type and frequency of events reported, for example with "*passive collection resulting in fewer recorded AEs*" and frequency of collection directly correlating with the number of events reported.[25, 85, 86] For example, more frequent assessment and longer follow-up will result in more events reported.[28] These are important factors to take into consideration when making conclusions about the harm profile. However, if inadequate information is provided on collection practices then such assessments are not feasible.

How the occurrence of an event is attributed to a drug was also poorly reported. Whilst this result is in line with the work of Favier et al. who examined trials that received funding from the French Programme Hospitalier de Recherche Clinique, Hum et al. showed that studies in paediatric acute otitis media had high levels of adherence to this component.[58, 79] The authors themselves comment that they observed higher levels of reporting adherence compared to previous trials but provided no explanation for this. Not reporting this information leaves the reader ill-informed to how causality was attributed for each event. A causal classification for each event is a requirement of the ICH E8 general considerations for clinical studies guidelines, which is adopted by both the EMA and FDA, and is important information to ensure each event is classified and escalated appropriately i.e. serious unexpected ADRs require a judgement about both causality and expectedness, and require expedited reporting to relevant bodies such as the sponsor.[118] Thus how this assessment is made and by whom is important as variations both within (i.e. across centres) and between trials will introduce a bias into the assessment. If this information is unavailable, it is not possible to assess this potential source of bias. Where this information was reported, it was clear there was a reliance on subjective assessments, which can be resource intensive and can lack consistency when required for every emerging event reported in a trial. The joint pharmaceutical/journal collaboration indicated that such attribution has "*limited value"* given the "*inherent subjectivity*" involved.[34] However, given the very point of collecting data on emerging harms is to aid identification of ADRs it seems that a causal assessment is necessary. More objective, efficient means by which to make such assessments will be explored in later chapters of this thesis.

Whilst the majority of trials in this review included a report of harms alongside benefit, many included generic summary statements regarding the harm profile such as "*the intervention was well tolerated"* or "*the intervention exhibited a good safety profile"* and in some instances provided no information at all on emerging harm outcomes. Sacks et al. criticised use of such reassuring descriptions which fail to acknowledged the more complex situation and are susceptible to varied

interpretations.[119] When such statements are based on 'non-significant' p-values (i.e. p-value > 0.5) from inappropriate hypothesis tests authors are also failing to recognise "*that absence of evidence is not evidence of absence*".[120, 121] In addition, the failure to report information on emerging harms restricts interpretation preventing a critical appraisal of harm and development of the harm profile, and thus an accurate risk-benefit assessment being made. As harm profiles are developed on accumulating evidence, it is important that each study report to the same standard and information be fully utilised. Cornelius et al. proposed that such summaries could be based on core outcome sets by drug class, serious events and any event leading to treatment discontinuation or study withdrawal and any pre-specified events.[56]

It is common for vast numbers of emerging harms to be experienced during a trial and there is a need to consider carefully how informative information regarding likely drug-event relationships can be better presented. Selection criteria used by authors to choose which events they include in articles were often arbitrary and inconsistent and the findings of this review are similar to other reviews, which demonstrated wide variation in selection criteria used.[25, 56, 57] This has important implications for systematic reviews and meta-analyses that synthesise data across studies to construct harm profiles. It also, as suggested by Mayo-Wilson et al., introduces the potential for investigators to "*cherry-pick*" results presented given the lack of evidence that selection criteria are prespecified.[57] Lineberry et al. provided some guidance on this, recommending that deaths, serious events and events leading to discontinuation of intervention should always be reported and criteria such as events of interest based on the disease(s) under investigation, comorbidities of the study population, intervention mechanism, and trial duration that should be considered when deciding what other events to report.[34] However, there is a lack of guidance to facilitate consistency and objectivity, and further research in this area is needed.

When undertaking any analysis it is important to be clear about what is to be estimated and how the results will be used to draw inferences. The importance of this in the harms setting is highlighted by recommendations from the CIOMS VI working group and recent work discussing harm outcomes in reference to the newly emerging estimand framework.[40, 62] For example, inferences drawn can be substantially impacted by the analysis population used. CONSORT recommend that analysis of harm data should be in the ITT population to maintain the random assignment.[28] However, this review revealed that this population label is not always appropriately or consistently applied when analysing harms and there was a substantial disparity seen between analysis populations used in industry funded studies compared to publicly funder studies. In addition, articles frequently reported that a modified ITT population is used without providing details about the modifications made. It has also been argued that the ITT or modified-ITT populations are inappropriate for harm outcomes as they are likely to underestimate the risk of an event by inflating the denominator with participants who may have never received the study drug.[122] Whilst there are scenarios where such estimates might be considered appropriate, for example, for health economic evaluations where estimates of the cost-effectiveness will inform policy level decisions regarding how to treat the population. A more appropriate population for analysis of harms to inform clinician and patient decisions may be those that receive at least some quantity of the intervention and this is a view supported by CIOMS VI and the SPERT initiative.[7, 40] Patson et al. also found variation in populations used when reviewing analysis of emerging harms in malaria trials but given the differences in guidelines it is perhaps unsurprising that such variations are observed.[123]

Information on withdrawals, the number of events participants experience, severity and duration can all be useful indicators of the impact of harms on patients. The number withdrawing for any reason was consistently reported which is likely to be as a result of it being a recommendation of the 2010 CONSORT statement, but information such as withdrawals due to harm were inconsistent.[83] Both results are in line with contemporary reviews from Favier et al. and Hum et al. but are at odds

with the earlier findings from Cornelius et al. who found over 80% of trials in neuropathic pain reported withdrawals due to harm, demonstrating the feasibility of including such information in articles.[56, 58, 79] Reporting this information would permit better evaluation of the impact of harms and would provide useful insight into the tolerability of the intervention to inform patients' and clinicians' treatment decisions. Only reporting numbers that experience at least one event and omitting information on recurrent events loses valuable information that may impact the patients' quality-of-life and affect any cost-effectiveness evaluation but was a common occurrence in this and previous reviews.[56, 58, 79] For example, "*chronic, repeated headaches over an extended duration*" with repeated service use will have an important detrimental impact on patients quality-of-life and associated health and economic costs compared to a single headache or headaches over a short duration but it is not possible to distinguish between these two scenarios when reported as 'at least one event'.[34] Information on severity of events was also often omitted and this could again conceal useful insights. For example, "*there would be a different impact on patients' quality-of-life with mild compared to severe nausea*", and such information could provide valuable insights to tolerable dosing regimens. Whilst information on serious events was typically well reported, more granular information on, for example, the time of likely onset of serious events that can be used to inform patient monitoring plans was omitted. For example, "*the documented risk of suicide and suicidal ideation within the first few weeks of starting an SSRI allows patients and clinicians to remain alert and plan for close monitoring over this period*".[45] Increasing the reporting of these items could facilitate a better understanding of the harm profile.[28] Online appendices and supplementary material provide a means by which to include this important information but there were examples where this was at the detriment of authors providing a balanced benefit harm assessment in the main journal article and others have reported that use of supplementary material did not "*improve the space devoted to safety*".[117]

Conclusions about the harm profile were frequently based on post-hoc statistical tests despite guidelines cautioning against such analysis practices.[34] The output of post-hoc tests under a hypothesis-testing framework are difficult to interpret as a lack of significance does not indicate that the intervention is not without harm and conversely multiple testing without adjustment will increase the number of significant differences due to chance.[120, 121] Whilst null hypothesis-testing is inappropriate in this setting, there is still a need for inferential statistics and formal analysis when analysing emerging harms.[22, 25] Subjective assessments of overwhelming amounts of data could easily lead to potential signals of harm being missed but formal assessments of events regarding stopping for emerging ADRs utilising statistical rules are rare. Calls for more careful consideration of the research question to be addressed and the framework for analysis of emerging harms in RCTs have been made.[25]

The majority of studies relied on presentations of proportions with no information or comment on observed follow-up/exposure time. Recent reviews have found similar findings and highlighted the potential negative consequences of this.[25, 123] This is problematic because if participants within treatment groups are not followed or exposed equally then simple proportions will provide biased summaries. Articles failing to account for this or acknowledging it makes it impossible for readers to assess the potential impact on the results presented.[44] Information on the time events occurred was also rarely reported but a small number did utilise time-to-event methods such as the Cox proportional hazards model and Kaplan-Meier plots to incorporate this information into their analysis. Use of such methods is an improvement but assumptions need to be assessed and they can only account for the time to first event. For example, the Cox proportional hazards model assumes proportional hazards in treatment groups, which is potentially invalid in the presence of drug-event relationships where there is a likely temporal relationship.[124] Methods that could detect time-dependent relationships will be explored in chapters six and seven.

Drug trials routinely screen continuous clinical and biological measures. This and the Patson et al. review found a pervasive practice of dichotomising such outcomes.[123] Dichotomising values into normal/abnormal groups can aid clinical interpretation and decision making e.g. values above a certain threshold indicate clinicians should intervene but it is well established that dichotomisation results in a loss of information and any subsequent statistical testing would suffer a reduction in statistical power in comparison to the equivalent continuous analysis.[112, 125] It can also increase the risk of a result being a false positive.[126] This review found that analysis methods that retain the continuous nature of clinical and biological outcomes such as linear mixed effects models are rarely used.

2.5.2 Differences between public and industry funded studies

Results were also presented by funding source, which served as a proxy measure to identify any potential differences in practice between sectors. Several important differences were identified specifically with regard to the results presented. The most striking being the difference between the analysis population used, with 70% of industry trials using the population who had taken a single dose compared to only 16% of publically funded studies. Industry studies also typically presented more comprehensive information including results on severity, seriousness, relatedness, duration and timing and is in line with the results of the earlier review from Haidich et al.[117] This could reflect a more predominant impact of regulatory guidance on industry trials where the focus is concentrated on gaining regulatory approval. For example, the ICH's E9 Statistical Principles for Clinical Trials adopted by the FDA and EMA states that "*for the overall safety and tolerability assessment, the set of subjects to be summarized is usually defined as those subjects who received at least one dose of the investigational drug. Safety and tolerability variables should be collected as comprehensively as possible from these subjects, including type of adverse event, severity, onset, and duration*."[6] Whereas trials "*initiated and led by researchers from academia, research institutes, or collaborative groups"* typically aim to find "*new therapeutic uses for existing medicines"*, a concept

known as repurposing, thus they may deem, inappropriately in my opinion, that reporting comprehensive harm profiles is of less importance because of existing background knowledge on the harm profile.[127] It should be noted when interpreting these results that defining the funding source was limited as it not possible to ascertain who had influence over analyses and published material, as industry funded studies are often sponsored and undertaken by an academic or NHS institution without access to industry resources such as bespoke safety teams. These differences will be investigated more rigorously in this thesis.

2.5.3 Limitations of this study

In line with the aim of the review to identify 'best' practice the review focused on the top four general medical journals as measured by impact factor, therefore, results are likely to be better than would be expected if RCTs reported across all journals were included. Included articles were from the years 2015–2016 and as such may not reflect the most current practice but ongoing, informal review of the literature since 2016 has failed to identify any notable changes and more recent reviews support these findings.[25, 80, 123] Independent verification of 10% of extracted data would not have removed subjectivity from the process but ongoing discussions between reviewers to clarify queries would have kept this to a minimum. Both phase II and phase III trials were eligible for inclusion in this review. However, it should be noted that whilst they share many common design features the overarching aim of phase II and III trials are different. Whilst definitions in the literature vary, it is widely accepted that phase II trials tend to focus on exploring therapeutic efficacy of an intervention in a narrow population that are closely monitored to provide further information on short term harms and generally have a smaller sample size than phase III trials. Phase III trials focus on confirming therapeutic benefit and real-world effectiveness of an intervention, whilst less focus is perhaps given to harm outcomes, their typically larger sample sizes, longer follow-up and wider inclusion criteria can help further characterise the harm profile.[26, 128, 129] The distinction in aims of the different phases of trials is likely to impact the focus of reporting and could in part explain some of

the variation identified in observed practices. I included both types of trials as they share similar considerations for analysis and presentation. It would have also been challenging to split results by phase II and III as definitions vary and few trials explicitly report their phase of research. Despite these limitations, this review characterises practices of those leading the field of clinical trials and provides some examples of good practice that could be adopted by trialists immediately for the analysis of emerging harm outcomes, as well as highlighting areas for improvements.

2.5.4 Changes for immediate adoption and future research

Whilst examples of good reporting and analysis practices have been observed, there is still much room for improvement. Immediate advantages could be made through several simple changes and a summary of recommendations to improve reporting is provided in table 2.14. Detailed strategies to improve reporting are discussed in chapter 8 section 8.3.

This review also highlighted a number of areas that are in need of further research. These include improving the consistency of reporting important harm outcomes across trials to facilitate comparison and synthesis. This could potentially be through the development of standard harm outcomes for a drug class as suggested by Cornelius et al. and others.[56, 57] Identification and evaluation of appropriate statistical methods for objective analysis of emerging harms would also enable recommendations to be made on analysis practices to undertake.[41]

2.5.5 Conclusions

RCTs provide a valuable source of information to establish the harm profile of interventions in the drug development pathway. However, analysis is frequently inappropriate and articles often provide insufficient and inconsistent information to allow a comprehensive summary of the harm profile to

be established. Inadequate reporting and analysis practices identified in this and contemporary reviews reveal opportunities to establish early harm profiles are likely being missed.

Table 2.14: Key components to include to improve the reporting of emerging harm outcomes in clinical trial publications

Article section	**Information to report**	**Recommendation**
Methods	How data was collected (mode of collection, timing) during the study.	Clearly specify the mode of collection e.g. spontaneous report, questionnaires, clinical and laboratory assessments
		Provide details of the timing of collection for each mode
Methods	Assessment practices to ascertain severity of the event or relatedness to the intervention.	Report who was responsible for assessment practice and how such assessments were made including any criteria (subjective or objective) used to make such decisions – seriousness, attribution to study drug, expectedness
Methods	Planned analysis including final and interim monitoring plans and analysis populations.	Report analysis plans for harm outcomes including both prespecified and emerging outcomes
		Make it clear which analysis was prespecified and which was post-hoc
		Clearly define exposure and specify a suitable analysis population for the analysis of harm outcomes
		Give clear criteria how the results of such analyses will be used to make inferences
Results	Selecting events for inclusion in the journal article.	Provide an informative summary of the harm profile in the main journal article
		Make it clear how events were selected for inclusion in main journal article and where details of other events can be found
		Avoid arbitrary rules to select events to report in main journal article
Results	Information to report	Reduce information loss, for example when count outcomes (such as repeated events within participants) and information on time of onset are available use appropriate statistical methods and consider avoiding dichotomising continuous data.

3. Statistical methods for the analysis of harm outcomes in RCTs: a scoping review

3.1 Introduction

The work presented in the previous chapter revealed that journal articles reporting the primary results of clinical trials predominantly rely on simple approaches such as tables of frequencies and percentages to report results for emerging harm outcomes.[45, 58] Comparative summaries such as the risk difference or risk ratio were rarely presented (6% and 7%, respectively), with more frequent use of statistical comparisons such as the chi-squared test or Fisher's exact test to analyse such data (22%). However, such statistical comparisons are known to be problematic for the analysis of harm outcomes. For example, analysing proportions with the chi-squared test or Fisher's exact test assumes that all participants are followed to the study end and fail to account for withdrawals. They also fail to take into consideration valuable information on the time that events occur or information on recurrent events. Regression and time-to-event approaches were found to be used infrequently but are a potentially useful alternative that can utilise information on exposure time (accounting for withdrawals or censoring), recurrent events and time of occurrence, which is information that is often ignored when analysing harm outcomes.[7] Whilst such approaches offer advantages, these approaches rely on underlying assumptions that may not be appropriate when analysing harm outcomes. For example, the Poisson regression model that can take account of exposure time, assumes that events occur at a constant rate over time, and the Cox proportional hazards model that can take account of censoring and time of the event, relies on the assumption of the hazard between groups being proportional. The assumptions of these models are unlikely to hold for ADRs where there is likely to be a time-dependent relationship.[124]

In light of the problems of the approaches outlined above and the complex nature of harm outcomes, a scoping review was undertaken to investigate what, if any, statistical methods have

been proposed specifically for the analysis of harm outcomes (both prespecified and emerging events) and whether they are used in practice. Knowledge of existing methods could be used to improve awareness and facilitate their use where appropriate. Identification of existing methods for the analysis of harm outcomes in the RCT setting will also reveal any potential gaps in the statistical tools available to researchers. Going forward this information can be used to motivate and inform the development of new methods and support recommendations for best practice.

I have published part of the work presented in this chapter in a BMC Medical Research Methodology paper.[130]

3.2 Aims

The aim of the work in this chapter is to identify and examine research methods that have been developed specifically to analyse harm outcomes in clinical trials that are not prominent in applied trial practice. The specific objectives were:

1) To identify and classify methods that have been specifically developed or adapted to analyse prespecified secondary harm outcomes and non-specific emerging events in clinical trials.
2) To describe the strengths and limitations of identified methods.
3) To identify a subset of methods that would make suitable candidates to be adopted by the clinical trial community.

Whilst it is important that any recommended methods are methodologically robust, to make a real change to practice, methods also need to be easy to implement via accessible software so that it is realistic for applied statisticians to implement alongside existing analysis demands.

3.3 Methods

Methods were identified by undertaking a scoping review of published approaches. The framework proposed by Arksey and O'Malley for conducting scoping reviews was followed.[131] Hence the strategy below was undertaken in an iterative manner to ensure all relevant articles describing methods were identified. A scoping review was proposed rather than a systematic review as this is in line with the aim to uncover all proposed methodology rather than perform a quantitative synthesis and allowed the results of a systematic search to be supplemented with information from alternative sources.

3.3.1 Search strategy

A systematic search of Medline and EMBASE databases via OVID, and the Web of Science and SCOPUS databases was performed in March 2018 to identify all relevant published articles. Web of Science and SCOPUS databases were included to help identify any work published in non-medical journals. Alerts were set up across each database to identify new articles throughout the duration of the thesis. No time restrictions were placed on the search. The search strategy was developed by studying key references in consultation with both experts in the field and experts in review methodology. Full details of the search terms can be found in appendix A3.1. Reference lists of all eligible articles, as well as 'special-issues' of key journals were hand-searched to identify any references that the database search may have missed. A search of the Web of Science database was also undertaken to identify citations of included articles.

Reviewers included my supervisors (VC and OS) and myself (RP). One reviewer (RP) screened titles and abstracts of all articles identified. Full texts of potentially eligible articles were retrieved and further scrutinised for eligibility by one reviewer (RP). All queries regarding eligibility were discussed with at least one other reviewer (VC or OS).

3.3.2 Selection criteria

The inclusion criteria were:

i. Articles that proposed original methods or the original application of existing methods applied to the analysis of harms in phase II-III trials. Specifically methods that aimed to:
    (a) detect signals for potential ADRs for example through statistical tests;
    (b) improve the estimation of between arm statistics but that did not necessarily perform a test e.g. descriptive methods or where a test is implied or easily derived.
ii. Methods that incorporated or utilised a concurrent control group.
iii. Methods suitable for implementation in a parallel group clinical trial design.
iv. Sufficient detail reported to allow the proposed method to be replicated.
v. Peer reviewed.

The exclusion criteria were:

i. Methods where harm outcomes were the primary or co-primary outcome such as risk-benefit methods.
ii. Methods that combined data on harm outcomes across trials, such as meta-analysis methods.
iii. Data-mining and machine-learning methods.
iv. Established methods designed to monitor efficacy outcomes, which could be used to monitor single prespecified harm outcomes, such as the methods of e.g. O'Brien and Fleming, Lan and DeMets.[132, 133]

The search was not restricted to English language articles and foreign language articles were translated where needed. Group sequential methods proposed specifically for harm outcomes were included.

3.3.3 Data extraction

One reviewer (RP) extracted data from eligible articles using a standardised pre-piloted data extraction form (appendix A3.2). Information was collected on methodological characteristics including: whether the method required the event to be prespecified or could be used to screen emerging events; whether it was applied to individual events or aggregate events; data type applicable to e.g. continuous, binary, count, time-to-event; whether any test was performed; what, if any, assumptions were made; if any prior or external information could be incorporated; and what the output included e.g. summary statistic, test-statistic, p-value, plot etc. All articles were discussed with at least one other reviewer and any queries or disagreements clarified with a third reviewer, if necessary.

3.3.4 Analysis

Published results are reported as per the PRISMA extension for scoping reviews and this reporting guideline has been used to ensure clear reporting of the work undertaken in this chapter (completed checklist can be found in appendix A3.3).[134, 135] Each statistical method was appraised in turn and a taxonomy was developed for classification of identified methods. Data analysis was primarily descriptive, and methods are summarised and presented by taxonomy.

The review was registered on PROSPERO a registry for reviews in October 2018.[136]

3.4 Results

3.4.1 Article selection

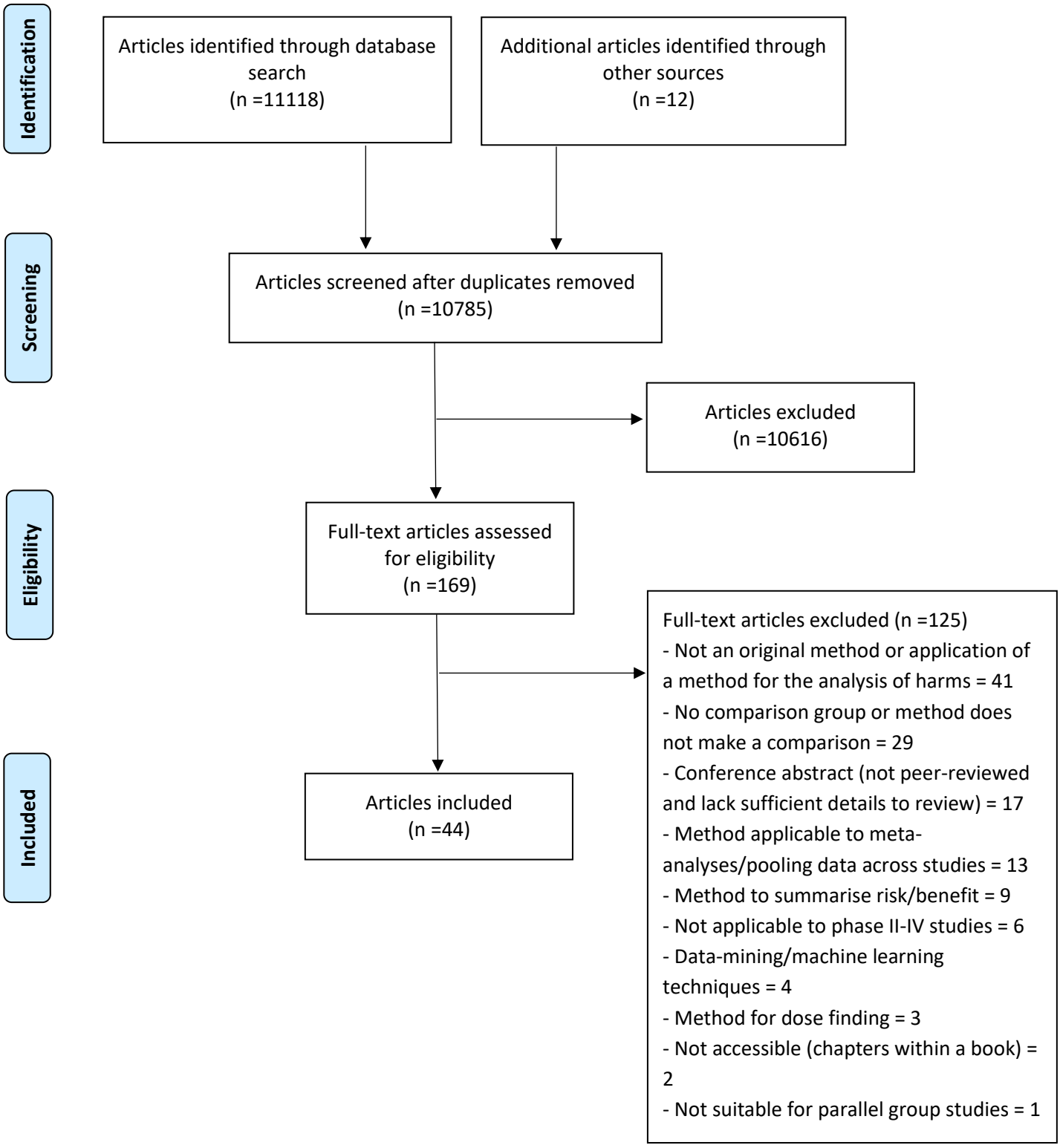
The search identified 11,118 articles. After duplicate articles were removed, 10,773 articles were screened. An additional ten articles were identified from the reference lists of eligible articles and two articles were identified through the search of citations of eligible articles. Review of titles and

abstracts reduced the number of articles for full review to 169. Review of full text articles resulted in a further 125 exclusions. The main reasons for exclusion after full text review were: the method presented was not original or the original application of a method to the analysis of harm outcomes (33%); there was no comparison group or comparison made (23%); articles were published conference abstracts and therefore were not peer-reviewed and/or lacked sufficient detail to undergo a full review (14%). This left 44 eligible articles for inclusion that proposed 73 individual methods (figure 3.1).

3.4.2 Characteristics of articles

Articles were predominantly published by authors working in industry (n=20 (45%)), eight (18%) were published by academic authors and four (9%) were published by authors from the public sector. Eight (18%) articles were from an industry/academic collaboration, two (5%) an academic/public sector collaboration, one (2%) an industry/public sector collaboration and one (2%) from an industry/academic/public sector collaboration.

Figure 3.1: Flow diagram describing the assessment of sources of evidence[††]



[††] Reprinted from *Phillips, R., et al. (2020). "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy" BMC Medical Research Methodology 20(1): 288 under a Creative Commons Attribution 4.0 International License* http://creativecommons.org/licenses/by/4.0/

3.4.3 Taxonomy of statistical methods for the analysis of harm outcomes

Due to the number and variety of methods identified a taxonomy to classify methods was developed. Four broad categories were identified and are described in the table 3.1 and figure 3.2.

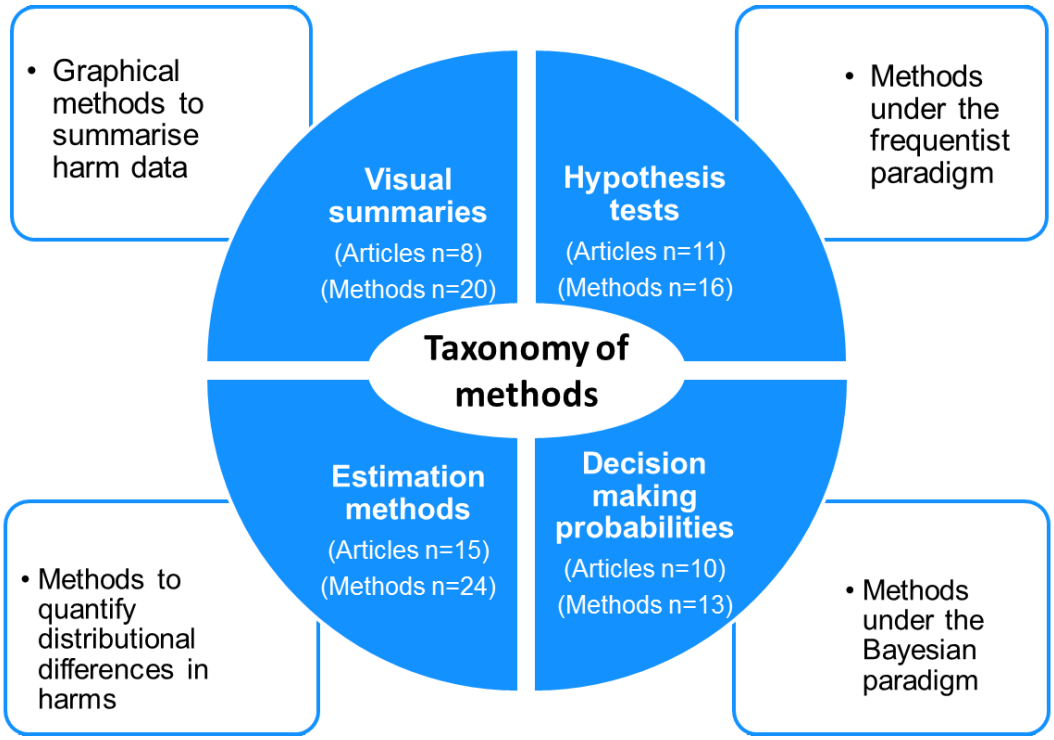
Table 3.1: Taxonomy of methods for the analysis of harm outcomes

	Method category	**Category description**
1	**Visual summary**	Methods that propose purely graphical approaches to view either single or multiple harm outcomes as the principal analysis method.
2	**Hypothesis tests**	Methods under the frequentist paradigm. These methods set up a testable hypothesis and use evidence against the null hypothesis in terms of p-values based on the data observed in the current trial.
3	**Estimation**	Methods that quantify distributional differences in outcomes between treatment groups with summary statistics but without a formal hypothesis test.
4	**Decision making probabilities**	Statistical methods under the Bayesian paradigm. The overarching characteristic of these methods is output of predicted or posterior probabilities regarding the chance of a predefined threshold of risk being exceeded based on the data observed in the current trial and/or any relevant prior knowledge.

All methods were further sub-divided into whether they were for use on prespecified events, which are listed in advance of a trial starting as harm outcomes of interest to follow-up. These would be events already known or suspected to be associated with the intervention, or followed for reasons of interest. The other category was for emerging (not prespecified) events that are reported and collected during the trial and may be unexpected (table 1.2 chapter 1). Further, distinctions were made between methods suitable for analysis of single events or methods that could handle multiple events at a time, where multiple event methods could either produce output for each event or use information on all events to make an overall comparison (multivariate approaches). Methods were also classified as either (group) sequential methods (methods to monitor accumulating data from ongoing studies) or methods for final/one analysis.

The number of articles and methods identified by type is provided in table 3.2. Articles most frequently proposed estimation methods (15 articles proposing 24 methods), followed by hypothesis-testing methods (11 articles proposing 16 methods). Ten articles proposed thirteen methods to provide decision-making probabilities and eight articles proposed 20 visual summaries. The majority of articles developed methods for emerging events (35 articles proposing 61 methods), single outcomes (25 articles proposing 46 methods) and final/one analysis (34 articles proposing 61 methods). Individual article classifications and brief summaries are presented in table 3.3.

Figure 3.2: Taxonomy of methods for the analysis of harm outcomes[‡‡]



[‡‡] Reprinted from *Phillips, R., et al. (2020). "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy" BMC Medical Research Methodology 20(1): 288 under a Creative Commons Attribution 4.0 International License* http://creativecommons.org/licenses/by/4.0/

Table 3.2: Summary level classifications of identified articles and methods

		Taxonomy of methods			
		Visual Articles N=8 [Methods N=20]	**Hypothesis testing** Articles N=11 [Methods N=16]	**Estimation** Articles N=15 [Methods N=24]	**Decision making probabilities** Articles N=10 [Methods N= 13]
Classification		n (%)	n (%)	n (%)	n (%)
Type of event	Prespecified	0 (0) [0 (0)]	5 (55.6) [7 (58.3)]	0 (0) [0 (0)]	4 (44.4) [5 (41.7)]
	Emerging	8 (22.9) [20 (32.8)]	6 (17.1) [9 (14.8)]	15 (42.9) [24 (39.3)]	6 (17.1) [8 (13.1)]
Designed for single outcomes, multiple outcomes or overall harm profile †	Single	2 (8.0) [10 (21.7)]	6 (24.0) [9 (19.6)]	13 (52.0) [22 (47.8)]	4 (16.0) [5 (10.9)]
	Multiple	4 (28.6) [10 (45.5)]	2 (14.3) [2 (9.1)]	2 (14.3) [2 (9.1)]	6 (42.9) [8 (36.4)]
	Single & Multiple	2 (100) [*]	0 (0) [-]	0 (0) [-]	0 (0) [-]
	Overall profile	0 (0) [0 (0)]	3 (100) [5 (100)]	0 (0) [-]	0 (0) [-]
Time of analysis	(Group) sequential¥	0 (0) [0 (0)]	5 (50.0) [6 (50.0)]	0 (0) [0 (0)]	5 (50.0) [6 (50.0)]
	Final/one-analysis	8 (23.5) [20 (32.8)]	6 (17.6) [10 (16.4)]	15 (44.1) [24 (37.5)]	5 (14.7) [7 (11.5)]

*Methods in these articles assigned to single or multiple outcome classification

† Methods suitable for analysis of single events (single) or methods that could handle multiple events at a time (multiple) or a summary of all events (overall profile)

¥ Methods to monitor accumulating data from ongoing studies

Table 3.3: Detailed article classifications[§§]

Authors	Year	Taxonomy [a]	Further classification variables			Brief summary
		V, HT, E, DMP	Prespecified or Emerging	Single or Multiple outcomes	(Group) Sequential (monitoring)	
Amit, Heiberger & Lane[137]	2008	V	Emerging	Single & Multiple	No	Dot plot for emerging AEs, Kaplan-Meier and hazard function for single events and cumulative frequency plots, boxplots and line graphs for continuous outcomes
Chuang-Stein, Le & Chen[138]	2001	V	Emerging	Single	No	Displays two-by-two frequencies graphically for emerging events, histograms and delta plots for continuous outcomes
Chuang-Stein & Xia[139]	2013	V	Emerging	Single & Multiple	No	Bar charts, Venn diagrams and Forest plots for emerging events, risk over time for single events and e-Dish plot for continuous outcomes
Karpefors & Weatherall[140]	2018	V	Emerging	Multiple	No	Tendril plot for emerging events
Southworth[141]	2008	V	Emerging	Single	No	Scatterplot with regression outputs for continuous outcomes
Trost & Freston[142]	2008	V	Emerging	Multiple	No	Vector plots for continuous outcomes, includes 3 outcomes per plot
Zink, Wolfinger & Mann[143]	2013	V	Emerging	Multiple	No	Volcano plot for emerging events
Zink, Marchenko, Sanchez-Kam, Ma & Jiang[42]	2018	V	Emerging	Multiple	No	Heat map for emerging events
Bolland & Whitehead[144]	2000	HT	Prespecified	Single	Yes	Alpha spending function
Fleishman & Parker[145]	2012	HT	Prespecified	Single	Yes	Alpha spending function, adjustment to significance threshold and conditional power
Lieu et al.[146]	2007	HT	Prespecified	Single	Yes	Likelihood ratio test
Liu[147]	2007	HT	Prespecified	Single	No	Non-inferiority test
Shih, Lai, Heyse & Chen[148]	2010	HT	Prespecified	Single	Yes	Likelihood ratio test
Agresti & Klingenberg[149]	2005	HT	Emerging	Overall profile	No	Multivariate likelihood ratio tests for the overall number of events
Bristol & Patel[150]	1990	HT	Emerging	Overall profile	No	Multivariate likelihood ratio test with Markov chains for the overall number of events, incorporating recurrent events
Chuang-Stein, Mohberg & Musselman[151]	1992	HT	Emerging	Overall profile	No	Multivariate test for the overall number of events with chi-squared distribution, incorporating severity and participant acceptability scores

[§§] Reprinted from *Phillips, R., et al. (2020). "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy" BMC Medical Research Methodology 20(1): 288 under a Creative Commons Attribution 4.0 International License* http://creativecommons.org/licenses/by/4.0/

Huang, Zalkikar & Tiwari[152]	2014	HT	Emerging	Single	Yes	Likelihood ratio tests for event rate (i.e. incorporating exposure time) incorporating recurrent events
Mehrotra & Adewale[153]	2012	HT	Emerging	Multiple	No	P-value adjustment
Mehrotra & Heyse[154]	2004	HT	Emerging	Multiple	No	P-value adjustment
Allignol, Beyersmann & Schmoor[155]	2016	E	Emerging	Single	No	Estimates cumulative incidence function in presence of competing risks
Borkowf[156]	2006	E	Emerging	Single	No	Confidence interval for difference in proportions
Evans & Nitsch[43]	2012	E	Emerging	Single	No	Proportions, incidences, odds ratios etc.
Gong, Tong, Strasak & Fang[157]	2014	E	Emerging	Single	No	Non-parametric estimate for mean cumulative number of recurrent events in presence of competing risks
Hengelbrock, Gillhaus, Kloss & Leverkus[158]	2016	E	Emerging	Single	No	Survival based methods to estimate hazard ratios for recurrent events
Lancar, Kramar & Haie-Meder[159]	1995	E	Emerging	Single	No	Non-parametric estimate for prevalence allowing for recurrent events
Leon-Novelo, Zhou, Bekele & Muller[160]	2010	E	Emerging	Multiple	No	Bayesian approach to estimate the probability of severity grading of events in treatment and control groups separately
Liu, Wang, Liu & Snavely[161]	2006	E	Emerging	Single	No	Confidence interval for difference in exposure adjusted incidence rates
Nishikawa, Tango & Ogawa[162]	2006	E	Emerging	Single	No	Estimates the cumulative incidence function in presence of competing risks and conditional estimate for recurrent events
O'Gorman, Woolson & Jones[163]	1994	E	Emerging	Single	No	Confidence intervals for difference in proportion
Rosenkranz[164]	2006	E	Emerging	Single	No	Survival based method to estimate dependence between event time and discontinuation time
Siddiqui[44]	2009	E	Emerging	Single	No	Non-parametric estimate for the cumulative mean number of events allowing for recurrent events
Sogliero-Gilbert, Ting, & Zubkoff[165]	1991	E	Emerging	Multiple	No	A score to indicate abnormal laboratory values
Wang & Quartey[166]	2012	E	Emerging	Single	No	Non-parametric estimate for mean cumulative event duration allowing for recurrent events
Wang & Quartey[167]	2013	E	Emerging	Single	No	Semi-parametric estimate for mean cumulative event duration allowing for recurrent events

Berry[168]	1989	DMP	Prespecified	Single	Yes	Bayesian approach to estimate the posterior probability that event rate or incidence rate (incorporating exposure time) is greater in the treatment group compared to control group
French, Thomas & Wang[169]	2012	DMP	Prespecified	Single	Yes	Bayesian logit model and a piecewise exponential model to give posterior probabilities that a predefined risk difference threshold is exceeded
Yao, Zhu, Jiang & Xia[170]	2013	DMP	Prespecified	Single	Yes	Bayesian beta-binomial model to give posterior probability that a predefined risk difference threshold is exceeded
Zhu, Yao, Xia & Jiang[171]	2016	DMP	Prespecified	Single	Yes	Bayesian Gamma-Poisson model to give posterior probability that a predefined risk difference (incorporating exposure time) threshold is exceeded
Berry & Berry[172]	2004	DMP	Emerging	Multiple	No	Bayesian hierarchical logit model to give posterior probability that event rate greater in treatment compared to control group
Chen, Zhao, Qin & Chen[173]	2013	DMP	Emerging	Multiple	Yes	Bayesian hierarchical logit model to give posterior probability that event rate greater in treatment compared to control group for interim analysis
Gould[174]	2008	DMP	Emerging	Multiple	No	Bayesian approach to estimate the posterior probability that events in treatment group produced by a larger process than events in control group
Gould[175]	2013	DMP	Emerging	Multiple	No	Bayesian approach to estimate the posterior probability that events in treatment group produced by a larger process than events in control group accounting for exposure time
McEvoy, Nandy & Tiwari[176]	2013	DMP	Emerging	Multiple	No	Bayesian multivariate approach to give posterior probability of difference in event rates based on indicator functions
Xia, Ma & Carlin[177]	2011	DMP	Emerging	Multiple	No	Bayesian hierarchical logit and log-linear (incorporating exposure time) models to give the posterior probability that the event rate in the treatment is greater than the event rate in the control group

[a] **V – Visual, HT – Hypothesis Testing, E – Estimation, DMP – Decision-Making Probabilities**

3.4.4 Summary of methods by taxonomy

1. Visual summaries – emerging events

The review identified eight articles published between 2001 and 2018 that proposed twenty unique methods to visually summarise harm data, including binary AEs and, continuous laboratory (e.g. blood tests, culture data) and vital signs (e.g. temperature, blood pressure, electrocardiograms) data (table 3.4).[42, 137-143] The majority of the proposed plots were designed to display summary measures of harm data (n=14) (e.g. figures 5.11, 5.12 or 5.13 in chapter 5) and the remaining plots displayed individual participant data (n=6) (e.g. figure 5.18 in chapter 5). None of the plots required the event to be formally prespecified but for some plots, prespecification of the event would be deemed good practice e.g. figure 5.15 chapter 5. Eight of the plots were designed to display multiple binary events (e.g. figures 5.11 or 5.12 in chapter 5).[143, 178] The remaining plots were proposed to focus on a single event, three of which proposed plots to display time-to-event data and nine proposed plots to analyse emerging, individual, continuous harm outcomes such as laboratory or vital signs data (e.g. figures 5.19, 5.20 or 5.21 in chapter 5).

Use in the applied literature

There is a variety of graphical options available for monitoring and analysing harm outcomes, but use in the published literature is rare. The systematic review described in chapter two found only 12% of RCT reports published in high impact journals included a graphical display of data on harm outcomes. A search of citations of the graphical options identified in this chapter using the Web of Science database found only one paper had implemented any of the aforementioned methods for the analysis of harm outcomes in a RCT.[179] However, trial publications seem to rarely cite methods for visual analysis as many are now commonly recognised and therefore it is impossible to gauge use of these particular methods.[180] In addition, some of the identified visualisations would be better suited to the monitoring of ongoing trials or surveillance in the post-marketing setting, which is

beyond the remit of this thesis, than final analysis (e.g. matrix of scatterplots figure 5.18 of chapter 5) and given such activities are largely performed confidentially, 'in-house', use is difficult to assess.

Software for implementation

Producing effective visualisations has become easier with advances in statistical software packages and development of user written commands in packages such as Stata and SAS and the widespread adoption of the freely available software package R. Several of the plots identified in this review are reproducible with standard commands built into software such as the stacked bar chart or line graphs. However, many require bespoke user written code/commands such as the tendril plot, the authors of which have developed a package to produce it in R.[140]

Summary

Visual summaries can be a powerful means to display complex data on harm outcomes to a range of audiences and wider use of graphical approaches for the analysis of harm outcomes from RCTs has been advocated in the literature.[139] Guidelines recommend plots as a space efficient way to display time-to-event and repeated measures on harms to help detect differences in event rates between treatment groups.[28, 181] Plots can also be used to identify signals for potential ADRs from the body of emerging harm outcomes. Visualisation are explored in detail in chapter 5 and recommendations for use are provided.

Table 3.4: Summary of visual approaches to summarise harm outcomes in phase II/III RCTs

Outcome	Data type	Plot	Reference	Brief Description
Emerging adverse events (multiple)	Binary	Volcano	Zink, Wolfinger & Mann 2013. Xia 2011 first proposed this method for systematic reviews but was not eligible for inclusion in this review.[143, 182]	Summarises and compares the incidence of each event reported by treatment group
	Binary	Dot	Amit, Heiberger & Lane, 2008. Cooper, 2008 also proposed but for pooled trials therefore not eligible for inclusion in this review.[137, 183]	Provides an absolute and relative measure compared across treatment groups for each event reported
	Time-to-event	Tendril	Karpefors & Weatherall, 2018.[140]	Provides a summary of time-to-event data by treatment group for each event reported
	Binary	Heat map	Zink, Marchenko, Sanchez-Kam, Ma & Jiang[42]	Visualises treatment effects for each event reported
	Binary	Bar chart	Chuang-Stein & Xia, 2013.[139]	Displays the frequency of events
	Binary	Venn diagram	Chuang-Stein & Xia, 2013.[139]	Presents frequencies highlighting the prevalence of overlapping events
	Binary	Two-by-two frequencies	Chuang-Stein, Le & Chen, 2001.[138]	Displays two-by-two frequencies graphically
	Binary	Forest plot	Chuang-Stein & Xia, 2013. Refers to Lewis & Clarke, 2001, which is not eligible in its own right as it is not specific to AEs.[139, 184]	Plots a relative measure compared across treatment groups for each event reported within a group such as body system
Emerging adverse events (single)	Time-to-event	Kaplan-Meier	Amit, Heiberger & Lane, 2008.[137]	Summarises time-to-event data highlighting absolute differences over time
	Time-to-event	Hazard function	Amit, Heiberger & Lane, 2008.[137]	Summarises time-to-event data highlighting the time at which differences emerge
	Time-to-event	Risk over time	Chuang-Stein & Xia, 2013.[139]	Summarises incidence of an event over time
Laboratory & Vital Signs	Continuous	Cumulative frequency plots/empirical cumulative distribution function	Amit, Heiberger & Lane, 2008.[137]	Provides a summary of the distribution e.g. change for individual participants over time
		Boxplots	Amit, Heiberger & Lane, 2008. Cooper, 2008 also proposed but for pooled trials therefore not eligible for inclusion in this review.[137, 183]	Provides a summary of the distribution e.g. change at specific time points
		Line graphs	Amit, Heiberger & Lane, 2008.[137]	Provides a summary of change at specific time points

		Histograms	Chuang-Stein, Le & Chen, 2001.[138]	Provides a summary of the distribution e.g. change at specific time points
		Scatter plots	Amit, Heiberger & Lane, 2008. Cooper, 2008 also proposed but for pooled trials therefore not eligible for inclusion in this review.[137, 183]	Provides a summary of change for individual participants over time
		Scatter plot with regression	Southworth, 2008.[141]	Provides a summary of change for individual participants over time highlighting any outliers
		Delta	Chuang-Stein, Le & Chen, 2001.[138]	Displays individual participant changes
		Vector plots	Trost & Freston, 2008.[142]	Simultaneously displays individual participant changes across three laboratory values
		e-Dish	Chuang-Stein & Xia, 2013.[139]	Scatter plot of peak bilirubin versus peak serum ALT or AST levels for individual participants to identify drug induced serious hepatotoxicity

2. Hypothesis tests – prespecified outcomes

Under a traditional frequentist statistical approach, a research question can be translated into a testable hypothesis. In the context of harm outcomes, this requires a prespecified event of interest with a predefined difference that if exceeded would indicate harm.

Five articles published between 2000 and 2012 present seven methods to analyse prespecified harm outcomes under a hypothesis-testing framework, these were predominantly designed to monitor participants in ongoing studies using sequential testing (appendix A3.4).[144, 145, 147, 148, 185] Methods specifically designed and promoted for sequentially monitoring prespecified harm outcomes included: two methods that incorporated an alpha-spending function (as originally proposed for efficacy outcomes)[133]; two that performed likelihood ratio tests; one that used conditional power to monitor the futility of establishing safety; and one proposed an arbitrary reduction in the traditional significance threshold when sequentially monitoring a harm outcome.[144, 145, 148, 185] In addition, one method not based on sequential testing proposed a non-inferiority hypothesis test approach for the final analysis of a prespecified harm outcome, requiring pre-specification of an acceptable 'safety margin' that that the intervention group does not exceed. This approach is the same as the traditional non-inferiority design used for efficacy outcomes whereby the confidence interval is used to appraise the evidence against the null hypothesis of excessive harm. The authors suggest that this approach should also be used to power the trial alongside the primary efficacy outcome.[147]

Application in applied literature

It is difficult to gauge use of these methods as outputs of sequential monitoring are reported in interim analyses and such reports are typically confidential and not accessible. Use in publications of final reports can be assessed by citation of the method although unless a trial is stopped early this will likely go unreported and thus be underreported. Method citations indicate only one of the

methods has been applied in a RCT setting to monitor harm outcomes (six citations according to Web of Science citations).[144] Results of chapter two also indicate that it is rare (2.7%) for trials to include a formal stopping rule for harm and in those studies that did include a stopping rule for harm, none used any of the identified methods (table 2.5 chapter 2).

Software for implementation

Implementation of the sequential methods discussed here are theoretically complex but should be implementable in standard statistical software and in addition, two authors provide reference to user written packages.[144, 148]

Summary

Sequential methods are useful when continuous monitoring of a prespecified event is important and they provide easy to interpret output. With careful preparation at the design stage stopping boundaries can be plotted in advance to better understand the scenarios that may lead to the trial being stopped. Unlike common sequential designs for efficacy outcomes, these approaches do not dictate sample size, which remains fixed by design for the primary efficacy comparison. They are also proposed to be implemented as one-sided tests indicating a trial should be stopped if evidence to date indicates the intervention group are experiencing excessive harm. However, they can be implemented as two-sided rules if trialists also wish to recommend stopping if the intervention is shown to be superior to the control. These methods have been included here for completeness but since monitoring of ongoing trials is not the focus of this thesis, only a short overview is provided. Further information on sequential approaches can be found in the literature, including an overview of methods by Whitehead.[186] The non-inferiority approach is a familiar method in the clinical trial setting for efficacy outcomes but is proposed here for a prespecified harm outcome. The chosen non-inferiority margin needs to have a clear rationale, this is often challenging in the efficacy setting

with one primary outcome and is therefore likely to prove as hard, if not more so, in the harm setting. In addition, the authors propose that the analysis should be powered alongside the primary efficacy outcome and as such are suggesting, without explicitly stating it, that the design requires co-primary outcomes, one to demonstrate efficacy and one to demonstrate an acceptable level of harm.

3. Hypothesis tests – emerging

Hypothesis tests can also be applied to the analysis of emerging outcomes. Six articles published between 1990 and 2014 suggest nine methods to perform hypothesis tests to analyse emerging data on harm outcomes, two of which were suitable for evaluating multiple events simultaneously, two were for single events and five were suitable for the overall profile (i.e. analysis of multiple events summarised into one outcome for comparison) (appendix A3.5).[149-154] All of the methods were designed for the final analysis, with one method incorporating an alpha-spending function, thus allowing the method to also be used to monitor ongoing studies. Methods are suggested for both binary and time-to-event data with several accounting for recurrent events.

Two methods proposed a p-value adjustment to account for multiple hypothesis tests across multiple outcomes to reduce the false discovery rate (FDR).[153, 154] These methods propose a two-stage approach that accounts for multiple hypothesis tests by adjusting the p-values based on the FDR.[154, 187] The method utilises coding structure where individual harm events are grouped within the system organ class or body system they occur. The FDR approach aims to control the expected proportion of incorrectly rejected null hypotheses i.e. the type I error. It flags events that achieve statistical significance after the FDR adjustment. Problems were identified with the resampling method used in the original FDR approach but an update proposed a modification so that resampling

is not necessary. The updated method allows better control of the FDR without compromising power and therefore supersedes the original method. The FDR method is a multi-stage approach to adjust p-values when analysing the entire body of emerging harm data, and allows a clear interpretation to aid the identification of events as potential ADRs for further monitoring.

One article proposed two likelihood ratio statistics to test for differences between treatment groups when incorporating time-to-event and recurrent event data for a single event.[152] The first method compares the time to first event using a likelihood ratio test on the relative risk, where the relative risk is the number of events over the sum of event time. The second method incorporates information on recurrent events where the number of events is assumed to follow a Poisson distribution and a likelihood ratio test is undertaken on the relative risk of all occurrences of the event. For multiple looks at the data across time, the authors propose an increasing or decreasing alpha-spending function to control the family-wise type I error.

Three articles adopted multivariate approaches to undertake global likelihood ratio tests to detect differences in the overall harm profile (i.e. using information on all events to make an overall comparison).[149-151] The first article utilised Markov chains of order one to compare recurrent events between treatment groups allowing the probability of an event at the current visit to be conditional on the presence or absence of an event at the preceding visit. Rejecting the null hypothesis allows the conclusion that the vector of probabilities between treatment groups are not equal. Transition probabilities (probability of event (or absence) given state in previous visit) and marginal probabilities (probability of event) can be calculated for each visit by treatment group to aid identification of differences. The second multivariate method utilises the structure of event classifications into body systems and accounts for severity of events by assigning weights to each severity grade of an event.[151] Participants can only be assigned to one grade per body system so

unlike the previous method does not account for recurrent events but does account for severity, which the first does not. The observed mean score incorporating weights is calculated for each body system. A vector of expected values of mean scores for each body system and the covariance vector is calculated and a multivariate test is used to compare vectors of expected mean scores between treatment groups, where the test statistic has an asymptotic chi-squared distribution under the null hypothesis. The third article proposed several global tests to compare equality of vectors of events between treatment groups for both frequencies and counts of events.[149] The first method uses a likelihood ratio test to test the marginal distributions of event counts, which are assumed to have an independent multinomial distribution. The likelihood ratio test uses a logit model to test for equality of two vectors for the marginal distributions. The second method is the same as that of Chuang-Stein et al., which proposed a test statistic using the variance and covariance of the marginal proportions. They also propose generalised estimating equations as a means to incorporate covariates using a Wald test statistic instead of a likelihood based approach. The third method tests the null hypothesis that the joint distributions are identical for the two treatment groups. It compares the overall proportion of events in each group by applying a likelihood ratio test using the exact permutation distribution. This considers all possible ways of splitting the subjects in each treatment group. The fourth method compares marginal distributions whilst modelling the joint distribution using a logistic normal random intercept model.

Application in applied literature

The FDR approach is the only method found to have been used in the applied literature for analysing harm data from RCTs identified through a Web of Science search of citations.[188, 189]

Software for implementation

The authors of the FDR method indicate that it is implementable in SAS.[153, 190] The likelihood ratio tests proposed by Huang et al. do not come with specific software recommendations but the methods should be implementable in most standard statistical software.[152] Agresti and Klingenberg who proposed several multivariate approaches state algorithms for implementation of these methods are available upon request but warn that each method becomes computationally more difficult as the number of events and/or sample size increases.[149]

Summary

Each of these methods provides clear output for interpretation i.e. reject or accept the null hypothesis to flag between-group differences for an individual event or the overall harm profile. However, the hypothesis test approach traditionally set-up for efficacy outcomes can be problematic for analysis of individual harm outcomes, specifically in the context of emerging events and could explain limited use in the literature. Any analysis in this context is data driven i.e. the events could not be prespecified. Therefore, the studies have not been powered for such analysis and run the risk of incorrectly concluding that the treatment is 'safe' due to insufficient power for the analysis undertaken. Therefore, despite these methods not requiring prespecification of events it seems reasonable that they are only used for prespecified events to prevent post-hoc data driven hypotheses testing. There are also likely to be multiple different emerging events (sometimes exceeding the number of trial participants) and under the frequentist paradigm, this raises the issue of multiple testing that could result in an event being incorrectly signalled as an ADR due to a chance difference. The FDR approach proposes one way of negating this and could be easily adopted into practice. Issues of multiplicity do not typically affect multivariate approaches as they require only one test to identify overall differences in harm profiles but the appropriateness of such an overall approach needs to be given careful consideration as it could mask important differences at the event

level. Whilst they could identify important differences in participant burden such analyses should still be accompanied by more event specific comparisons. Caution in interpretation is needed for any hypothesis test approach, with researchers remaining mindful of the limitations of null hypothesis testing for harm outcomes in trials designed around primary efficacy outcomes. Formal comparisons between the different likelihood approaches and different multivariate approaches would be helpful to inform recommendations on use.

4. Estimation – emerging

Data can be used to quantify distributional differences for emerging harms rather than formally test the data. Fifteen articles proposing 24 methods to estimate between group statistics for emerging events were published between 1991 and 2016 (appendix A3.6).[43, 44, 155-160, 162-167, 191] These estimates incorporate a range of characteristics collected on harm outcomes, outputting estimates such as point estimates for incidence, measures of precision, or estimates of the probability. They rely on subjective comparisons of differences to identify potential treatment effects.

Point estimates such as the risk difference, risk ratio and odds ratio to compare treatment groups with corresponding confidence intervals are simple approaches for the analysis of binary harm outcomes.[43, 45] Two articles proposed alternative means to estimate confidence intervals for differences in proportions for harm outcomes and one article proposed methods to estimate confidence intervals for exposure adjusted incidence rates that follow a Poisson distribution.[156, 163, 191]

Eight articles provided methods to calculate estimates that take into account characteristics of harm outcomes, such as recurrent events, exposure-time, time-to-event information, and duration, which

if taken into consideration can help develop a clearer profile of the overall burden of harm.[44, 155, 157-159, 162, 164, 166, 167] Methods such as the mean cumulative function and mean cumulative duration propose non-parametric approaches to estimate the cumulative mean number of events and the cumulative mean duration of events at different time points, both allowing for recurrent events. Time-to-event methods such as the counting process model originally proposed by Andersen-Gill and the conditional Cox model originally proposed by Prentice-Williams-Peterson can also be used to estimate treatment effects accounting for recurrent events.[158, 192, 193] The Aalen-Johansen cumulative incidence function and parametric time-to-event models such as that of Fine and Grey can be used to estimate the probability of an event in the presence of competing risks.[155, 194] Several of these methods incorporated plots that can highlight when differences between treatment groups start to emerge, which would otherwise be masked by single point estimates. In addition, a Bayesian approach was developed to estimate the probability of experiencing different severity grades of each event, accounting for the structure of events within body systems.[160] Only one article developed an approach for the analysis of continuous laboratory values, with the aim of flagging abnormalities if values were found to be outside of the normal reference ranges.[165]

Application in applied literature

There is some evidence that the cumulative frequency and duration methods have been used in practice.[195, 196] Whilst there is not widespread evidence of time-to-event based techniques being used for harm outcomes, the paper by Proctor et al. provides a useful example on the implementation of such approaches.[197]

Software for implementation

Authors indicate that standard commands in SAS can be used to estimate the mean cumulative function and duration. Whilst no specific software is recommended for calculation of confidence intervals or time-to-event estimates, including estimates from competing risk models, these estimates should be implementable in standard statistical software.

Summary

There is a variety of methods proposed to estimate different characteristics of the emerging harm profile, with many ensuring an efficient use of the data collected. These vary in complexity and ease of use. In their simplest form, they avoid the problems of multiple testing and insufficient power but do rely on a conclusion being made from a subjective comparison and when presented with a confidence interval there is a strong temptation to interpret this as a hypothesis test based on whether the interval crosses the summary statistic value of no difference. Methods that provide estimates of between group differences with a measure of uncertainty could be easily incorporated into practice, for example, differential follow-up and recurrent events can be accounted for via the incidence rate ratio. Estimates of cumulative frequency and duration with accompanying plots are easy to implement and interpret and could be easily incorporated into analysis of harm outcomes to provide a concise summary of the overall harm profile that account for recurrent events. Several existing time-to-event based techniques have been recommended for the analysis of individual events. These take into account the time of occurrence of the event with some allowing for recurrent events and others account for competing events such as death and/or withdrawal. Time-to-event methods for the first occurrence of an event will be investigated further in chapters five to seven.

5. Decision making probabilities – prespecified outcomes

As well as using existing knowledge to prespecify harm outcomes for monitoring under a frequentist framework, prior or accumulating information about these outcomes can be formally incorporated into the analyses of ongoing studies under a Bayesian framework. Such analyses can provide evidence to aid decisions about the conduct of ongoing trials or future trials based on the emerging harm profile.

The review identified four articles suggesting five unique Bayesian approaches to monitor prespecified harm outcomes (appendix A3.7).[168-171] The first paper was published in 1989 by Berry and was the forerunner for these methods, giving the general principles for monitoring prespecified harm outcomes under a Bayesian framework for binary or a time-to-event outcomes. No further research was published in this area until 2012; the last paper was published in 2016. Methods include the beta-binomial and gamma-Poisson models. The beta-binomial model assumes the number of events in each treatment group follows a binomial distribution with beta priors for the event rate.[170] The parameter values for each beta prior are based on the number of events and the total participants observed per treatment group from historical data, which could be based on a single previous trial or an aggregate summary of data from multiple trials. At each analysis, these distributions are updated with the number of events observed and number of participants enrolled up until that point to give posterior beta distributions. The gamma-Poisson method follows similar principles, assuming that the event rates follow a Poisson distribution with gamma priors for the event rates per unit time for each treatment group.[171] The parameter values for each gamma prior are based on the number of events and total participant exposure time taken from historical data. At each analysis, these distributions are updated with the number of events observed and total participant exposure time up until that point to give posterior gamma distributions. The posterior distributions are then used to calculate posterior probabilities that some predefined threshold of risk

difference has been crossed and are used to guide the decision whether to continue with the study based on the harm outcome.

An additional article proposes two alternative methods. One method parametrises the event rate e.g. proportion of participants with an event with the logit model and the second method parametrises the event rate per unit time, with a piecewise exponential model.[169] Normal prior distributions are specified for the logit of the event rate in the control group and the difference in the logit event rate between treatment groups. The piecewise exponential model splits the study period into distinct intervals, allowing each period to have its own constant baseline hazard therefore allowing a non-constant hazard over time. The treatment effect (hazard ratio) is calculated by fitting a proportional hazards model that assume that the baseline hazard is a piecewise constant. The baseline hazards are assumed to have gamma priors with parameter values based on the number of events and total participant exposure time from historical data. The hazard ratio is assumed to have a normal prior distribution. Each of these methods was designed for use in interim analyses to monitor ongoing studies but could be used for the final analysis without modification. They could be implemented for continuous monitoring (i.e. after each observed event) or in a group sequential manner after several events have occurred. These methods require a prespecified event, an assumption about the prior distribution of this event, a 'tolerable risk difference' and an 'upper threshold probability' to be set at the outset of the trial.[170] At each analysis, the probability that the 'tolerable risk difference' threshold is crossed is calculated and if the predetermined 'probability threshold' is crossed then the data indicate a predefined unacceptable harmful effect. The tolerable risk difference should be chosen based on clinical and participant input, alternatively simulations exploring various scenarios and possible outcomes could be used as is common in the efficacy setting but the practicalities of this in the harm setting are likely unfeasible.

Application in applied literature

These approaches do not appear to have been used in the applied literature for analysis of harm outcomes in the RCT setting, with citations in Web of Science limited to single arm or historical control group studies.[198, 199]

Software for implementation

Each of these methods can be implemented in the specialist Bayesian software OpenBUGS (or WinBUGS) with specific coding details provided in the paper by French et al.[169] Recent developments have also made implementation in standard statistical software such as Stata possible. Going forward this should make wider use of these methods in the applied literature more practical.

Summary

The output from these methods is intuitive and has a clear interpretation. Both the gamma-Poisson and piecewise exponential models can incorporate information on time-to event. The piecewise exponential model overcomes the potentially problematic assumption of constant event rate by splitting the study period up and fitting separate models to each. However, how the study period is split across time is arbitrary and requires selection based on judgement. Alternative parametric models that account for non-constant hazards or disproportionality between treatment groups may offer a better solution for the harm setting where such characteristics would be indicative of a temporal relationship and potential adverse reaction, and will be explored further in chapters six and seven.[124]

For each of these methods the decision regarding whether to raise a signal regarding a prespecified outcome is based on posterior probabilities about a predefined difference. Scenarios that would

lead to the decision to terminate the study can be explored in advance and discussed with the research team prior to implementation. Sensitivity of the results to the prior assumptions can also be undertaken. These methods could also be adapted for analysis of emerging outcomes using non-informative priors and such an approach will be explored in chapters six and seven.

6. Decision making probabilities – emerging outcomes

Trials produce an abundance of emerging data on harms and under a traditional framework, performing multiple hypothesis tests runs the risk of raising a false signal of an ADR due to a chance imbalance. Bayesian methods do not suffer from such problems and as such have been proposed as a potential solution for the analysis of emerging harm data.

Six articles published between 2004 and 2013 proposed eight unique Bayesian methods to analyse the body of emerging data on harm outcomes (appendix A3.8).[172-177] Under a Bayesian framework the issues around multiplicity for different outcomes can be overcome by utilising the structure of coding systems such as MedDRA, allowing biologically similar events (i.e. those classified within the same body-system) to borrow strength from each other under the assumption of exchangeable parameters that allows 'similar' events to share a common prior.[200, 201] Five of the methods utilise a Bayesian framework to borrow strength from medically similar events. Berry and Berry were the first, proposing a Bayesian three-level random effects model.[172] The method allows events classified within the same body system to be more alike and information can be borrowed both within and across systems. For example, within a body system a large difference for an event amongst events with much smaller differences will be shrunk toward zero. This work was extended to incorporate person-time adjusted incidence rates using a Poisson model and to allow sequential monitoring.[173, 177]

Two articles developed alternative approaches that follow similar principles. The first proposed a multivariate method for summarising the overall harm profile where individual event rates are assumed to follow a binomial distribution.[176] Each event is assigned an indicator function to indicate if the risk of an event is different between treatment and control group. The indicator function is assumed to follow a binomial distribution with beta prior. Across all events a vector of indicators of difference is produced. Dependencies amongst related events is incorporated using a binary Markov Random Field, known as the Ising prior. The prior is set for the indicator functions and not the difference in event rates. The posterior distribution follows a beta distribution and can be sampled to give posterior probabilities of, for example, a difference in event rates for a specific event of interest.

The second article proposed two methods that do not look at the probability of specific estimates but instead look at the distributions that produce events. These methods still output a posterior probability about emerging events to indicate a signal for a potential ADR. The first method assumes that the event rate follows a binomial distribution with a beta prior in the control group.[174] It then examines whether the intervention group event rate is generated by the same process or a larger process with its own beta parameters. The posterior distribution for the treatment group event rate is a mixture distribution. The mixture distribution is the same as for the control group if there is no difference between treatment and control groups or larger if there is a difference. The mixture distribution contains an indicator function to allow for a potential difference. The indicator follows a Bernoulli distribution with a beta prior distribution therefore the posterior is a beta-binomial density. The probability that the indicator function equals zero (i.e. treatment and control group rate generated by the same process) in the mixture posterior distribution is the parameter of interest. A low probability that the indicator function equals zero would indicate that events in the treatment group are unlikely to have been generated by the same process as the control group. A similar

method was also developed that instead assumed events followed a Poisson distribution.[175] The assumption is that if the trial is large enough and event rate low enough that events can arise from a Poisson distribution instead of binomial distribution rather than it allows for exposure time. The priors are assumed to follow a gamma distribution, which when combined with the Poisson likelihood gives a gamma posterior distribution for event rate in the control group. Again, this method examines whether the intervention group event rate is generated by the same or a larger process.

Application in applied literature

There has again been limited uptake of these methods in the applied literature, and the only citations are limited to research articles. For example, one article compares the method of Berry and Berry to data-mining techniques in a RCT setting.[202]

Software for implementation

None of the methods initially provided code for implementation but each should be implementable in specialist software such as OpenBUGS (or WinBUGS) and a follow-up paper in 2018 provided R code for the Bayesian screening method proposed by Gould.[203]

Summary

Each of the methods output Bayesian posterior probabilities to guide the decision whether to raise a signal for events to undergo further investigation or future monitoring. Many are proposed as a solution to the issue of multiple hypothesis tests by utilising the structure of the data allowing borrowing of information from related events. Whilst these methods could be useful for analysing emerging events, they are reliant on the assumption that intervention effects are on the 'system'

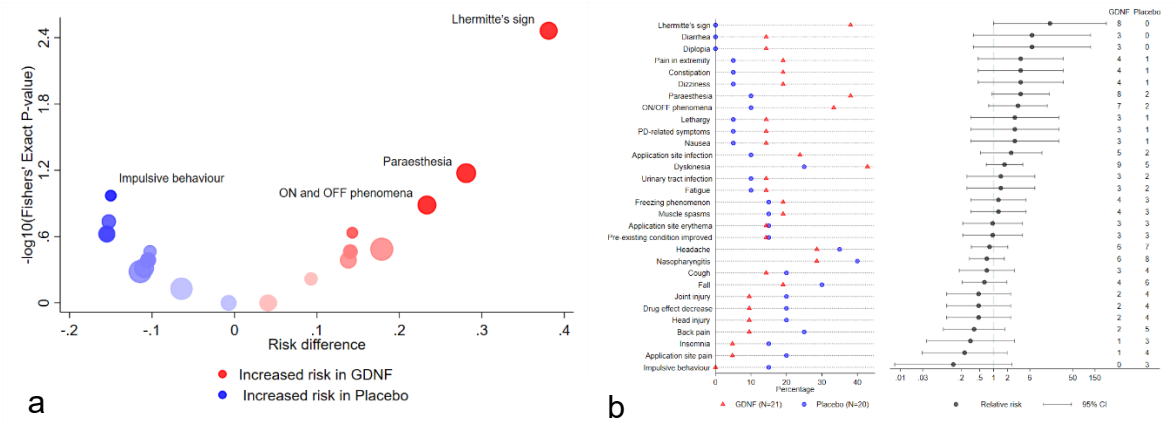
level and are unlikely to detect harms that occur in isolation. They also rely on correct and consistent classification of events within trials. In addition, at present implementation in standard software is not straightforward and as such these methods are not easily implemented by applied statisticians.

3.4.5 Software development for selected visualisations

Two visualisations identified in this review introduced in section 3.4.4 (1), the dot plot and volcano plot, were selected for further exploration as both were felt to offer potential benefit for the presentation of data on harm outcomes from RCTs in journal articles. Both display information on multiple events simultaneously, offer flexibility on the chosen metric to present being able to display point estimates for binary, count or time-to-event data, and were assessed as being intuitive to interpret as well as visually appealing. They also allowed presentation of differing characteristics of the data allowing for different emphasis and the potential to affect the inferences drawn. The volcano plot displays a between arm (absolute or relative) summary statistic on the x-axis against the –log10 p-value, from a test of the researcher's choice, on the y-axis. Under the null hypothesis of no difference, the plot would display a symmetrical shape around the null value of no difference; detection of asymmetry in the plot can be used to identify events for further investigation. In contrast, the dot plot explicitly presents, for each event, both an absolute and relative measure (with accompanying 95% confidence intervals). To enable wider use of the volcano and dot plots for presentation of data on the body of emerging harm outcomes, in collaboration with a colleague, easy to use Stata commands (`aedot` and `aevolcano`) for production of these plots were developed and made available to the Stata community.[204, 205] An article demonstrating their implementation and potential utility has also been published.[206] Figure 3.3 displays examples of each of these plots using published summary level data on emerging harms from Whone et al.[178] To account for the multiple tests undertaken to produce the volcano plot, the authors proposed implementing a multiplicity adjustment, specifying the use of the FDR method as proposed by

Mehrotra and Adewale and described in section 3.4.4 (3).[153] To replicate the original plot, code for implementation of the FDR method was incorporated into the Stata code for the volcano plot and also made available as a stand-alone command.[190]

Figure 3.3: Volcano and dot plot for emerging harms experienced by at least three participants in either treatment group from summary results presented in Whone et al.[178] ***



a. Volcano plot: The x-axis represents the difference in proportions of participants experiencing each event between the treatment groups (intervention – placebo). The y-axis represents the p-value from a Fisher's exact test on the -log10 scale. The centre of the bubble indicates the coordinates for each event. The size of the bubble is proportional to the total number of events for both treatment groups combined. Colour is used to indicate direction of treatment effect with red indicating greater risk in the intervention group and blue indicating greater risk in the placebo group. Colour saturation corresponds to the -log10(p-value) for each event. **b. Dot plot:** The left side of the figure displays the percentage of participants experiencing an event (labelled on the y-axis) in the intervention group with a red triangle and placebo group with a blue circle. The central panel displays the relative risk and corresponding 95% confidence interval on the log10 scale. On the far right is a data table including the number of participants with an event and the total number of events for each treatment group.

*** *Reprinted with minor format modifications from Cornelius, V., et al. (2020). "Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials." Trials 21(1): 1028 under a Creative Commons Attribution 4.0 International License* http://creativecommons.org/licenses/by/4.0/

3.5 Discussion

3.5.1 Summary

This review provides a comprehensive picture of statistical methods that have been specifically developed or adapted for analysis of harm outcomes in RCTs, building on and updating existing summaries of the literature.[42, 137, 139, 207] It reveals that there are a broad range of published statistical methods available that account for the complexities of harm outcomes.[208, 209] Many of which could be adopted into current practice with relative ease and that could potentially lead to improved analysis of harm outcomes. Chuang-Stein and Xia have proposed examples of industry strategies adopting such methods.[139] However, based on the review undertaken in chapter two and low citations of these articles, evidence of the application of these approaches for the analysis presented in the primary results publications is limited.[45, 208, 209] The complex nature of harm outcomes collected in studies designed around efficacy and statistical issues raised under a traditional frequentist approach (e.g. of multiple testing and insufficient power) are sometimes used to defend the continued practice of simple analysis approaches for harm outcomes in RCTs. Given the array of methods available, some of which, to some extent, address these issues, there is an opportunity for researchers to move away from the prevalent simplistic approaches identified in chapter two.

Under the frequentist paradigm, performing multiple hypothesis tests increases the likelihood of incorrectly raising a signal for an ADR due to a chance imbalance (i.e. inflated type I error). This could be considered less problematic , if incorrectly raising a signal for an ADR simply means that it undergoes closer monitoring in ongoing or future trials.[6] However, evidence indicates that investigators incorrectly interpret the output of such tests as evidence of a definitive difference or conversely as an absence of harm.[210] A switch in thinking to that of signal detection as advocated by some in the literature, using the p-value as an indicator to raise a signal of potential harm, instead of null hypothesis testing to make definitive conclusions, offers an alternative paradigm for analysis of

harms.[25 202] Strict control of type I error can also be problematic if it increases the risk of missing important signals that would otherwise have undergone closer monitoring. A recent update to the NEJM statistical guidelines to authors, indicates that this is an approach they support, stating "*Because information contained in the safety endpoints may signal problems within specific organ classes, the editors believe that the type I error rates larger than 0.05 are acceptable*". Issues of multiplicity are not typically an issue for multivariate approaches identified in this review. The approaches identified typically aimed to identify overall differences in harm profiles, which could help to identify any differences in the burden of harm participants' experience. However, a global approach, looking for overall difference could mask important differences for specific events. Therefore, such approaches are likely to be useful in conjunction with more specific event-based analysis.

Despite a lack of power to undertake formal hypothesis tests for harm outcomes such an approach is common, the results of which are often accompanied with inappropriate conclusions that a treatment is 'safe' or 'well-tolerated'. Prespecified analysis plans for prespecified events of interest could prevent post-hoc, data-driven, hypotheses testing, as well as setting out the purpose of any analysis to be undertaken on emerging events. Nevertheless, most analysis of harm outcomes is undertaken without a clear objective.

Many of the methods identified in this review are not reliant on the hypothesis-testing framework and thus are not as affected by issues of power and multiplicity outlined above. In addition, they were predominantly suitable for the analysis of emerging events. Consequentially these methods provide a multitude of useful, alternative ways to analyse emerging harm outcomes, where it could be argued suitable methods are most needed. Graphics have much to offer, they can help simplify complex data into digestible summaries suitable for delivering messages to a variety of audiences

and they can help to detect signals for potential ADRs from the body of emerging events (recommendations on appropriate plots are explored in chapter five).[139] Similarly, estimation methods, many of which incorporate often under-utilised information on, time of occurrence or recurrent events, can be used to quantify distributional differences in the harm profile between treatment groups. Both estimation and visual approaches rely on subjective assessments regarding a decision whether to flag a signal for a potential ADR. As such, they both provide a useful means to support analysis of harm outcomes but might be most useful if used in conjunction (whether alongside or in future studies) with objective approaches such as statistical tests or Bayesian decision-making methods, which provide clear output for interpretation to detect differences between treatment groups.

Evidence on a drug harm profile is accumulated over the entire development pathway, and such information can be used to prespecify suspected harm outcomes for monitoring. Bayesian decision-making approaches can then be used to formally incorporate existing information into analyses. Outputs can be used to aid objective decisions about the conduct of ongoing trials or future trials based on the emerging harm profile. Bayesian approaches that incorporate prior and/or accumulating knowledge into ongoing analyses are an efficient use of existing evidence that do not suffer to the same extent with issues of insufficient power or multiplicity as hypothesis test approaches.[200, 201, 211] In the Bayesian paradigm, type I and II error rates are not relevant and as such make this a potentially useful approach for analysis of harm outcomes where such 'up-front' decisions are not always feasible. In fact, unlike the traditional hypothesis test approach, issues of multiplicity have been shown to be less problematic in the Bayesian paradigm. In the case of sequential data analysis (repeated tests on same data) analysing the data at multiple points under a Bayesian framework does not need to account for type I errors through for example an alpha-spending function.[211] In the case of analysing multiple parameters (i.e. lots of different outcomes being tested at the same time), Bayesian analysis assumes one of the following scenarios:

1) Identical parameters so data can be pooled – e.g. same outcome across different trials;
2) Each parameter is independent – e.g. each has its own prior;
3) Exchangeable parameters – the parameters are 'similar' with a common prior

The last scenario is useful when analysing the body of emerging harm data. In this scenario, using a hierarchical model shrinks effects towards the prior and therefore each other, which gives comparisons that are more conservative than estimates in the frequentist paradigm.[200, 201] A Bayesian power analysis could be calculated for predefined outcomes, however it has been proposed that such calculations are not necessary and Bayesian trials should simply increase their sample size until enough evidence exists to make a decision.[212] For studies designed around a primary efficacy outcome it is unlikely to be feasible to keep increasing the sample size to analyse the emerging harm profile. However, when analysing emerging events the power of a decision based on Bayesian posterior probability can be calculated post-hoc under the assumed data generating process and sample size, which could be helpful when comparing to methods under the frequentist paradigm. Bayesian methods are sensitive to the prior information incorporated, the impact of which can be explored, with careful consideration given to the appropriateness of the source of prior knowledge and its applicability.[213] Bayesian approaches can also be used to analyse emerging events where it has not been possible to use historical data to inform priors, in this scenario, non-informative priors can be used and this will be explored further in chapters six and seven.

The most appropriate method for analysis will depend on whether events are prespecified or are emerging, as well as the aims of the analysis. Statistical analysis strategies could be prespecified at the outset of a trial, as is done for efficacy outcomes, for both prespecified and emerging harm outcomes. With a multitude of specialist methods for the analysis of harm outcomes, it is likely that a combination of approaches would be most suitable and the unwavering reliance on solely presenting tables of frequencies and percentages no longer necessary. Whilst consumers of clinical

trial articles are likely to still require access to frequencies of events, exploration of how the use of these more specialist analysis methods could be utilised and how best to incorporate them into reported results, alongside or supplementing frequency tables is needed.

3.5.2 Recommendations for immediate adoption

Without formal quantitative comparisons of the methods identified within this review, it is unclear which are the most optimal for the analysis of harm outcomes. Some of these approaches will be explored in subsequent chapters of this thesis but immediate advantages could be made through incorporation of prespecified, objective methods alongside clinical review to monitor harms. As with efficacy outcomes, specifying clear objectives setting out the purpose of the analysis to be performed for both prespecified harms and emerging events, along with clear analysis strategies at the design stage of trial development will improve transparency. Recommendations for immediate adoption are summarised in table 3.5 and more specific analysis strategies that could be adopted immediately are discussed in chapter 8 section 8.3.1.

Table 3.5: Recommendations for analysis of harm outcomes

	Recommendation
Design	Specify clear objectives setting out the purpose of the analysis to be performed for both prespecified harms and emerging events
	Specify a clear analysis strategy for both prespecified harms and emerging events
Analysis	Incorporate more objective methods into the evaluation of harms
	Consider if methods to control type I error are needed and avoid using p-values and 95% confidence intervals as a proxy for null hypothesis tests (specifically when analysing multiple, undefined, emerging events)
	Reduce information loss when analysing at participant level by analysing events within participants and not just participants with at least one event and utilise methods that incorporate information on recurrent events and time of occurrence
	Consider methods that incorporate prior knowledge on the harm profile
	Use visualisations to explore data and help summarise large amounts of data to communicate important messages

3.5.3 Limitations

The performance of all of the individual methods identified in this review have not been formally examined and no quantitative comparisons have been made. Recommendations and methods taken forward for further research have been chosen based on perceived ease of use and implementation, the information that they take into consideration and their objectivity. Methods not specifically proposed but that could be applied to the analysis of harm outcomes were not considered in this review for practical reasons. Attention was restricted to methods specifically designed or adapted for harm outcomes to gain a better understanding on what has been done to prevent duplication in future work and to identify unknown and potentially useful methods. There were also methods identified that have been proposed for the analysis of harm outcomes in RCTs that were excluded as they did not meet the eligibility criteria. This included methods designed to be used in the RCT setting that did not utilise the control group, combining treatment arms in an effort to preserve blinding .[214, 215] Whilst these methods offer alternative, objective ways to detect potential harms they were excluded from this review in line with the aims to identify methods that utilise the control group to enhance inference.

3.5.4 Areas to explore further within this thesis

In light of this review, the following areas would benefit for further research and I will explore them in subsequent chapters:

i) The reasons for low uptake of available methods are unknown but warrants further investigation to gain a better understanding of current practices beyond what is reported in journal articles. As well as gaining insights into any potential barriers to implementation and raising awareness of these and new methods, where appropriate, to improve the analysis of harm outcomes in RCTs.

ii) It is not clear which analysis methods trialists should be using. Exploration of analysis methods for objective analysis of harm outcomes focusing on different data types including binary, count, time to event and continuous outcomes will allow us to recommend methods for adoption. Including:

a. Exploring methods identified in this review such as the beta-binomial model and gamma-Poisson models that can incorporate prior information and account for information on recurrent events and time of exposure, as well as survival methods that incorporate time of occurrence as candidate methods for an objective means to analyse emerging harm outcomes.

b. Investigating methods designed to detect a time dependent relationship between intervention exposure and events, which has been shown to be a powerful method to raise signals for harms in the observational setting but this review highlighted have not been considered in the clinical trial setting.[124] For example, detecting a disproportionality in the proportion of events between intervention and control groups would potentially be indicative of a causal relationship and warrants exploration.

3.5.5 Conclusions

Analysis of harm outcomes in clinical trials is complex and there is a reliance on simple approaches that do not fully utilise available data. In this chapter, I undertook a review that revealed a multitude of methods specifically designed to overcome some of these complexities but evidence of application of any of these methods in clinical trial publications is limited. This is supported by the results of both the systematic review of journal reports (chapter 2) and the citation search for application of these methods through Web of Science referenced in this chapter. A quantitative evaluation of these methods will enable researchers to navigate which are the most appropriate; and gaining a better understanding as to why there is a reliance on simplistic approaches will enable

a strategy to be developed to tackle suboptimal practice and ultimately improve analysis of harm outcomes.

4. Understanding current practice, identifying barriers and exploring priorities for the analysis of harm outcomes in RCTs: a survey of academic and industry statisticians

4.1 Introduction

The comprehensive methods review described in chapter three revealed a broad range of published statistical methods proposed to analyse harm outcomes for both the interim and final analysis of clinical trials.[130] Many of these could be adopted into current practice with relative ease. Recommendations for immediate adoption to improve analysis practices are provided in table 3.5 and discussed in the final chapter of this thesis, as well as the 2013 article by Chuang-Stein and Xia, which recommends a core set of methods for analysis of harm outcomes with clinical examples.[139] However, the review described in chapter two demonstrated that these methods are not used for the final analysis presented in the primary results publication, and chapter three revealed that there are minimal citations of these published methods applied in the RCT setting, which further suggests uptake of these methods is low.[45, 208, 209] As a consequence of this finding I sought to explore the level of awareness of these methods and the reasons for low uptake. Understanding the reasons for this low uptake will help identify potential solutions to improve the analysis of harm outcomes in RCTs and will provide crucial information on appropriate next steps for this work.

Some of the work presented in this chapter has been published in the BMJ Open.[216] This chapter acknowledges the support of Dr. Suzie Cro of Imperial College London who along with my supervisor, Dr. Victoria Cornelius, helped facilitate the workshop at the UKCRC CTU network's biannual statistician's operations group meeting to disseminate the results of this chapter.

4.2 Aims

The overarching aim of the research in this chapter is to gain a better understanding of analysis practices and the rationale for the methods used for the analysis of harm outcomes by clinical trial statisticians working in the UK academic setting and industry. Specifically the objectives were to:

1) Survey clinical trial statisticians to identify their current practice for the final analysis of harm outcomes in pharmacological RCTs to gain an understanding of practices beyond those evident in journal articles.
2) Assess the awareness of methods designed specifically for the analysis of harm outcomes.
3) Explore statisticians' priorities, concerns and identify any perceived barriers when analysing harm outcomes.
4) To compare the results from academic and industry participants, identifying differences of note and highlighting any areas for cross industry learning opportunities.
5) To identify priorities for future work.

4.3 Methods

4.3.1 Study design

A cross-sectional, online survey of clinical trial statisticians working in UK Clinical Research Collaboration (UK CRC) clinical trial units (CTU) and invited industry statisticians including both pharmaceuticals and clinical research organisations (CROs). This was followed by an open invitation to statisticians not already targeted, which was promoted at the Statisticians in the Pharmaceutical Industry (PSI) conference and via social media. The format of an online survey was chosen for ease of administration, ability to reach a wide geographical audience, and cost and time efficiencies. This was a cross-sectional survey so once participants completed and submitted their responses there was no further follow-up.

4.3.2 Sample size

The aim was to recruit a minimum of one statistician from each of the UKCRC registered CTUs of which there were 51 at the time, and from a sample of pharmaceutical companies and CROs in the UK. As the aim of the survey was to gain a better understanding of analysis practices and measure awareness and opinions of proposed methods to analyse harm outcomes with no hypothesis tests planned, no formal sample size calculation was undertaken but more than 50 participants would provide a good estimate of the score distribution.

4.3.3 Development of content and structure

The survey content was developed in-line with the research aims and was based on recommendations from current guidance on reporting standards for harm outcomes and previous research that examined barriers to the uptake of new methodology.[28, 34, 217, 218] Content was discussed with supervisors and questions refined following these discussions. Questions were developed to cover the following themes:

i) Current practice for analysing harm outcomes.

ii) Factors influencing analysis performed such as preferences and trial characteristics.

iii) Barriers encountered when analysing harm outcomes including factors relating to knowledge, resources, priorities and trial characteristics.

iv) Awareness and opinions of methods specifically designed or proposed for analysing harm outcomes.

v) Concerns and barriers of implementing methods specifically designed for analysing harm outcomes such as limitations of methods or trial team expectations.

vi) Opinions on potential solutions to support a change in analysis practices for harm outcomes.

So that completion of the survey was not overly burdensome and to help ensure high completion rates the number of questions was limited so that it would take no more than approximately 15 minutes to complete. Questions were predominantly closed form to enable ease of completion by participants as well as to ease the quantification of the results. Where appropriate open-ended questions were included to allow for more detailed responses and comments, enabling a deeper understanding of current practice. Closed form responses were measured using Likert scales. Survey questions for UKCRC CTU and industry statisticians were identical and can be found in appendix A4.1.

Survey questions specifically asked about practices relating to the final analysis of emerging harms, and clearly stated that answers should not relate to prespecified harm outcomes or interim analyses. This was for practical purposes so that the remit remained clear and did not become overly onerous which would likely have resulted in incomplete submissions and a reduced response rate.

The survey was piloted remotely on a small sample of clinical trial statisticians (n= 6) prior to launching nationwide. This pilot included four statisticians from within Imperial Clinical Trials Unit, the current chair of the UKCRC statistics operations group and an external academic statistician based in a health services department. The pilot involved sending a test version of the survey via the online platform, SurveyMonkey, where the survey would be hosted, for review and written feedback. Running a pilot helped to ensure sufficient coverage of questions, understanding of the questions, whether sufficient response categories had been included, and if certain questions were consistently left unanswered, as well as to test the usability and functionality of the online platform before it was launched.[219] Feedback received highlighted ambiguities with question phrasing and important omissions with question responses that were rectified before the survey was finalised and circulated.

4.3.4 Sampling and recruitment

Statisticians known to be predominantly involved in the analysis of clinical trials were targeted. Specifically people were eligible to participate in the survey if:

i) Their current role was as a senior statistician or equivalent at a UKCRC CTU or pharmaceutical or CRO;

ii) They had experience of planning and preparing final analysis reports for pharmacological RCTs.

Statisticians were sampled from three population via the following:

1) The UKCRC CTU network's statistics operation group supported the survey and contacted each registered CTUs' senior statistician regarding the survey in April 2019. The UKCRC CTU network is a group of academic clinical trials units. Units are accredited to run high quality clinical trials by an international panel of experts and UK funders such as the National Institute for Health Research (NIHR) and Medical Research Council (MRC) encourage collaboration with an accredited UKCRC CTU to undertake any trials they fund. A benefit of using the UKCRC CTU network is the anticipated high response rate, which has been achieved in previous similar surveys within this group.[218, 220]
2) Email invitations were also sent directly to a convenience sample of senior statistical contacts working in UK based pharmaceuticals and CROs (a copy of the email invitation sent to CTUs and industry contacts can be found in appendix A4.2). Without access to a similar network of industry statisticians', personal contacts were relied on for this sample. The pharmaceutical industry has the highest research and development expenditure of any industry within the UK (as reported for year 2018, the most recent available figures) and statisticians working within this industry are amongst those at the forefront of pharmacological research within the UK.[221]

Email invitations sent to both groups requested that one statistician within the unit or organisation complete the survey. Reminder emails were sent to non-responders and the platform remained open for 8 weeks from the point of emails being sent.

3) An open platform to complete the survey was also created for statisticians not already targeted. This was promoted in June 2019 at the annual PSI conference in London. PSI is an organisation that is "*dedicated to leading and promoting the use of statistics within the healthcare industry for the benefit of patients*" with a membership primarily comprised of statisticians working in industry. Participation via the open platform was also advertised on the Effective Statistician podcast, which was broadcast in July 2019 and via Twitter and LinkedIn platforms over the same period. The open platform remained open for 10 weeks from the initial launch at the PSI conference. This platform remained open slightly longer to ensure a sufficient timeframe for completion was available after the final promotion activity.

Statisticians placed within CTUs and pharmaceutical companies have a wealth of experience of analysing clinical trials. They were specifically targeted for this survey (over statisticians working on trials outside of industry or CTUs) to provide insight into high quality trials, given their (inter)nationally recognised reputation for conducting rigorous and well-designed research, as well as their ability to set standards that are adopted by the wider trial community. Early engagement with these groups also has the potential to help dissemination, as well as raise awareness and adoption of any recommendations that are a product of this research. Targeting both academic and industry statisticians would enable a deeper exploration of differences in practices between sectors identified in chapter two. As well as providing an opportunity for potential cross sector learnings. Completion of the survey automatically entered participants into a prize draw to win £50 worth of gift vouchers.

4.3.5 Ethics and consent

The Head of Imperial Clinical Trials Unit and Imperial College Joint Research Compliance Office (JRCO) (which has subsequently been renamed the Research Governance and Integrity Team) reviewed the survey and granted it ethical approval on 20th February 2019 (ICREC reference: 19IC5067). The invitation to participate in the study included the participant information sheet, which was repeated at the beginning of the survey before participants formally entered (the participant information sheet sent can be found in appendix A4.3). Participants were encouraged to read the information sheet and discuss the study with others or myself if they wished. If invitees were happy to enter into the trial at that point their consent was taken as implied upon submission of the completed survey.

4.3.6 Analysis

Descriptive analysis was undertaken, including frequencies and proportions for each questionnaire item, and where appropriate was accompanied with visual summaries. Results are presented for each possible item response plus the frequency and proportion of participants that showed support for an item was summarised by combining the 'always' and 'often' or 'strongly agree' and 'agree' categories. Free text comments were classified into themes and discussed with one supervisor (VC) to ensure agreement. Participants were classified according to affiliation into either CTU/public sector (referred to from here as public sector participants) or industry sector and analysis was stratified by sector. Response rates were calculated for the public and industry samples, it was not possible to make such a calculation for the open-platform, as the denominator for this sample was unknown.

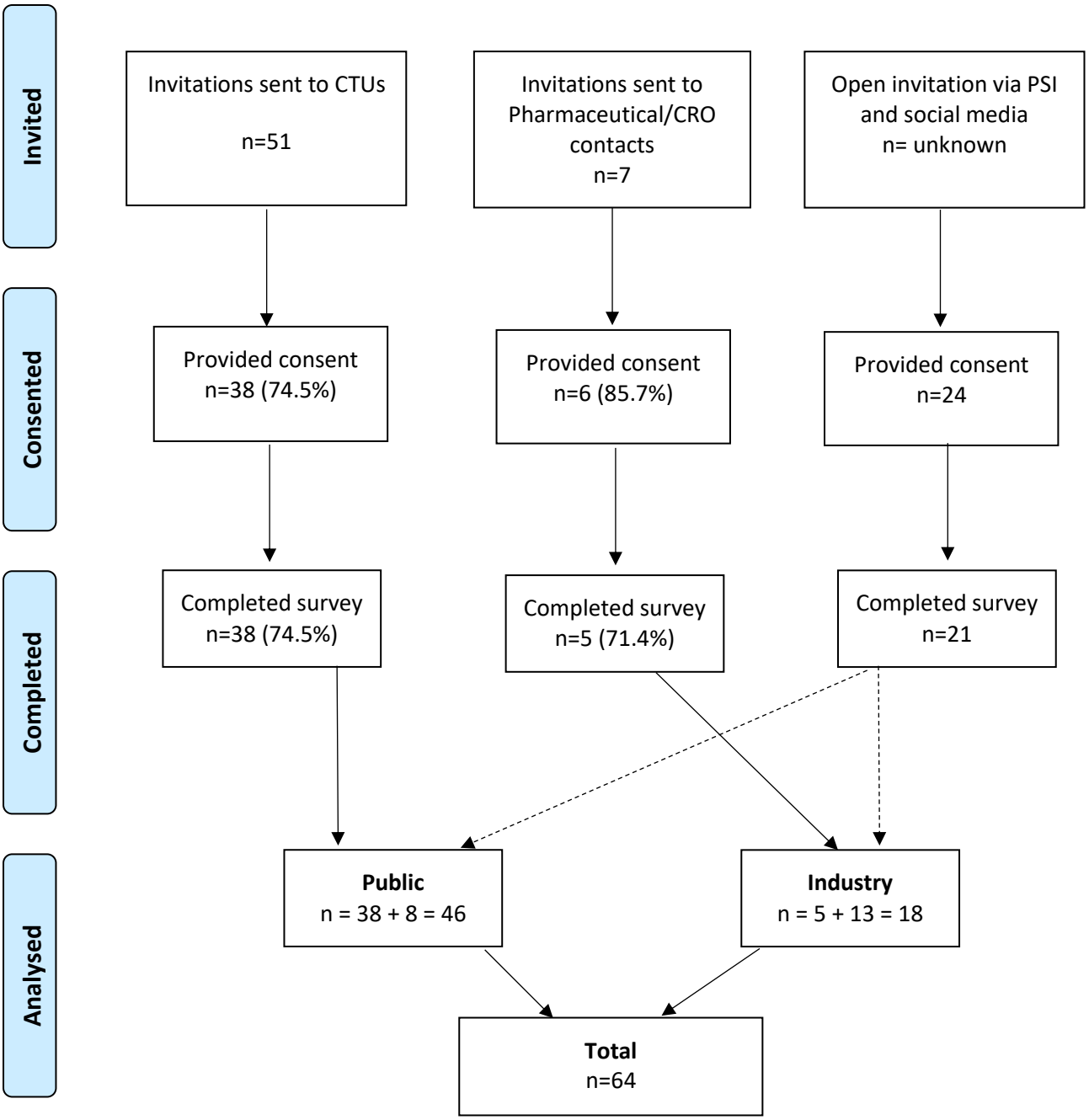
4.4 Results

4.4.1 Participant flow

Emails inviting participation in the online survey were sent to contacts at fifty-one CTUs via the UKCRC CTU network and seven personal contacts working in the UK pharmaceutical industry (Astra-Zeneca, Boehringe-Ingelheim, Glaxo-Smith-Kline (GSK), Novartis and Roche) and CROs (Cytel and IQVIA). Thirty-eight (75%) participants from CTUs and six (86%) industry contacts consented to participate in the study. One industry contact failed to complete the survey after providing consent giving an overall response rate of 74%.

Twenty-four people consented to participate via the open platform, of which three failed to complete the survey after providing consent. Of the 21 participants who completed the survey, eight indicated they worked in the public sector and were grouped with the CTU participants and thirteen indicated they worked for a pharmaceutical or CRO and were grouped with the industry sector participants for analysis. In total 64 participants took part in the survey, of which forty-six were grouped as public sector participants and eighteen were classified as industry participants (figure 4.1).

Figure 4.1: Flow diagram of participation in the online survey[†††]



Acronyms: CTUs – clinical trial units; CRO – clinical research organisations; PSI - Statisticians in the Pharmaceutical Industry

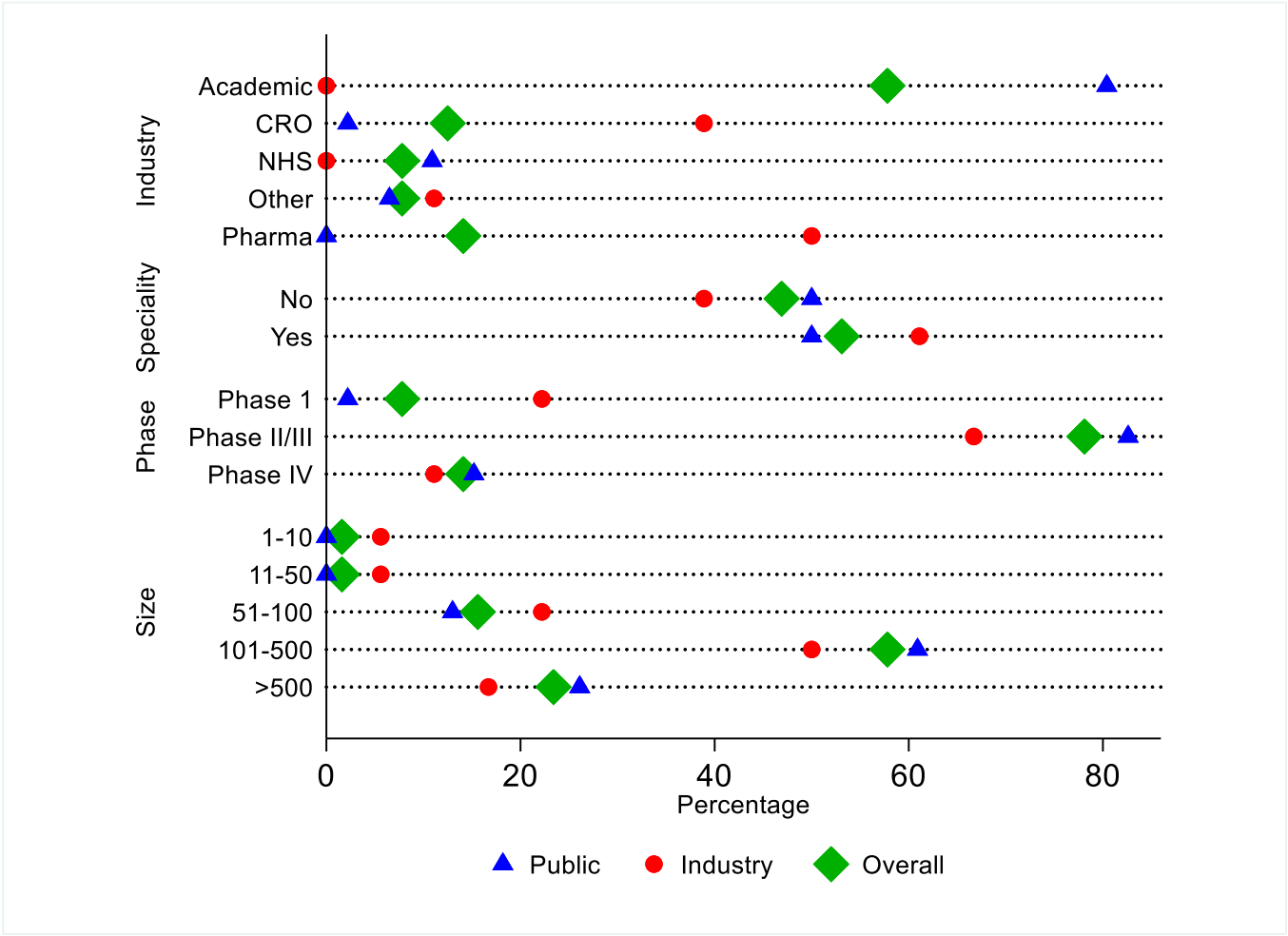
[†††] *Reprinted from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

4.4.2 Participant characteristics

Figure 4.2 provides a visual summary and table 4.1 provides summary statistics on participant characteristics by sector and for the overall sample. Eighty-one percent of participants indicated that they typically worked on studies of more than 100 participants, and 80% typically worked on phase II/III trials. When examining the results by sector it was revealed that a greater proportion of industry participants typically worked on phase I/dose finding trials compared to public sector participants (22% vs 2%). Participants had a mean number of years of experience of 12.8 (SD 8.3) (median 11.5 years, range (1-35 years)), with industry participants having slightly more experience (mean 14.7 years (SD 10.7)) compared to public sector participants (mean 12.0 years (SD 7.2)).

The majority of participants indicated that they predominantly worked on oncology trials (public sector n=8 and industry n=7 (44% of those indicating a clinical area)). This is unsurprising given the dominance of oncology trials in the UK where the number of new commercial oncology trials has consistently doubled the number of new trials in other clinical areas (such as immune, nervous and cardio-metabolic diseases) over the period 2012 to 2017.[222] Other participants reported that they worked across a range of different therapeutic areas (n=14), as well as healthcare settings (n=3), types of interventions including psychological (n=1), complex interventions (n=1) and non-CTIMPs (n=1), and different trial designs, such as adaptive designs (n=1) (full details are provided in appendix A4.4).

Figure 4.2: Participant characteristics by employment sector and overall‡‡‡



Acronyms: CRO: Clinical Research Organisation; NHS: National Health Service; Pharma: Pharmaceuticals
Speciality: Participants were asked if there was a clinical area they predominantly worked on e.g. cancer, surgery or whether they work across clinical areas.
Size: average size of clinical trials they typically work on

‡‡‡ *Reprinted from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.1: Participant characteristics by employment sector and overall[§§§]

		Public (N=46)		Industry (N=18)		Overall (N=64)	
Characteristics		**n**	**%**	**n**	**%**	**n**	**%**
Work setting	Academic institution	38	82.6	0	0.0	**38**	**59.4**
	CRO	1	2.2	7	38.9	**8**	**12.5**
	NHS trust	5	10.9	0	0.0	**5**	**7.8**
	Pharmaceutical	0	0.0	9	50.0	**9**	**14.1**
	Other	2	4.3	2	11.1	**4**	**6.3**
Speciality*	No	23	50.0	7	38.9	**30**	**46.9**
	Yes	23	50.0	11	61.1	**34**	**53.1**
Typical trial phase	Phase I/Dose-finding	1	2.2	4	22.2	**5**	**7.8**
	Phase II/III	38	82.6	12	66.7	**50**	**78.1**
	Phase IV	7	15.2	2	11.1	**9**	**14.1**
Typical trial size**	1-10	0	0.0	1	5.6	**1**	**1.6**
	11-50	0	0.0	1	5.6	**1**	**1.6**
	51-100	6	13.0	4	22.2	**10**	**15.6**
	101-500	28	60.9	9	50.0	**37**	**57.8**
	>500	12	26.1	3	16.7	**15**	**23.4**
Years of experience	Mean (SD)	12.0	(7.2)	14.7	(10.7)	**12.8**	**(8.3)**
	Median (min, max)	12.0	(1, 30)	15.5	(1, 35)	**11.5**	**(1, 35)**

Acronyms: CRO: Clinical Research Organisation; NHS: National Health Service; SD: standard deviation; min: minimum; max: maximum

*Participants were asked if there was a clinical area they predominantly worked on e.g. cancer, surgery or whether they work across clinical areas.

**Size: average size of clinical trials they typically work on.

4.4.3 Results

Current analysis practice for emerging harm outcomes

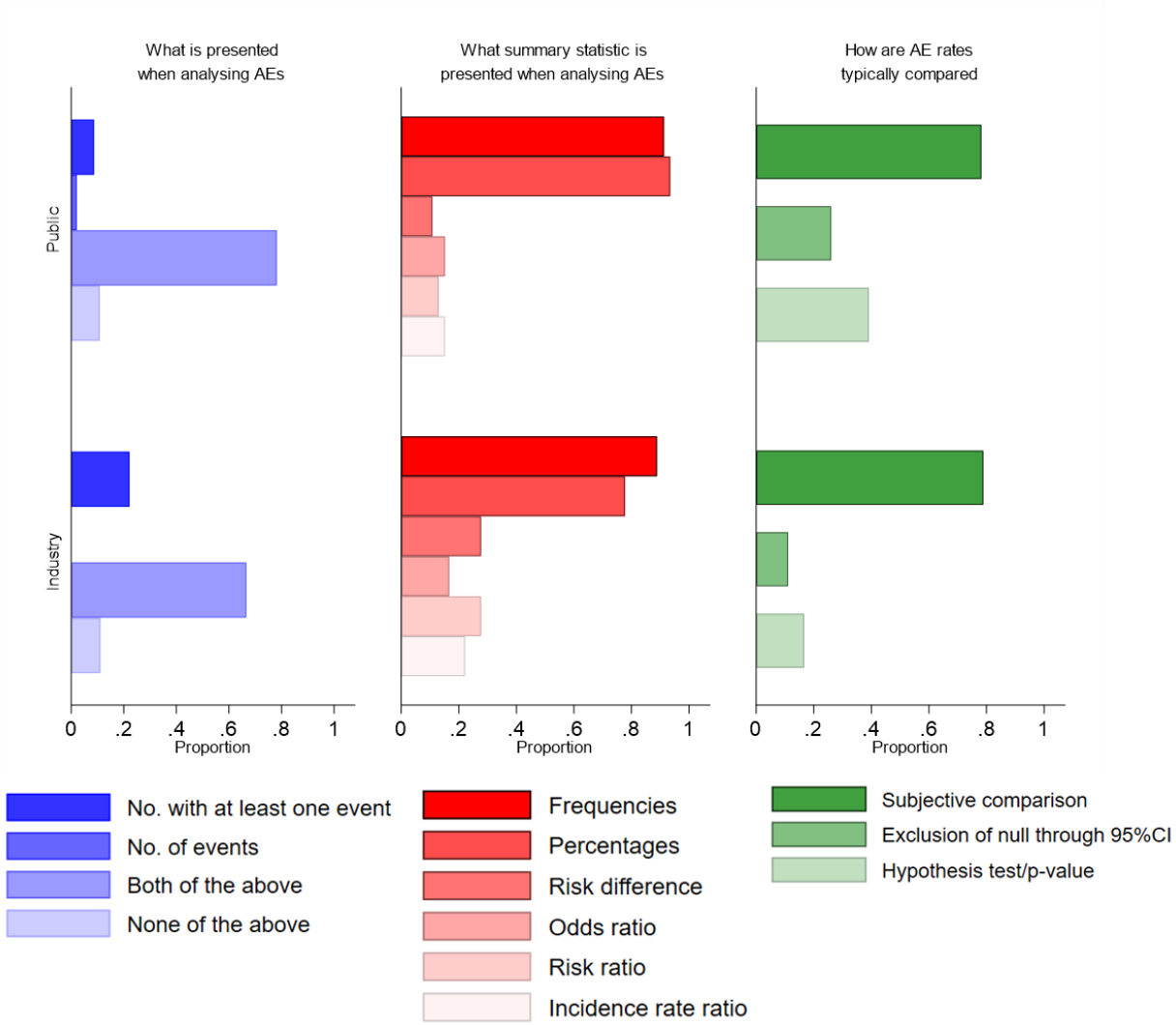
Thinking about the final analysis undertaken, participants were asked what information they typically presented on emerging harm outcomes. Three-quarters reported presenting both '*the number of participants with at least one event'* and '*the number of events'*, 13% indicated only presenting '*the number with at least one event'*, 2% indicated that they only present *'the number of events'* and 11% reported not presenting this information. Results were broadly similar across sectors but slightly more industry participants presented '*the number of participants with at least one event'* compared to public sector participants (22% vs. 9%) (figure 4.3 and table 4.2).

[§§§] *Reprinted from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Participants were also asked what summary statistic they would typically use when presenting information on emerging harm outcomes. Ninety percent reported that they use frequencies and percentages, less than 20% indicated they presented risk differences (16%), odds ratios (16%) or risk ratios (17%), and just under a quarter reported use of incidence rate ratios (23%) (figure 4.3 and table 4.2). Sixty-one percent of participants reported frequencies and/or percentages with no accompanying relative measure (odds ratios, risk ratios or incidence rate ratios). Comparative summaries were more widely reported by industry participants compared to public sector participants, with nearly 30% reporting use of risk differences and risk ratios compared to less than 15% using such measures in the public sector and nearly 40% presented incident rate ratios compared to 17% in the public sector. Five participants indicated that the summary statistic they present would depend on the specific study being analysed.

When asked how they would typically compare event rates between treatment groups, 80% of participants reported relying on subjective comparisons, 33% indicated they would compare rates using hypothesis tests, and 22% stated they would use 95% confidence intervals as a means to examine the null hypothesis of no difference. Comparing results by sector revealed that public sector participants indicated a wider use of both hypothesis tests (39% public sector versus 17% industry) and 95% confidence intervals (26% public sector versus 11% industry) to compare event rates between treatment groups. Fourteen percent of participants reported another means of comparison table 4.2, two of these related to the calculation of confidence intervals for precision, one indicated use of a graphical summary and four comments indicated reservations about using statistical tests for comparisons of emerging harm outcomes e.g., "*statistical testing is rarely requested and raises multiple testing concerns".*

Figure 4.3: Visual summary of analysis practices of survey respondents by employment sector



Acronyms: No.: number; CI: confidence intervals

Table 4.2: Information on emerging harms typically presented by employment sector and overall[****]

	Public (N=46)		Industry (N=18)		Overall (N=64)	
Information presented	**n**	**%**	**n**	**%**	**n**	**%**
Number of participants with at least one event	4	8.7	4	22.2	**8**	**12.5**
Number of events	1	2.1	0	0.0	**1**	**1.6**
Both of the above	36	78.3	12	66.7	**48**	**75.0**
None of the above	5	10.9	2	11.1	**7**	**10.9**
Other[1]	16	34.8	6	33.3	**22**	**34.4**
Summary statistic†						
Frequencies	42	91.3	16	88.9	**58**	**90.6**
Percentages	43	93.5	14	77.8	**57**	**89.1**
Risk difference	5	10.9	5	27.8	**10**	**15.6**
Odds ratio	7	15.2	3	16.7	**10**	**15.6**
Risk ratio	6	13.0	5	27.8	**11**	**17.2**
Incidence rate ratio[2]	8	17.4	7	38.9	**15**	**23.4**
Other[3]	6	13.0	4	22.2	**10**	**15.6**
Comparison of events †						
Subjective comparison	36	78.3	15	83.3	**51**	**79.7**
Exclusion of null through 95% confidence interval	12	26.1	2	11.1	**14**	**21.9**
Hypothesis test/p-value	18	39.1	3	16.7	**21**	**32.8**
Other[4]	4	8.7	5	27.8	**9**	**14.1**
Awareness of any published methods specifically to analyse harm outcomes¥						
No	25	56.8	4	23.5	**29**	**47.5**
Yes	11	25.0	12	70.6	**23**	**37.7**
Don't know	8	18.2	1	5.9	**9**	**14.8**
Undertaken any specialist analysis not mentioned in your previous response*						
No	38	88.4	14	82.4	**52**	**86.7**
Yes	5	11.6	3	17.6	**8**	**13.3**

† Participants were able to provide multiple responses to this question

¥Two public sector and one industry participant failed to answer this question

*Three public sector and one industry participant failed to answer this question

[1] Other ways of presenting information on harm outcomes included presenting information on: overall number of events (n=2); number of patients experiencing 0, 1, 2 etc. events and number of events per patient (n=2); duration (n=1); relatedness (n=1) and severity (n=7) (full free text comments in appendix A4.5).

[2] Incorporates free text comments that described summaries synonymous with incidence rate ratios

[3] Included a comment that a participant presents the "median number (IQR)" of events.

[4] Other comments related to the calculation of confidence intervals for precision (n=2), one indicated use of a graphical summary (n=1) and four cautioned against the use of testing.

Thirty-eight percent of participants indicated an awareness of methods published specifically for the analysis of harm outcomes in RCTs (table 4.2). Methods mentioned were classified into one of four

[****] *Reprinted from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

groups summarised in table 4.3 and away from the individual trial setting two participants indicated they would use meta-analysis for harm outcomes e.g. *"Meta-analysis of rare events"*. Six participants also mentioned specific theoretical and applied examples they were aware of in the literature.[34, 62, 141, 155, 197, 223] Full free text comments are reported in appendix A4.6.

Table 4.3: Methods participants mentioned they were aware of specifically for the analysis of harm outcomes

Method classification	**n**	**Example**
Modelling approaches appropriate to different data types	6	*"Bayesian methods to analyse low frequency event data"* (public sector participant)
Incidence rates	5	"*crude incidence rates, exposure-adjusted incidence rates, mean cumulative function (MCF)*" (public sector participant)
Graphics	2	"*Graphics for biological parameters (ellipse ci)*" (public sector participant)
Bayesian approaches	1	*"Bayesian methods to analyse low frequency event data"* (public sector participant)

Despite modest levels of awareness of specific methods for the analysis of harms, only thirteen percent of participants reported undertaking specialist analysis of harm outcomes (table 4.2) and 8% of participants provided details of the analysis undertaken. Responses were summarised into groups which are presented in table 4.4 (full text comments are reported in appendix A4.7). There was a greater awareness of such methods by industry participants (71%) compared to public sector participants (25%), however use of these methods were similar between sectors (18% in industry compared to 12% in public sector).

Table 4.4: Participants' use of specialist methods for analysis of emerging harm outcomes

Method classification	n	Example
Time-to-event analysis	2	"*In characterising safety signals I have used Time to Event, Event rates, prevalence"* (industry participant)
Data visualisations	1	*"Data visualisation (which is more or less equivalent to frequencies and percentages)"* (industry participant)
Bayesian methods	1	*"Bayesian methods for sparse adverse events data meta-analysis"* (public sector participant)
Incorporating repeated events	1	*"For within-patient repeated events we have produced comparisons with a 2-d frequency table (arm vs # events)"* (public sector participant).

Of the participants who reported that they were aware of specialist methods for the analysis of harm outcomes, opinions on why such methods were not more widely used were sought. Twenty-seven percent thought limited use was due to technical complexity and when examining results by sector this belief was more common amongst industry participants (42%) than public sector participants (10%). Over a third of participants thought trial characteristics such as unsuitability of sample sizes (36%) and the number of different events experienced in trials (36%) contributed to limited use and just under half (46%) thought methods were not suitable for typical event rates observed; and 46% believed methods to be too resource intensive (table 4.5). Again these beliefs were more common in industry participants, with 50% of industry participants thinking methods were not suitable for the number of different events experienced across a trial and 59% believing methods were unsuitable for typical event rates observed compared to 20% and 30% of public sector participants holding the same beliefs, respectively.

Additional reasons for the lack of use of specialist methods were given by 77% of participants and included comments relating to: concerns with the suitability of methods in relation to trial characteristics and nature of data (n=7); opposition and a lack of understanding from clinicians (n=5); a lack of need for such methods (n=3); a desire to keep analysis consistent with historical analysis (n=3); and a lack of training and resources to implement these methods (n=1). Table 4.6 displays the participants' comments attributed to each group.

Table 4.5: Reasons specialist methods are not used (by participants who were aware of such methods)[††††]

Available methods are:		Public (N=11)*		Industry (N=12)		Overall (N=23)	
		n	%	n	%	n	%
Technically too complex	Strongly disagree	1	10	0	0	**1**	**4.5**
	Disagree	7	70	6	50	**13**	**59.1**
	Agree	1	10	4	33.3	**5**	**22.7**
	Strongly agree	0	0	1	8.3	**1**	**4.5**
	Don't know	1	10	1	8.3	**2**	**9.1**
Too resource intensive	Strongly disagree	1	10	0	0	**1**	**4.5**
	Disagree	4	40	7	58.3	**11**	**50**
	Agree	5	50	4	33.3	**9**	**40.9**
	Strongly agree	0	0	1	8.3	**1**	**4.5**
Not suitable for typical trial sample sizes	Strongly disagree	1	10	0	0	**1**	**4.5**
	Disagree	5	50	4	33.3	**9**	**40.9**
	Agree	1	10	3	25	**4**	**18.2**
	Strongly agree	2	20	2	16.7	**4**	**18.2**
	Don't know	1	10	3	25	**4**	**18.2**
Not suitable for the number of different events typically experienced across a trial	Strongly disagree	1	10	1	8.3	**2**	**9.1**
	Disagree	6	60	4	33.3	**10**	**45.5**
	Agree	1	10	6	50	**7**	**31.8**
	Strongly agree	1	10	0	0	**1**	**4.5**
	Don't know	1	10	1	8.3	**2**	**9.1**
Not suitable for typical event rates observed	Strongly disagree	1	10	1	8.3	**2**	**9.1**
	Disagree	6	60	4	33.3	**10**	**45.5**
	Agree	3	30	6	50	**9**	**40.9**
	Strongly agree	0	0	1	8.3	**1**	**4.5**
Other reasons why those methods are not used	No	0	0	3	25	**3**	**13.6**
	Yes	9	90	8	66.7	**17**	**77.3**
	Don't know	1	10	1	8.3	**2**	**9.1**

*One participant failed to answer these questions despite indicating that they were aware of published methods specifically to analyse harm outcomes

[††††] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.6: Classification of participants' comments on the reasons for a lack of use of specialist methods for the analysis of emerging harm outcomes[‡‡‡‡]

Classification of reasons given for the lack of use of specialist analysis methods for emerging harm outcomes	Participant comment	Sector
1.Concern with the suitability of methods in relation to trial design characteristics and nature of data	*"…These analyses methods may also not be appropriate if there are doubts about the robustness of AE data…"*	Public
	"The strongest driver is sample size and multiplicity with multiple endpoints, limiting the power of any such analysis."	Public
	"AEs not the primary objective of trial, Pharmaceutical companies focused not on most powerful analyses, issues around multiplicity, recurrent events, low incidence of events"	Industry
	"…Most AE signals will not result in a statistically significant difference (due to low rates and trial size) and therefore a fear of testing exists, as statisticians we do not want to give the impression that the signal is not real as p>0.05!! Few trials are designed to specifically look at safety, the above methods are used on safety studies."	Industry
	"…safety analyses typically lack a scientific hypotheses to direct where to look for signals."	Public
	"…2) Multiple testing issues: The multiplicity of AEs that may arise in a RCT makes it also not really appropriate to use statistical tests because of inflated false positive error rates resulting from multiple testings. …3) Even if 1 or 2 AEs of special interest are selected for statistical testing, detecting a statistically significant difference across treatment arms requires to power the trial and calculate the sample size accordingly."	Industry
	"Appropriateness of methods depends on many factors including underlying distribution, prevalence of repeated events, whether participants were followed up for the same duration, etc. For example, if repeated events are rare and participants were followed up for the same duration then simple number and percentages of participants who experienced at least one event is sufficient. On the contrary, this will obscure the true picture if repeated events are prevalent and participants were follows up for varying periods. So I would say there is a range of statistical methods that are appropriate depending on the situation."	Public
2. Opposition and a lack of	*"Lack of emphasis placed by clinicians on the need for appropriate statistical methods to analyse adverse events data."*	Public

[‡‡‡‡] *Reprinted from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

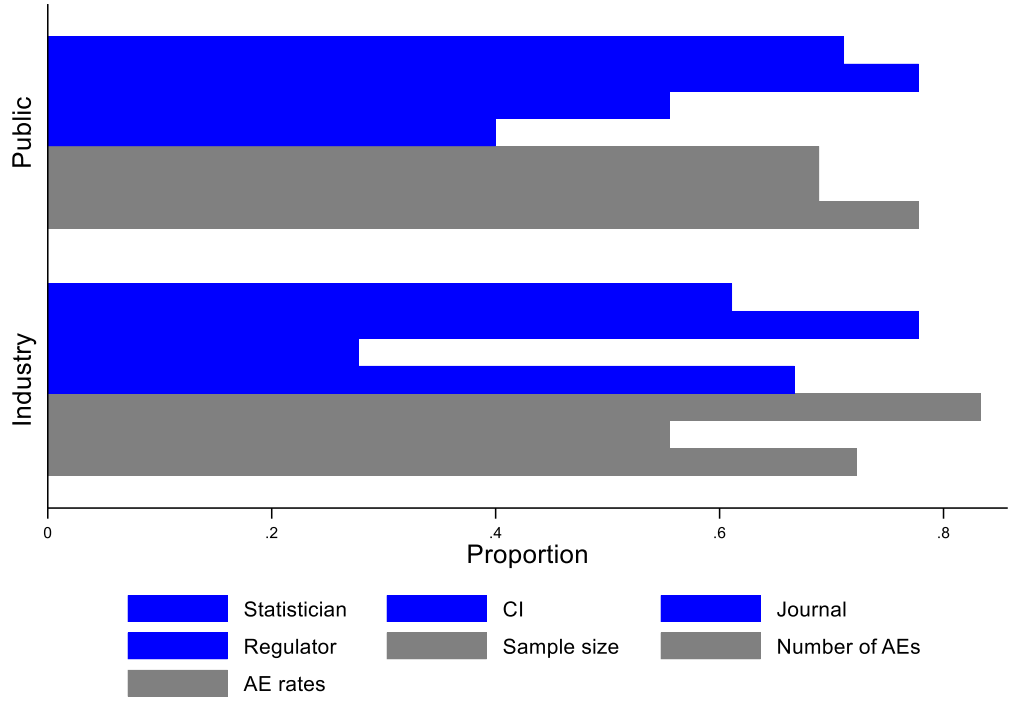
understanding from clinicians		
	"The standard approach of looking at g3+ AEs only is so accepted, there is little motivation to explore other methods. In addition, persuading clinicians to embrace other methods, can be difficult."	Public
	"*Most medical leads on clinical trials do not understand statistical analyses and only prefer a list of AEs with their percentages to be presented"*	Industry
	"A tendency to oversimplify reporting of safety signals, to make them easier to understand to non-stats people (e.g. % are easier than incidence rates)"	Industry
	"The template for reporting AEs is too basic. In the pharmaceutical industry the statisticians have little to no input into the trial paper"	Public
3. Not deemed to be needed by statisticians	"*Not required/ wanted*."	Public
	"Don't want to report additional information in CTR"	Public
	"They are perhaps not used as they are no required or appropriate for that type of trial. There is no point in applying a complex method when it is not needed (eg when AEs are collected for a well established drug; when the trial is not attempting to define a safety profile)."	Public
4. A desire to keep analysis consistent with historical analysis	"*Easiness to present always the same tables"*	Public
	"*1) High level of standardization in reporting of results of RCTs. AE tables are pretty standard and there are requirements to meet ICH3 CSR recommendations…"*	Industry
	"Consistency of analysis across trials in a development programme is often paramount. So, if AEs from a previous study have been analysed using a frequency/percentage approach, so would later trials."	Industry
5. Lack of training and resources	"*Training. Availability of code*."	Industry

Influences, barriers, and concerns

Participants were asked to assess how often key stakeholders (including statisticians, chief investigators, journals, and regulators) and trial characteristics influenced the analysis they performed on emerging harm outcomes. The most common was the chief investigator's preference for simple approaches (78%), the observed event rates (76%) and the size of the trial (73%). Over 60% of participants indicated that the statistician's preference for simple approaches was always or

often an influence (68%), and the number of different events experienced in a trial were influential (65%). Less than 50% of participants indicated that they believed that journals (48%) or regulators (48%) preference for simple approaches always or often influenced the analysis performed but when examining results by sector there was a notable difference. A greater proportion of industry participants indicated a belief that regulators' preference for simple approaches always or often influenced analysis (67% versus 40%); and a greater proportion of public sector participants indicated a belief that journals' preference for simple approaches always or often influenced analysis (56% versus 28%) (table 4.7 and figure 4.4).

Figure 4.4: Influences on analysis performed on emerging harm outcomes by employment sector (always and often categories combined)

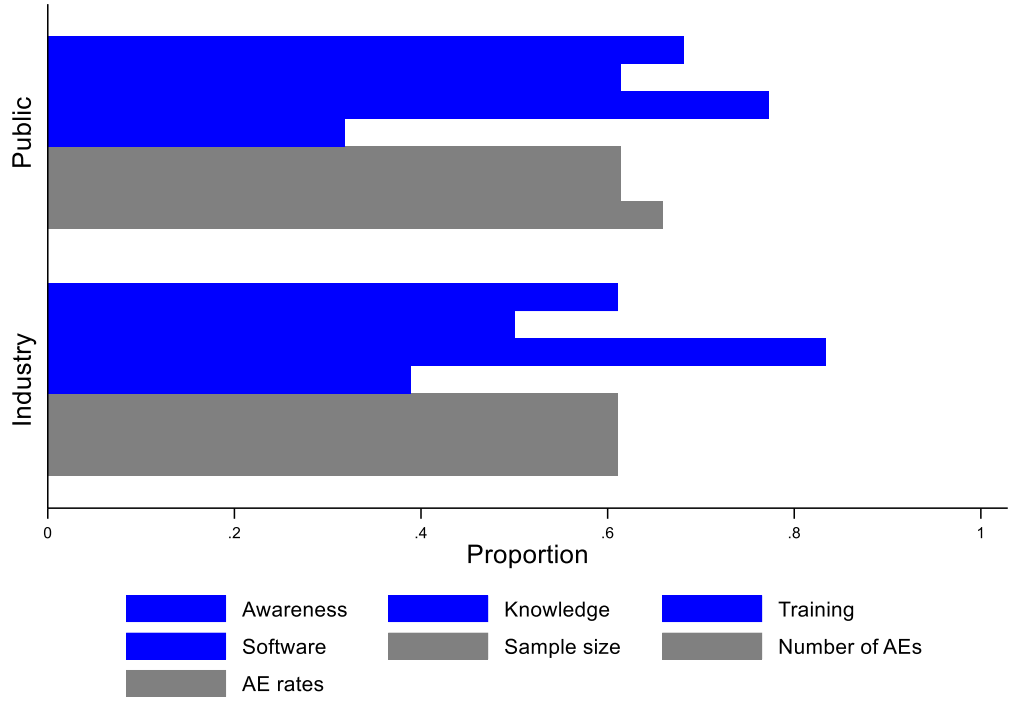


Acronyms: AE: adverse event; CI: chief investigator

Participants were asked to indicate if any of the following had been a barrier to the analysis they performed on emerging harms: lack of training opportunities (indicated by 79% of participants); lack of awareness of appropriate methods (indicated by 66% of participants); lack of knowledge to

implement appropriate methods (indicated by 58% of participants); a lack of statistical software/code to implement appropriate methods (indicated by 34% of participants); and trial characteristics, including trial sample size (61%), number of different events experienced (61%) and event rates (65%) (table 4.8 and figure 4.5).

Figure 4.5: Barriers when analysing emerging harm outcomes by employment sector (strongly agree and always agree responses combined)

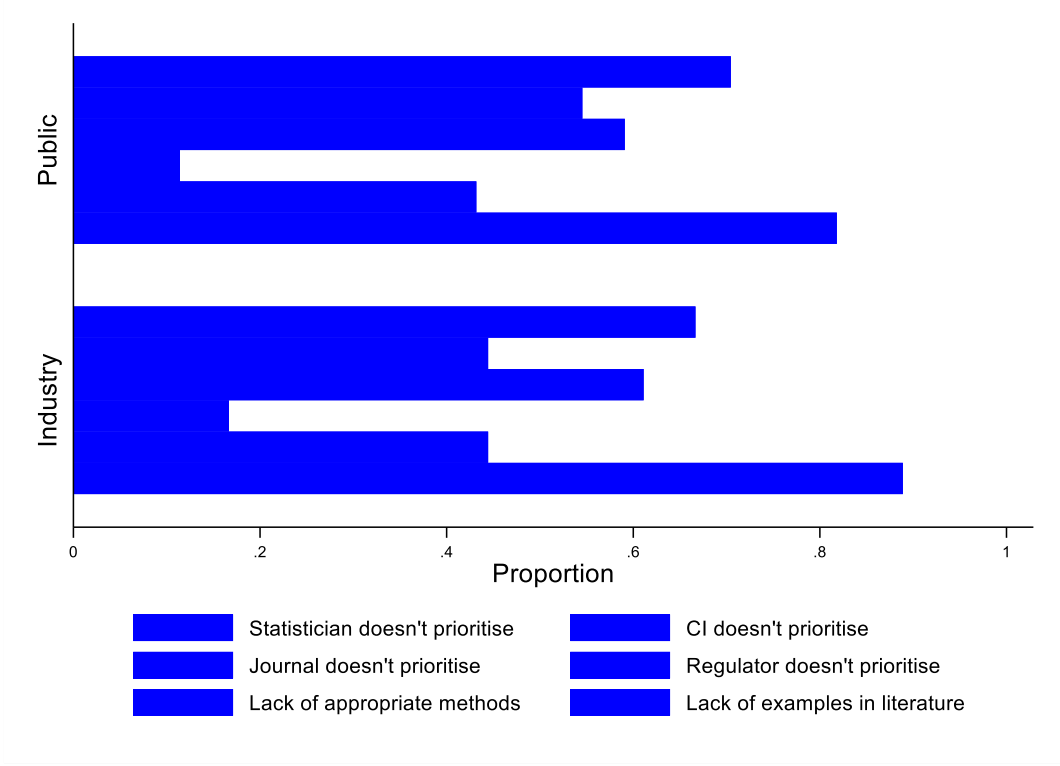


Acronyms: AE: adverse event

Participants were asked to indicate how strongly they agreed with a series of statements about the analysis of emerging harm outcomes. The majority of participants (84%) agreed or strongly agreed that there are a lack of examples for appropriate analysis methods in the applied literature. Less than half of participants (44%) agreed or strongly agreed that there are a lack of appropriate analysis methods. Over half of participants agreed or strongly agreed that statisticians (69%), journals (60%) and chief-investigators (52%) do not give data on harms the same priority as the primary efficacy outcome. Thirteen percent of participants agreed or strongly agreed that regulators do not prioritise

data on harms but nearly a quarter (24%) felt unable to comment on regulators priorities (table 4.9 and figure 4.6).

Figure 4.6: Opinions about analysis of emerging harm outcomes by employment sector (agreed and strongly agreed categories combined)



Acronyms: CI: chief investigator

Table 4.7: Influences the analysis performed by employment sector and overall[§§§§]

		Public (N=45)*		Industry (N=18)		Overall (N=63)	
INFLUENCE		**n**	**%**	**n**	**%**	**n**	**%**
Statistician prefers simple approaches e.g. tables of frequencies and percentages	Never	3	6.7	1	5.6	**4**	**6.3**
	Not very often	10	22.2	6	33.3	**16**	**25.4**
	Often	26	57.8	10	55.6	**36**	**57.1**
	Always	6	13.3	1	5.6	**7**	**11.1**
Chief investigator prefers simple approaches e.g. tables of frequencies and percentages	Never	1	2.2	0	0	**1**	**1.6**
	Not very often	8	17.8	2	11.1	**10**	**15.9**
	Often	23	51.1	12	66.7	**35**	**55.6**
	Always	12	26.7	2	11.1	**14**	**22.2**
	Don't know	1	2.2	2	11.1	**3**	**4.8**
Journal prefers simple approaches e.g. tables of frequencies and percentages	Never	3	6.7	2	11.1	**5**	**7.9**
	Not very often	9	20	5	27.8	**14**	**22.2**
	Often	21	46.7	4	22.2	**25**	**39.7**
	Always	4	8.9	1	5.6	**5**	**7.9**
	Don't know	8	17.8	6	33.3	**14**	**22.2**
Regulator prefers simple approaches e.g. tables of frequencies and percentages	Never	1	2.2	0	0	**1**	**1.6**
	Not very often	8	17.8	4	22.2	**12**	**19**
	Often	15	33.3	11	61.1	**26**	**41.3**
	Always	3	6.7	1	5.6	**4**	**6.3**
	Don't know	18	40	2	11.1	**20**	**31.7**
Trial sample size	Never	5	11.1	1	5.6	**6**	**9.5**
	Not very often	7	15.6	1	5.6	**8**	**12.7**
	Often	24	53.3	9	50	**33**	**52.4**
	Always	7	15.6	6	33.3	**13**	**20.6**
	Don't know	2	4.4	1	5.6	**3**	**4.8**
The number of different events experienced across the trial	Never	4	8.9	2	11.1	**6**	**9.5**
	Not very often	9	20	5	27.8	**14**	**22.2**
	Often	25	55.6	6	33.3	**31**	**49.2**
	Always	6	13.3	4	22.2	**10**	**15.9**
	Don't know	1	2.2	1	5.6	**2**	**3.2**
Event rates	Never	1	2.2	2	11.1	**3**	**4.8**
	Not very often	8	17.8	2	11.1	**10**	**15.9**
	Often	28	62.2	9	50	**37**	**58.7**
	Always	7	15.6	4	22.2	**11**	**17.5**
	Don't know	1	2.2	1	5.6	**2**	**3.2**

*One participant failed to answer the questions on influence

[§§§§] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.8: Barriers when analysing emerging harm outcomes by employment sector and overall[*****]

		Public (N=44)*		Industry (N=18)		Overall (N=62)	
BARRIERS		**n**	**%**	**n**	**%**	**n**	**%**
Lack of awareness of appropriate methods	Strongly disagree	1	2.3	0	0	**1**	**1.6**
	Disagree	10	22.7	7	38.9	**17**	**27.4**
	Agree	27	61.4	9	50	**36**	**58.1**
	Strongly agree	3	6.8	2	11.1	**5**	**8.1**
	Don't know	3	6.8	0	0	**3**	**4.8**
Lack of knowledge to implement appropriate methods	Strongly disagree	2	4.5	0	0	**2**	**3.2**
	Disagree	13	29.5	8	44.4	**21**	**33.9**
	Agree	25	56.8	7	38.9	**32**	**51.6**
	Strongly agree	2	4.5	2	11.1	**4**	**6.5**
	Don't know	2	4.5	1	5.6	**3**	**4.8**
Lack of training opportunities to learn what methods are appropriate	Strongly disagree	2	4.5	0	0	**2**	**3.2**
	Disagree	5	11.4	3	16.7	**8**	**12.9**
	Agree	28	63.6	13	72.2	**41**	**66.1**
	Strongly agree	6	13.6	2	11.1	**8**	**12.9**
	Don't know	3	6.8	0	0	**3**	**4.8**
Lack of statistical software/code to implement appropriate methods	Strongly disagree	5	11.4	1	5.6	**6**	**9.7**
	Disagree	16	36.4	10	55.6	**26**	**41.9**
	Agree	14	31.8	6	33.3	**20**	**32.3**
	Strongly agree	0	0	1	5.6	**1**	**1.6**
	Don't know	9	20.5	0	0	**9**	**14.5**
Trial sample size	Strongly disagree	2	4.5	3	16.7	**5**	**8.1**
	Disagree	11	25	4	22.2	**15**	**24.2**
	Agree	19	43.2	6	33.3	**25**	**40.3**
	Strongly agree	8	18.2	5	27.8	**13**	**21**
	Don't know	4	9.1	0	0	**4**	**6.5**
The number of different events experienced across the trial	Strongly disagree	1	2.3	3	16.7	**4**	**6.5**
	Disagree	14	31.8	4	22.2	**18**	**29**
	Agree	20	45.5	7	38.9	**27**	**43.5**
	Strongly agree	7	15.9	4	22.2	**11**	**17.7**
	Don't know	2	4.5	0	0	**2**	**3.2**
Event rates	Strongly disagree	1	2.3	4	22.2	**5**	**8.1**
	Disagree	13	29.5	3	16.7	**16**	**25.8**
	Agree	24	54.5	6	33.3	**30**	**48.4**
	Strongly agree	5	11.4	5	27.8	**10**	**16.1**
	Don't know	1	2.3	0	0	**1**	**1.6**

*Two participants failed to answer the questions on barriers

[*****] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.9: Opinions regarding analysis of emerging harms by employment sector and overall[†††††]

		Public (N=44)*		Industry (N=18)		Overall (N=62)	
OPINIONS		**n**	**%**	**n**	**%**	**n**	**%**
Statisticians don't give harm outcomes the same priority as the primary efficacy outcome	Strongly disagree	3	6.8	3	16.7	**6**	**9.7**
	Disagree	10	22.7	3	16.7	**13**	**21**
	Agree	22	50	7	38.9	**29**	**46.8**
	Strongly agree	9	20.5	5	27.8	**14**	**22.6**
Chief investigators don't give harm outcomes the same priority as the primary efficacy outcome	Strongly disagree	3	6.8	1	5.6	**4**	**6.5**
	Disagree	17	38.6	6	33.3	**23**	**37.1**
	Agree	21	47.7	6	33.3	**27**	**43.5**
	Strongly agree	3	6.8	2	11.1	**5**	**8.1**
	Don't know	0	0	3	16.7	**3**	**4.8**
Journals don't give harm outcomes the same priority as the primary efficacy outcome	Strongly disagree	4	9.1	1	5.6	**5**	**8.1**
	Disagree	8	18.2	3	16.7	**11**	**17.7**
	Agree	20	45.5	7	38.9	**27**	**43.5**
	Strongly agree	6	13.6	4	22.2	**10**	**16.1**
	Don't know	6	13.6	3	16.7	**9**	**14.5**
Regulators don't give harm outcomes the same priority as the primary efficacy outcome	Strongly disagree	9	20.5	6	33.3	**15**	**24.2**
	Disagree	16	36.4	8	44.4	**24**	**38.7**
	Agree	4	9.1	2	11.1	**6**	**9.7**
	Strongly agree	1	2.3	1	5.6	**2**	**3.2**
	Don't know	14	31.8	1	5.6	**15**	**24.2**
There are a lack of appropriate analysis methods	Strongly disagree	1	2.3	2	11.1	**3**	**4.8**
	Disagree	14	31.8	6	33.3	**20**	**32.3**
	Agree	15	34.1	6	33.3	**21**	**33.9**
	Strongly agree	4	9.1	2	11.1	**6**	**9.7**
	Don't know	10	22.7	2	11.1	**12**	**19.4**
There are a lack of examples of the use of appropriate analysis methods in the applied literature	Strongly disagree	1	2.3	0	0	**1**	**1.6**
	Disagree	4	9.1	1	5.6	**5**	**8.1**
	Agree	28	63.6	12	66.7	**40**	**64.5**
	Strongly agree	8	18.2	4	22.2	**12**	**19.4**
	Don't know	3	6.8	1	5.6	**4**	**6.5**

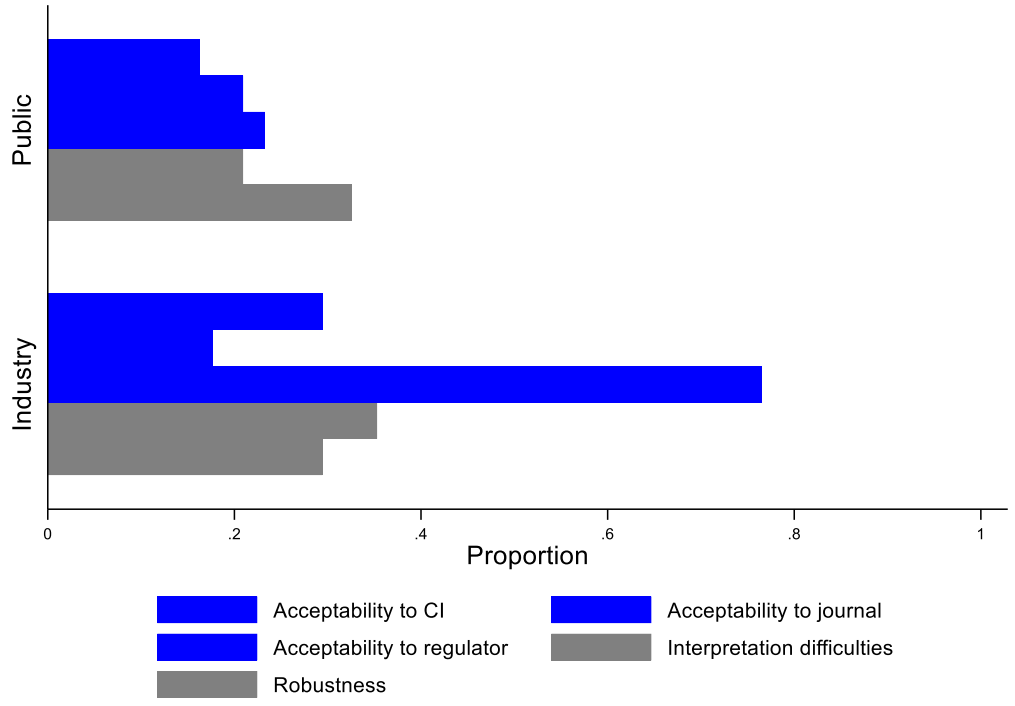
*Two participants failed to answer the questions on concerns

[†††††] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Concerns and solutions

Participants were asked to think about a series of statements on available methods and indicate their level of concern for each. The statement supported by most, was a concern for the acceptability of methods to regulators, with 38% of participants moderately to extremely concerned. Examining the results by sector indicated a substantial difference with 23% of public sector participants being concerned about the acceptability of methods to regulators compared to 77% of industry participants. Twenty percent of participants agreed they were moderately to extremely concerned about the acceptability of methods to the chief investigator and journals and 32% indicated they agreed that the robustness of available methods was a concern. These results were broadly similar across sector (table 4.10 and figure 4.7).

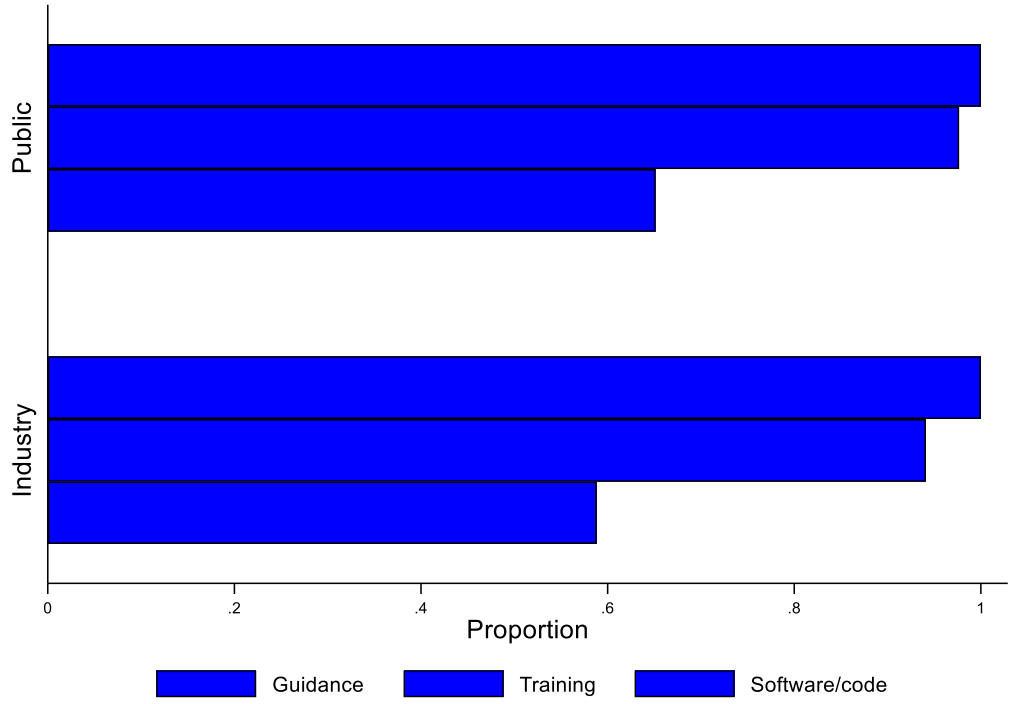
Figure 4.7: Concerns about current analysis practice for emerging harm outcomes by employment sector (moderately to extremely concerned categories combined)



Acronyms: CI: chief investigator

Participants were asked to indicate how much they agreed with a series of prespecified solutions that could potentially support a change in analysis practices for emerging harm outcomes. One-hundred percent of participants agreed or strongly agreed that guidance on appropriate methods for the analysis of harms is needed, 97% agreed or strongly agreed that training specifically for the analysis of harms is needed, and 63% indicated that they agreed that new software or code is needed. Results were similar across sector (table 4.11 and figure 4.8).

Figure 4.8: Solutions to support a change in analysis practices for emerging harm outcomes by employment sector (strongly agree and agree categories combined)



Participants were also asked to provide their own thoughts on potential solutions to support change in analysis practices for emerging harms and thirty-two percent of participants provided further feedback. Suggestions included improved standards or calls for change from journals, registries and regulators (n=8); development of guidance, education and engaging with the medical community (n=9); and proposed analysis practices to be adopted (n=3). Table 4.12 provides the participant comments attributed to each group.

Table 4.10: Concerns regarding the analysis of emerging harm outcomes by employment sector and overall ‡‡‡‡‡

		Public (N=43)*		Industry (N=17)†		Overall (N=60)	
CONCERNS		**n**	**%**	**n**	**%**	**n**	**%**
Difficulties in interpreting the results/output	Not at all concerned	4	9.3	4	23.5	**8**	**13.3**
	Slightly concerned	16	37.2	4	23.5	**20**	**33.3**
	Somewhat concerned	14	32.6	3	17.6	**17**	**28.3**
	Moderately concerned	5	11.6	5	29.4	**10**	**16.7**
	Extremely concerned	4	9.3	1	5.9	**5**	**8.3**
Robustness of methods	Not at all concerned	1	2.3	3	17.6	**4**	**6.7**
	Slightly concerned	15	34.9	3	17.6	**18**	**30**
	Somewhat concerned	13	30.2	6	35.3	**19**	**31.7**
	Moderately concerned	11	25.6	2	11.8	**13**	**21.7**
	Extremely concerned	3	7	3	17.6	**6**	**10**
Acceptability of methods to chief investigator	Not at all concerned	13	30.2	4	23.5	**17**	**28.3**
	Slightly concerned	13	30.2	2	11.8	**15**	**25**
	Somewhat concerned	10	23.3	6	35.3	**16**	**26.7**
	Moderately concerned	5	11.6	0	0	**5**	**8.3**
	Extremely concerned	2	4.7	5	29.4	**7**	**11.7**
Acceptability of methods to journal	Not at all concerned	11	25.6	5	29.4	**16**	**26.7**
	Slightly concerned	17	39.5	6	35.3	**23**	**38.3**
	Somewhat concerned	6	14	3	17.6	**9**	**15**
	Moderately concerned	8	18.6	3	17.6	**11**	**18.3**
	Extremely concerned	1	2.3	0	0	**1**	**1.7**
Acceptability of methods to regulator	Not at all to concerned	10	23.3	2	11.8	**12**	**20**
	Slightly concerned	13	30.2	1	5.9	**14**	**23.3**
	Somewhat concerned	10	23.3	1	5.9	**11**	**18.3**
	Moderately concerned	7	16.3	7	41.2	**14**	**23.3**
	Extremely concerned	3	7	6	35.3	**9**	**15**

*Three public sector participants failed to answer the questions about concerns around available methods for the analysis of emerging harms

† One industry sector participants failed to answer the questions about concerns around available methods for the analysis of emerging harms

‡‡‡‡‡ *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.11: Solutions to support a change in the analysis of emerging harms by employment sector and overall §§§§§

		Public (N=43)*		Industry (N=17)†		Overall (N=60)	
Solutions to support a change		**n**	**%**	**n**	**%**	**n**	**%**
Software/code development is needed	Strongly disagree	1	2.3	0	0	**1**	**1.7**
	Disagree	8	18.6	6	35.3	**14**	**23.3**
	Agree	20	46.5	7	41.2	**27**	**45**
	Strongly agree	8	18.6	3	17.6	**11**	**18.3**
	Don't know	6	14	1	5.9	**7**	**11.7**
Training specifically for the analysis of harm outcomes is needed	Strongly disagree	0	0.0	0	0.0	**0**	**0.0**
	Disagree	1	2.3	1	5.9	**2**	**3.3**
	Agree	30	69.8	11	64.7	**41**	**68.3**
	Strongly agree	12	27.9	5	29.4	**17**	**28.3**
Guidance on appropriate analysis methods for the analysis of harms is needed e.g. case studies and tutorials in open access journals	Strongly disagree	0	0.0	0	0.0	**0**	**0.0**
	Disagree	0	0.0	0	0.0	**0**	**0.0**
	Agree	24	55.8	8	47.1	**32**	**53.3**
	Strongly agree	19	44.2	9	52.9	**28**	**46.7**
Are there any other solutions in addition to those stated above that would support a change in analysis practices for harm outcomes?	No	34	79.1	7	41.2	**41**	**68.3**
	Yes	9	20.9	10	58.8	**19**	**31.7**

*Three public sector participants failed to answer the questions about solutions to support a change in analysis practices for emerging harms

† One industry sector participants failed to answer the questions about solutions to support a change in analysis practices for emerging harms

§§§§§ *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

Table 4.12: Classification of participants' comments on solutions to support change in analysis practices for emerging harm outcomes[******]

Classification of solutions to support a change in analysis practices for emerging harm outcomes	Participant comment	Sector
1. Improved standards or calls for changes from journals, registries and regulators	*"Influencing journals to pay more attention to this"*	Public
	"…we presented incidences because they represented a fairer picture due to differential follow-up and repeated incidences per person. The reviewer and the editor said they prefer proportions and don't understand what we presented. I explained in lay terms and pushed back their request because it was flawed. This shows that Statisticians can defend a certain position and educate others even if they have their own preferences. *Regulatory repositories/registries such as EUDRACT has a fixed format of presenting results so you have to go with what is required even though you know it's flawed in certain situation. Flexibility of such registries is very important to allow people to present both proportions and incidences where appropriate."*	Public
	"Asked by the authorities"	Industry
	"Strong regulatory direction is always good for changing practices within the industry!"	Industry
	"engaging the … regulators"	Industry
	"*The biggest driver of a change in behaviour is usually a regulator requesting it*."	Industry
	"Regulators to be more demanding in analytical approaches, don't require more than summaries. That's far removed from discussions on efficacy"	Industry
	"Would have to be able to upload the results to EUDRACT for CTIMPS."	Public

[******] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

2. Development of guidance, education and engaging with the medical community	*"Best practice guidance although that would depend on trial type and phase, sample size, whether only SAEs/related AEs are being captured/important, ... particularly important to reflect on complex interventions vs CTIMP, etc"*	Public
	"There needs to be consensus that a change is needed. What are the issues in current AE reporting? There needs to be better guidance re collection of AE data. Can we collect it in a more robust way? We need to differentiate between examining pre-specified hypotheses and trying to identify issues we don't know about (eg in early phase trials). We need agreement re standards for different phases and types of trials (eg Phase 1 vs Phase 4, explanatory vs pragmatic, regulatory submissions vs investigator led exploratory trials on marketed products)"	
	"Published case studies"	Industry
	"engaging the medical community …. and Better education on the pros of using proper stats methodology. If the benefits of using effective statistical analysis methods over frequencies and percentages can be demonstrated, there might be more interest"	Industry
	"demonstration of the benefits of these methods over existing ones, and when they are appropriate"	Public
	"Open discussions with clinical community (e.g. open forums, etc) on alternative methods to avoid them being scared off"	Industry
	More focus on safety analyses in the E9 addendum"	Industry
	"Application of CONSORT harms"	Public
	"*Evolution of standard reporting requirements in clinical trials (ICH E3, and maybe CONSORT Statement ?)"*	Industry
3. Proposed analysis practices to be adopted	*"IPD meta analysis of AEs"*	Public
	"In addition to 'methods' there perhaps need to be discussion about populations/datasets on which to base AE analyses."	Public
	"Inferential analysis based on small numbers of adverse events, but of great influence on the patient health."	Industry

Participants were also given the opportunity to express any other thoughts they had on current practice for the analysis of harm outcomes. Thirty percent of participants took this opportunity.

Opinions expressed covered the following themes: minimum summary information that participants would expect to be reported for emerging harm outcomes such as "*numbers and percentages*" (n=2); changes to analysis practice that could or have already been made such as "*use of graphical methods*" (n=8); concerns about the quality and collection of data on emerging harm outcomes (n=3); and general comments and criticisms about current analysis and reporting practices for harm outcomes (n=4). Table 4.13 provides the participant comments attributed to each theme.

Table 4.13: Classification of participants' general comments raised regarding analysis practices for emerging harm outcomes[††††††]

Classification of suggestions raised for analysis of emerging harm outcomes	Participant comment	Sector
1. Minimum summary information participants would expect to be reported for emerging harm outcomes	*"Different analysis approach are useful for interpretation when reporting AEs/SAEs. As a starting point, I would like to know the numbers and proportions experiencing at least one SAE by group, between group differences with uncertainty. In addition, I would like to know the incidences per group and incidence rate ratio with uncertainty. The later is not always necessary depending on the situation.."*	Public
	"I think in general reporting numbers and percentages is appropriate. The argument being that, if we were clinicians or patients we would want to know what is the chances of me having this event and how bad will it get, which is essentially what the frequency tables give you."	Public
2. Changes that could or have been made to analysis practice	*"No best practice guidance although revised CONSORT does help remind of importance of AE reporting"*	Public
	"There was a great talk at SCT 2017 on using graphical methods to summarise AEs and I have been trying to implement graphical methods to summarise the many dimensions of AE reporting as a way forward"	Public
	"Use of graphical methods in reporting to compare treatments ought to be standard, as per BMJ article. They are easy enough to apply… *…The format of the source data, typically free text, is a pain to code into MedDRA. Methods to make this easier would be very valuable: some sort of AI machine learning maybe?...* *…Meta-analysis should be very important to apply to safety data, given how under-powered individual trials may be for safety comparisons. Finding tools to automate, maybe using results entered on EudraCT might be an idea."*	Public
	"We have increased our use of graphics. I find benefit risk plots a very powerful way of summarising data. Allows key efficacy and safety to be displayed on one page and is a really useful summary of a drug's profile."	Industry
	"*Current practice will need to turn to methods of detecting signals as real-time data come from trials."*	Industry

[††††††] *Reprinted with minor modifications from Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875 under a CC BY 4.0 License: http://creativecommons.org/licenses/by/4.0/*

	"Signal detection method"	Public
	"I'm interested in knowing more about risk factors of occurrence of serious or really frequent AEs of chemotherapies, beyond receiving protocol x."	Industry
	"… not many medical leads understand statistical analysis of AEs or count or rate data and only insist on percentages and frequencies. Better methods exist but are not utilised due to lack of knowledge of PIs or medical advisors"	
3. Concerns about the quality and collection of data on emerging harm outcomes	*"This definitely gets overlooked. I always worry about how systematically the data have been collected too as well as the validity of lumping very different events together in the same analysis."*	Public
	"I think a big factor in what analysis we choose is how the data is collected. If the data is not detailed enough some only simple methods may be appropriate - this has often been my feeling when analysing our data. this may change in current/future trials as we are changing how we collect some AE data"	Public
	"*My concerns start with the quality of AE data collected. Is it complete? Is it robust? There is recall bias, variability between centres, investigators etc. There may also be variability with respect to coding. We all have experience of stating up front what should NOT be recorded as AE, to see such things recorded multiple times. One of my major concerns is the listing of AEs each with associated p-values (obviously the CI would insist on this and not the statistician). Completely meaningless as it doesn't take into account sample size, rate, number of events within a participants, severity of event etc etc. Also of concern is the use of more complex methodologies on such data as it implies that the data are robust. I think that the simple approach is often acceptable so long as the data are presented in different ways (see Q16). The main issue is about defining what you are trying to detect from the collection of AE data. If we can do this better then perhaps additional required methodology will come."*	Public
4. General comments and criticisms about current analysis and reporting practices for emerging harm outcomes	*"Somewhat arbitrary grouping of AEs. Not always clear whether numbers are subjects or events are presented in published papers."*	Public
	"In my 8.5 years of experience I have not seen many studies where they have spoken much about AE data analysis."	Industry
	"People do the most powerful test for efficacy - no barrel goes unscraped - and the least powerful for safety"	Public
	"It can be improved!"	Industry

4.5 Discussion

4.5.1 Summary of findings

This survey of statisticians from the UK public and private sectors has established a more detailed picture of clinical trial statisticians' analysis practices for harm outcomes, specifically emerging harm outcomes. This builds on the work of chapter two which evaluated analysis practices for harm outcomes reported in journal articles.[45] It identified that results presented in journal articles are likely to reflect a subset of results produced by statisticians and that more thought is likely being given to this area than published results reflect. Results also suggested that acceptability of methods to chief investigators (e.g. there's a need to "*engage the medical community")* and specifically in industry, regulators (e.g. "*Strong regulatory direction is always good for changing practices within the industry!")*, and in academia, journals (e.g. "*Influencing journals to pay more attention to this"),* preference for simple approaches could one be one possible reason for a lack of progress. It also helped identify priorities and concerns regarding analysis of harm outcomes, some of which will be addressed in this thesis.

Specifically, in later chapters I will look at ways to incorporate time into the analysis of emerging events to detect signals for potential ADRs which was supported in free text comments such as "*In characterising safety signals I have used [time-to-event analysis]"* and comments highlighting the importance of using an appropriate statistic due to "*due to differential follow-up".* I will also explore how visualisations could be used to improve the communication of information regarding harm outcomes, which was endorsed in free text comments such as "*Use of graphical methods in reporting to compare treatments ought to be standard",* as well as the unanimous endorsement that guidance is needed to support change. Ultimately, with the aim of exploring and developing methods to improve the suboptimal analysis practices for harm outcomes identified here and the earlier review described in chapter two.

Survey results were broadly similar across public and industry sectors with striking similarities in terms of analysis practices, barriers when analysing harms, opinions on priorities and unanimous support across sectors that guidance and training on appropriate analysis for harm outcomes is needed. Differences of note included the greater use of hypothesis tests and 95% confidence intervals as a means to compare event rates between treatment groups by public sector participants, in-line with the findings of the review of journal articles presented in chapter two. In addition, industry participants more often believed that regulators preferred simple approaches for the analysis of harm outcomes, and more industry participants were concerned about the acceptability of methods to regulators.

4.5.2 How does self-reported practice compare to that reported in the literature?

Reviews of published articles (including the results in chapter two) have found that between 1% and 9% of articles report both the number of participants with at least one event and the total number of events.[45, 56, 75] This is in contrast to the substantially higher figure of 75% of survey participants that indicated they produce tables with both. Reporting information on the number of events rather than just those with at least one event can give a better summary of impact on patients' quality-of-life and is important information to report. However, reviews identified only 6% to 7% of published articles report this information.[45, 79] Reported use of between group statistics such as risk differences or risk ratios were greater than what was reported in journal articles, where our review identified no more than 10% of articles reported such summaries. Survey participants' reported use of 95% confidence intervals that were comparable to that reported in journal articles (22% compared to 20%) but use of hypothesis tests was less than what was found in journal articles (32% compared to a range of 38% to 47% across reviews of journal articles).[45, 56, 75] Reasons for the identified differences between reported results (in journal articles) and reported practices (as per survey responses) could include journals editors requesting such analyses be presented to compare groups,

or at the request of the chief investigator, both of which are supported by survey responses indicating a "perceived" preference for simple approaches from both groups. It could also be that the survey participants were restricted to those working in CTUs and industry and are perhaps not fully representative of those undertaking and reporting clinical trial results. It is important to note that reasons for these disparities were not directly sought from participants.

Results of chapter three revealed that there are many methods that have been specifically proposed for the analysis of harm outcomes in RCTs.[130] Survey responses indicated that there was a moderate level of awareness of these methods amongst clinical trial statisticians (40%) but in line with the review of journal articles described in chapter two uptake was minimal (13%).[208, 209] The results of this survey are also closely aligned with the results of a 2016 survey of industry statisticians and clinical safety scientists, that indicated a reliance on the use of descriptive statistics and frequentist approaches when analysing harm outcomes.[224] Whilst I do not endorse null hypothesis testing for harms, especially for unspecified emerging events, there is still a need for inferential statistics to enable between treatment group comparisons and less reliance on simple descriptive summaries.

Responses in free text comments indicated that a similar proportion of participants used graphs to present information on harm outcomes as identified in the review of journal articles (9% vs 12%).[45] In contrast, the 2016 survey of industry statisticians found that 37% of participants used graphs when analysing and reporting on harm outcomes.[224] This disparity could reflect the use of graphical approaches for internal reports, as advocated in work by Davis et al. and Furey et al., rather than in the dissemination of results to the wider trials community, which would not be captured by a review of journal articles.[225, 226] It is perhaps also unsurprising that a survey targeting industry statisticians alone indicated a wider use of graphics than respondents to this present survey, that includes both public and industry statisticians, given the widespread investment in data visualisation by industry

and the expanding in-house expertise. This is evidenced by the emergence of departments dedicated to data visualisations within many pharmaceuticals and development of bespoke graphical analysis software and tools. This is also supported by the increasing literature published by industry in this field, with the results of chapter three indicating novel visualisations were predominantly published by authors working in industry.[130, 227, 228] There is potentially much to be learned from industry in their attitude towards incorporation of visualisations for clinical trial reports and the complex nature of harm outcomes are likely to be an area that could benefit from greater adoption of visualisations.

4.5.3 Priorities for future work as highlighted by research participants

This and earlier chapters have identified the prevalence of suboptimal analysis practices and highlighted the need for change. In addition, as part of early dissemination and feedback activities a workshop was held following completion of the survey as part of the UKCRC CTU network's biannual statisticians' operations group. This allowed richer information to be gathered from CTU statisticians with discussions focusing on priorities for improvement and ideas for how such improvements could be brought about. When survey participants were asked about different potential solutions to support a change in analysis practices for harms, training and guidance for statisticians and trialists about appropriate methods were both overwhelmingly supported. This support was reiterated by workshop attendees e.g. "*demonstration of the benefits of these methods over existing ones, and when they are appropriate*" and *"best practice guidance although that would depend on trial type and phase".* Whilst there already exist guidelines on how harm outcomes should be reported e.g. the harms extension to CONSORT, the pharmaceutical industry standard from SPERT, and the joint pharmaceutical/journal editor collaboration guidance on reporting of harm data in journal articles; recommendations for analysing harm outcomes are limited.[7, 28, 34] In addition, there is evidence to suggest that adherence to existing reporting guidelines is suboptimal. For example, a recent review of adherence to the CONSORT harms extension has shown that improvements since its publication

in 2004 have been limited.[80] In addition, the review of analysis practices for harm outcomes reported in chapter two indicates uptake of suggestions from both Lineberry et al. and Crowe et al. such as "*reporting CIs around absolute risk differences*" and to "*include both the number of events (per person time) and the number of patients experiencing the event*" has been limited.[45, 58, 75, 78, 79] Others have argued that existing guidelines do not go far enough, failing to account for the complex nature of data collected on harm outcomes and that a "*lack of standardized guidelines for safety data analysis … may limit the ability to draw rich conclusions about the safety of the investigational product*".[123] In other fields guidance has taken the form of tutorial papers and case studies detailing examples of appropriate analysis e.g. the tutorial paper from Morris et al. on designing, performing, analysing, and reporting simulation studies; and the practical guide from Cro et al. on using controlled multiple imputation in clinical trials with missing outcome data that includes code for implementaion.[229, 230] It is believed that such resources help to achieve wider adoption of methods and could ultimately lead to improvements in analysis practices. Development of such resources for harm outcomes was raised by survey participants e.g. a need for *"published case studies"* and highlighted as a priority by workshop attendees. Awareness of good practices and alternative methods are essential to harness change. Guidance, tutorial papers and training can be useful to increase knowledge and arm trialists with the skills to implement 'good practice' methods, but wide dissemination and promotion of such resources is essential if practices are to change. The failure of CONSORT harms to result in changes to reporting practices could in part be due to the lack of endorsement by journals and a resulting lack of awareness of its existence by the wider trials community.[115] Journals and regulators hold an unrivalled position in their ability to promote good practice and influence statisticians and trialists practice through policy change. A universal journal initiative endorsing existing guidelines, that could be achieved through the mandatory submission of the CONSORT harms checklist has been proposed as one simple, initial step towards change.[117]

Participant feedback indicated that it is not only clinical trial statisticians that need to be persuaded that analysis practices need to change, but that engagement and endorsement by the clinical trial community is also essential e.g. *"Open discussions with clinical community (e.g. open forums, etc.) on alternative methods to avoid them being scared off"* and *"Better methods exist but are not utilised due to lack of knowledge of PIs or medical advisors".* Engagement and feedback from clinical researchers will be sought on later development work to help with this.

Survey participants endorsed improved analysis for harm outcomes and mentioned exploration of time-to-event analyses, data-visualisations, and Bayesian methods to achieve this. In this thesis, the aim is to explore adoption of existing or development of more appropriate methods for the analysis of harm outcomes, which may help identify signals for potential ADRs and enable a clearer harm profile to be presented. Such an approach is supported by the earlier findings of Colopy et al. who concluded that statisticians should help "*minimize the submission of uninformative and uninterpretable reports*" and thus present more informative information regarding likely drug-event relationships, and this was also endorsed by workshop attendees.[224]

Survey participants and workshop attendees voiced concerns about the quality and reporting of data on harm outcomes from RCTs. As with any outcome, if the data collection is not robust the analysis approach used is redundant as the results will not be accurate. Procedures should be put in place at the trial design stage to mitigate problems with data collection, including, for example, development of validated methods for collection and clear, standardised instructions and training for those involved in the detection and collection of data, as well as monitoring procedures to ensure the accuracy of what is recorded.[56, 85] The quality of harms data is not the focus of this thesis and will not be explored in detail but it is important to be mindful of these issues when summarising and making conclusions from any data. I have outlined some key components to ensure transparent reporting of

harm outcomes in chapter 2 table 2.14, which should enable an assessment of the quality of the data collected on harm outcomes. These recommendations are discussed further in the final chapter of this thesis.

4.5.4 Strengths and limitations

A high response rate for the survey was achieved through support of the UKCRC CTU statistics operation group and utilisation of personal contacts. Participants recruited via the open platform were self-selected, therefore, there is a possibility that these participants had an increased interest in the analysis of harm outcomes and might not represent a typical sample from the clinical trial community. It was not feasible to collect information on non-responders, thus it is not possible to characterise any potentially relevant differences that could affect the generalisability of these results. Nonetheless, this survey provides valuable insights into practices and perceptions from senior clinical trial statisticians with a wealth of experience and has identified key starting points to focus on to support a change to improve analysis practices for harm outcomes. In addition, the workshop attendees who represented more of a general interest group echoed many of the opinions raised in the survey.

4.5.5 Plans for future work

Considering survey results, work to take forward in the immediate term include:

i) Development of guidance for appropriate statistical methods, in collaboration with key stakeholders including CTU and industry statisticians, clinical researchers and journal representatives. Including development of software for implementation as necessary.

ii) Development of case studies and tutorials to promote more objective analysis of harm outcomes.

4.5.6 Conclusions

The result of this survey revealed that the results presented in journal articles are likely to reflect a subset of both the results produced by statisticians and their wishes. More thought is likely being given to this area by statisticians than published results reflect but key stakeholders (e.g. chief investigators, journal editors and regulators) are likely influencing observed practices, in part to remain consistent with historical practice. However, they also suggest that, analysis practices for harm outcomes in RCTs are sub-optimal and confirms that despite a moderate level of awareness of more sophisticated statistical methods for analysis of emerging harm outcomes, uptake is minimal. This research highlights that improvements are needed and that clinical trial statisticians require guidance on appropriate methods for analysis of harm outcomes with training to aid change and that engagement with the wider trial community is needed to ensure support for any changes. However, further research is still needed to identify the most appropriate statistical methods for analysis of harm outcomes from all those available.

5. Recommendations for visualising harms in RCT publications: a national consensus

5.1 Introduction

A well-designed graphic can effectively communicate a message to diverse audiences and help identify patterns in data that might otherwise be missed.[231] In 1983 Tufte stated, "*of all methods for analyzing and communicating statistical information, well-designed graphics are usually the simplest and at the same time the most powerful*".[232] In clinical trials, when analysing harms where there is an abundance of complex data, graphics can be potentially useful to help summarise harm profiles and identify potential ADRs. As Harrell says in his book on the principles of graph construction, "*graphical displays should make large datasets coherent*".[233] Trial reporting guidelines such as the CONSORT extension to harms, the 2016 recommendations to improve the reporting of harms from industry representatives and journal editors, a pharmaceutical industry standard from SPERT and guidance from regulators on statistical principles in clinical trials (ICH E9) encourage the use of visualisations for exploring harm outcomes.[6, 7, 28, 234] In addition, the review presented in chapter three along with the work from Amit et al., Cooper et al. and Chuang-Stein et al. demonstrated that there are an abundance of visualisations available for the analysis of harm outcomes but use in journal articles is limited.[130, 137, 139, 183] Results presented in chapter two found that only 12% of journal articles made use of visual summaries for harm data, and this finding was reinforced by the survey of UKCRC CTU and industry statisticians undertaken in 2019 reported in chapter four.[45, 216] However, an earlier, independent survey of industry statisticians from 2016 suggests in-house practice in this sector differs.[224] In addition, an overall appraisal of the quality of graphics published in high impact journals found over a third were rated as poor.[235] In the context of harm outcomes, evidence to date suggests that there remains a prevailing practice to present data in simple and often long tables of frequencies and percentages despite the advantages visualisations offer.[236]

Advances in computer software has improved trialists capability at producing visualisations but there is lack of guidance on what and how to visually display complex harm data in journal articles. The results presented in chapter four revealed that researchers want guidance on appropriate methods for the analysis of harm outcomes, as well as case studies detailing examples of use. There have also been independent calls from the statistical community for direction on "*how to decide which of many possible graphics to draw*".[216, 237] Following an invitation to present the work on visualisations for harms (identified in chapter three) to key stakeholders at a BMJ research editors meeting, it was suggested by the BMJ editors that I develop recommendations on which visualisations to use in journal publications. This also included advice from senior editors that whilst prescriptive guidance would help instigate a change this would need consensus from the community to ensure adoption. Therefore, with a range of visualisation options available and the increasing ease in which they can be implemented, I sought to lead a consensus to develop a set of recommendations to support researchers in their choice of visualisations for the presentation and analysis of harm data collected in clinical trials. This work was undertaken in collaboration with the UKCRC CTU statistics operations group.

5.2 Aims

The aim of the work in this chapter is to provide recommendations on which visualisations researchers should consider including in the publication of their main research findings. Specifically the objectives are:

1. To undertake a series of consensus meetings with experienced clinical trial statisticians from academia and industry:
    a. To review and critically appraise visualisations for the analysis of harm outcomes.
    b. To reach agreement on which graphics to endorse and identify any potential modifications to improve the graphic.

c. To develop a set of recommendations for each chosen visualisation, highlighting any limitations or cautions of use.

d. To highlight areas for further work, including the development of new visualisations and code for implementing existing plots where not already available.

2. To seek feedback from clinical investigators on the endorsed graphics, incorporating their feedback into the recommendations for use.
3. To produce a guidance document with examples to facilitate and promote use of visualisations for harms outcomes in journal articles and clinical trial reports.

5.3 Rationale for consensus approach

Recommendations produced via a consensus offer a number of advantages. They take into account a broader range of knowledge and experience and provide an opportunity for discussions and an exchange of viewpoints, which can lead to decisions that are more robust. In addition, group decisions are likely to carry more authority than those of an individual, which in turn leads to increased likelihood of adoption of any recommendations by the wider community.[238] The work described in this chapter was designed in line with the recommendations from the CONSORT group executive regarding development of reporting guidelines.[239] Whilst development of reporting guidelines are not completely aligned with the aims of this chapter, the CONSORT group recommendations offer guiding principles that serve as important considerations.

5.4 Methods

5.4.1 Study design

Consensus meetings

A series of three half-day virtual consensus meetings of predominantly UKCRC CTU and industry statisticians plus a health economist based at an academic population health department and a data graphics designer that sits on the multimedia team at the BMJ took place in July 2020. Guidance from the CONSORT group executive recommend a minimum of one day for developing guidelines. Given the need to run these sessions virtually due to the COVID-19 pandemic, the meetings were split over three half-day sessions over three consecutive weeks. This helped to ensure attendees were not overly fatigued and remained energised throughout each meeting, provided participants multiple opportunities to contribute (whilst attendees were encouraged to attend all three sessions it was not mandatory) and provided adequate time to cover the material and time for reflection between meetings.

Clinician interviews

Three clinical investigators experienced in clinical trials were invited to take part in one-to-one, virtual, semi-structured interviews to review the outputs of the consensus meetings. Each interview lasted approximately one and a half hours. This allowed sufficient time to cover all the material but was not too onerous for clinicians already balancing clinical work alongside research.

5.4.2 Sample size

Consensus meetings

As per the CONSORT recommendations for developing guidelines, an initial list of core participants (n=16) to invite was drawn up and availability of this core group was taken into consideration when scheduling the meetings.[239] Once initial scheduling was in place an open invitation to a statistical mailing list seeking expressions of interest was sought as described in section 5.4.3. An initial cap of twenty attendees was set due to limits on venue capacity; however, the change to virtual attendance allowed this to increase. Twenty-seven participants were invited to attend (n=16 from

core list and n=11 after expressing interest through the open invitation), which remained in line with the CONSORT recommendation to limit meetings to no more than 30 participants.

Clinician interviews

Input from clinical researchers was deemed vital as they lead clinical trials and are the ones whom statisticians need to commute initial research findings to. Clinicians are also the primary audience for clinical trial reports and journal articles, and act on the data presented to them, therefore it is vital that they understand and support any recommendations. Three clinicians with a variety of clinical expertise and trials research experience were identified and invited to participate. This was a pragmatic approach based on availability of clinicians and resources available.

5.4.3 Sampling and recruitment

Consensus meetings

Initial emails were sent to contacts in academia and industry with expertise in clinical trials and/or a known interest in visualisations (n=16). In addition, expressions of interest were sought from members of the UKCRC CTU statisticians' operations group and an advert was placed on the PSI visualisation special interest group homepage. Potential participants were asked to describe their applied experience of analysing trial data and whilst participants were not required to be experts in the analysis of harm outcomes, we asked that they describe their interest in this area.

Clinician interviews

Email invitations were sent to recommended clinicians and collaborators with a variety of clinical trials experience and one of whom holds a senior editorial position at a leading medical journal.

5.4.4 Participant eligibility

Consensus meetings

Researchers were eligible to participate if:

i) Their current role was as a statistician or other quantitative based health-care research role at a UKCRC CTU or academic institution or UK pharmaceutical company or CRO;
ii) They had extensive applied experience of analysing clinical trial data;
iii) A demonstrable interest in the analysis of harm outcomes in RCTs but not necessarily an expert in harms.

A data graphics designer that sits on the multimedia team at the BMJ was also invited to attend to provide expertise on visualisations in journal articles.

Clinician interviews

Clinicians were eligible to participate if:

i) They were or had been a chief investigator of a RCT or had extensive experience of undertaking RCTs;
ii) They had a strong interest in clinical trial methodology.

5.4.5 Meeting overview

Pre-meeting

Visualisations for analysing harm outcomes identified in the review described in chapter three were taken forward for evaluation at the consensus meeting. In addition, alternative visualisations that could be adapted to the harm setting brought to my attention through dissemination activities were also considered. Each visualisation was categorised according to the type of outcome they supported e.g. displaying multiple binary events, or single time-to-event outcomes etc. Prior to the first meeting individual plots were discussed with a data graphics designer who provided initial feedback and some sketches of alternative graphics. These alternative sketches were also included for consideration in the meeting and are named using the preface 'alternative' followed by the name of the plot they are an alternative to in the descriptions below. Examples of implementation of each

plot with detailed summaries and ideas for initial potential adaptations were shared with participants a week in advance of the first meeting (figures for initial consideration are presented as thumbnails in figures 5.1 to 5.6 and full images with summaries are included in figures A5.1 to A5.38 of appendices A5.2 to A5.7). This also included a call for participants to suggest any plots for consideration that may have been inadvertently omitted.

A draft framework for appraisal was developed taking into consideration work from Ballarini et al. who proposed a framework to assess the properties of graphics for subgroup analysis, principles for producing effective visualisations proposed by Gordon and Finch, and discussions with supervisors regarding the important components to communicate when analysing harm outcomes.[235, 240] The initial draft framework is displayed in table 5.1. This was discussed amongst participants in the first meeting and edited in real time based on feedback and group endorsement. Participants were asked to raise any items they thought had been omitted, whilst bearing in mind that the list should not become overly burdensome, as each visualisation would need to be appraised against it to highlight any items they thought were unnecessary or whether further clarification was needed for any of the items. Participants were also asked to consider whether item seven, which related to limitations in the number of events displayed should contribute to the overall score. This was because including it would mean that visualisations that could only display a limited number of events would be disadvantaged, but in some settings, this might be through design and is likely to be advantageous. Participants were happy with the proposed items and agreed that item seven should not be included in the calculation of the overall score. In addition, participants thought it was important that an item on adaptability to multi-arm or adaptive designs was included and that a ranking of the graphics would be useful. It was also decided that participants would not formally appraise the objective items (11-13) in table 5.1 but that they would considered each when appraising and discussing each of the plots.

The final criteria to appraise each of the graphics is included in table 5.2. It comprised of eight items related to plot content and presentation such as, does the plot clearly display an effect size for each event; does it clearly display a robust measure of uncertainty; and does it require supplementary data presentation. Two further items related to usage e.g. suitability for use in journal articles, interim or final analysis reports, and whether it was suitable for explanatory or exploratory analysis, where exploratory analysis was defined as visualisations suited to data exploration to help identify potential signals for ADRs and explanatory analysis was defined as visualisation suited to communicate a message about the data. Participants were asked to score each of these items on a scale of 1 to 5, with lower scores indicating negative responses such as 'very unclear' or 'very difficult' or 'strongly disagree' and higher scores indicating positive responses such as 'very clear' or 'very easy' or 'strongly agree'. Participants were also asked to rank each plot in relation to other plots within the same category.

Table 5.1: Draft framework for assessing the properties of graphical displays

Item	**Subjective criteria for appraisal**
1	Effect size: Does it clearly display an effect size for each event?
2	Treatment effect: Does it clearly display the direction of the treatment effect?
3	Uncertainty: Does it clearly display a robust measure of uncertainty such as CI or SEs?
4	Does it require supplementary data presentations? I.e. does it stand-alone or does it need additional data presented alongside it?
5	Can you understand the plot without a detailed explanation?
6	Do you think non-statistical colleagues i.e. clinicians can understand the plot without a detailed explanation?
7	Are there limitations around the number of events displayed?
8	Overall score (sum of items 1-7)
9	Is it suitable for the journal article, final study report, interim analysis report? (all that apply)
10	Is it useful for exploratory or explanatory communication? (all that apply)
	Objective criteria for appraisal
11	Does it allow presentation of absolute effects or relative effects or both?
12	Is there flexibility regarding the summary statistic displayed? e.g. OR, RR, IRR etc.
13	Is it easily produced across a variety of software?

Acronyms: CI - confidence interval; SEs - standard errors; OR - odds ratio; RR - risk ratio; IRR - incidence rate ratio

Table 5.2: Framework for assessing the properties of graphical displays

Item	Criteria for appraisal	Responses
1	Effect size - Does it clearly display an effect size for events?	1: no/very unclear, 2: unclear, 3: unsure, 4: clear, 5: yes/very clear
2	Treatment effect - Does it clearly display the direction of the treatment effect?	1: no/very unclear, 2: unclear, 3: unsure, 4: clear, 5: yes/very clear
3	Uncertainty – Does it clearly display a robust measure of uncertainty such as CI or SEs?	1: no/very unclear, 2: unclear, 3: unsure, 4: clear, 5: yes/very clear
4	Does it require supplementary data presentations? I.e. Does it stand-alone or does it need additional data presented alongside it?	1: yes/extremely likely, 2: likely, 3: neutral, 4: unlikely, 5: extremely unlikely/stand-alone
5	Can you understand the plot without a detailed explanation?	1: very difficult, 2: difficult, 3: neutral, 4: easy, 5: very easy
6	Do you think non-statistical colleagues i.e. clinicians can understand the plot without a detailed explanation?	1: very difficult, 2: difficult, 3: neutral, 4: easy, 5: very easy
7	How adaptable is the plot for multi-arms/adaptive trials?	1: very difficult, 2: difficult, 3: neutral, 4: easy, 5: very easy
8	Are there limitations around the number of events displayed?	1: yes/extremely limited, 2: very limited, 3: moderately limited, 4: slightly limited, 5: not at all/unlimited
	Total for 1-7*	
9	Is it suitable for inclusion in a:	
i	Journal article	1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree
ii	Final study report	1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree
iii	Interim analysis report	1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree
10	Is it best suited to†:	
i	Exploratory analysis	1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree
ii	Explanatory analysis	1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree
11	Ranking	1 - most preferred through to least preferred
12	Comments: Please indicate if you support any of the amendments proposed for this plot or any other comments on this plot that are not captured elsewhere	Please provide details of the amendment in case of multiple suggested amendments

Abbreviations: CI - confidence interval; SE – standard error

*Total score is a sum of scores assigned to questions 1 through to 7. Scores to question eight were not included as whilst we thought this was important for consideration, we did not wish to disadvantage plots that could only present a limited number of events, as this might be through design and in fact in some settings is likely to be an advantage. The group discussed this point before the decision was made.

† Exploratory analysis was defined as visualisations suited to data exploration to help identify potential signals for adverse drug reactions and explanatory analysis was defined as visualisation suited to communicate a message about the data.

Consensus meeting one

In the first meeting, a brief overview of current practice and its shortcomings (identified in work undertaken in earlier chapters) was provided, evidence that visualisations could offer a potential

solution to some of these issues was presented and the aims of the meeting and the project were outlined.

I presented summaries of each of the graphics for multiple binary outcomes, single time-to-event outcomes and single binary outcomes and each was discussed in turn amongst the group over the course of the first meeting. Participants were encouraged to raise any queries they had regarding each plot, to highlight what they liked or disliked about it, consider in which research contexts they thought it might useful and to raise any potential problems or opportunities for causing confusion or over-emphasis of effects. Participants could use the audio and the chat function to raise these comments during the virtual meetings. Time was given to participants to complete their appraisals (using the finalised framework in table 5.2) for each graphic and encouraged to include free text comments if they endorsed any of the recommendations or adaptations discussed or if they wanted to raise any concerns.

Consensus meetings two and three

Following the presentations and discussions in meeting one, appraisals for the first three categories of graphics were returned for analysis before further discussions in meeting two. Summaries of appraisal scores were shared with participants in advance of subsequent meetings. In meeting two, graphics for single time-to-event, multiple continuous, and single continuous outcomes were presented and discussed. In addition, appraisals from the first meeting were presented and were used to guide discussions about which plots to retain. After examining the results of the initial appraisals participants engaged in further discussions and were encouraged to champion low scoring plots if they felt strongly that they were underscored, but asked to consider where possible adaptations might be needed. Participants were advised that appraisal scores should be used to guide and focus discussions rather than being used as a formal guide for selection and that is was

possible for low scoring plots to be taken forward if deemed appropriate. As well as looking at the overall score for each plot, participants were encouraged to examine the individual items that drove the overall score so that they had a better understanding of what others considered the strengths and weaknesses of each of the plots.

Once discussions were concluded the software tool Mentimeter was used to obtain participants votes on whether they wished to take plots forward for further discussion regarding recommendations for use and refinements. Results of these votes were presented back to the group in real-time and unclear results were revisited and discussed, with further votes undertaken if necessary.

In meeting three, results from appraisals for single time-to-event, multiple continuous, and single continuous outcomes were presented and discussed. Participants had the opportunity to champion poorly performing plots and highlight where they thought adaptations might be needed. Finally, participants voted on each of the plots using Mentimeter to decide which should be taken forward. Once the final plots to endorse had been decided, discussions and free text comments from the appraisals were summarised and presented back to the group and participants were given the opportunity to raise any other important points they felt had been omitted. These were focused on comments about potential adaptations, where each plot should be recommended for use and what, if any, cautions or limitations should be included within the recommendations. Mentimeter was then used to garner endorsement for each of the adaptations, and to finalise the appearance of the endorsed plots and accompanying recommendations.

The majority of the meeting time was dedicated to decisions around the plots to include rather than finalising the wording of the recommendations as recommended by the CONSORT group executive.[239] Specific recommendations were drafted following the meetings and further feedback was sought via email. All meetings were recorded with the consent of participants and minutes transcribed upon completion.

Clinician interviews

One-to-one, semi-structured, in-depth interviews were undertaken with two of the three invited clinicians to gather feedback on the plots endorsed by the consensus group. This was also an opportunity to gather their insights on the utility and interpretation of each plot, which could then be used to structure the explanatory information provided in the recommendations. Work to date was outlined, results and recommendations from the consensus meetings were presented and feedback sought. Topics to guide discussions included:

a. Opinions on the finalised plots, including the merits of each and whether they were likely to use/endorse these plots in practice.
b. Whether they thought any of the plots were unclear and/or required further explanation that could be incorporated into the recommendations.
c. Whether they thought any modifications were required to any of the plots.

Questions asked were open-ended to allow for unconstrained responses and comments (appendix A5.8 lists the specific questions clinicians were asked to consider for each plot). Clinicians were also encouraged to raise any other comments they had that were not prompted from the topic areas and were given opportunity to provide written feedback following the meeting if they wished to. All meetings were recorded with the consent of clinicians and minutes transcribed upon completion.

5.4.6 Analysis

Consensus meetings

Appraisal scores were summarised with means and standard deviations (SD) and medians and inter-quartile ranges (IQRs). These summaries were presented back to the group in subsequent meetings using both tables and graphs. Free text comments regarding recommendations and adaptations were grouped into themes according to whether they related to amendments, limitations, cautions or advantages for each of the recommended plots before being presented back to participants. Results of the Mentimeter votes were summarised using frequencies and percentages. Default graphical summaries produced in Mentimeter were presented to the consensus attendees in the meeting once voting was completed. A cut-off of 60% was used to indicate endorsement. Scores in the 50-60% range were revisited for further discussions and votes retaken until a consensus could be reached.

Clinician interviews

No formal analysis of these interviews were undertaken but free text comments were summarised and themed according to whether they related to amendments, limitations, cautions or advantages for each of the recommended plots and were incorporated into the final recommendations of endorsed plots.

5.5 Consensus results

The results of the consensus meetings are described below, this includes a summary of the initial appraisal scores, group discussions and outcome of the Mentimeter votes to decide final recommendations and modifications. Full descriptions of the final plots chosen for recommendation are presented in section 5.6.

5.5.1 Participant characteristics

Twenty-three participants contributed to at least one of the sessions over the course of the three meetings. This included 20 statisticians from 15 UKCRC registered CTUs, a health economist based at an academic population health department, a statistician at a UK pharmaceutical company, and a graphic designer who sits on the multimedia team of leading academic medical journal (the BMJ) (a full list of names can be found in appendix A5.1).

Clinician interviews included a consultant in intensive care medicine with over twenty years of research experience developing and leading RCTs, and a consultant hepatologist new to developing and leading RCTs but with extensive experience running trials. One clinician declined to participate due to limited availability because of the COVID-19 pandemic.

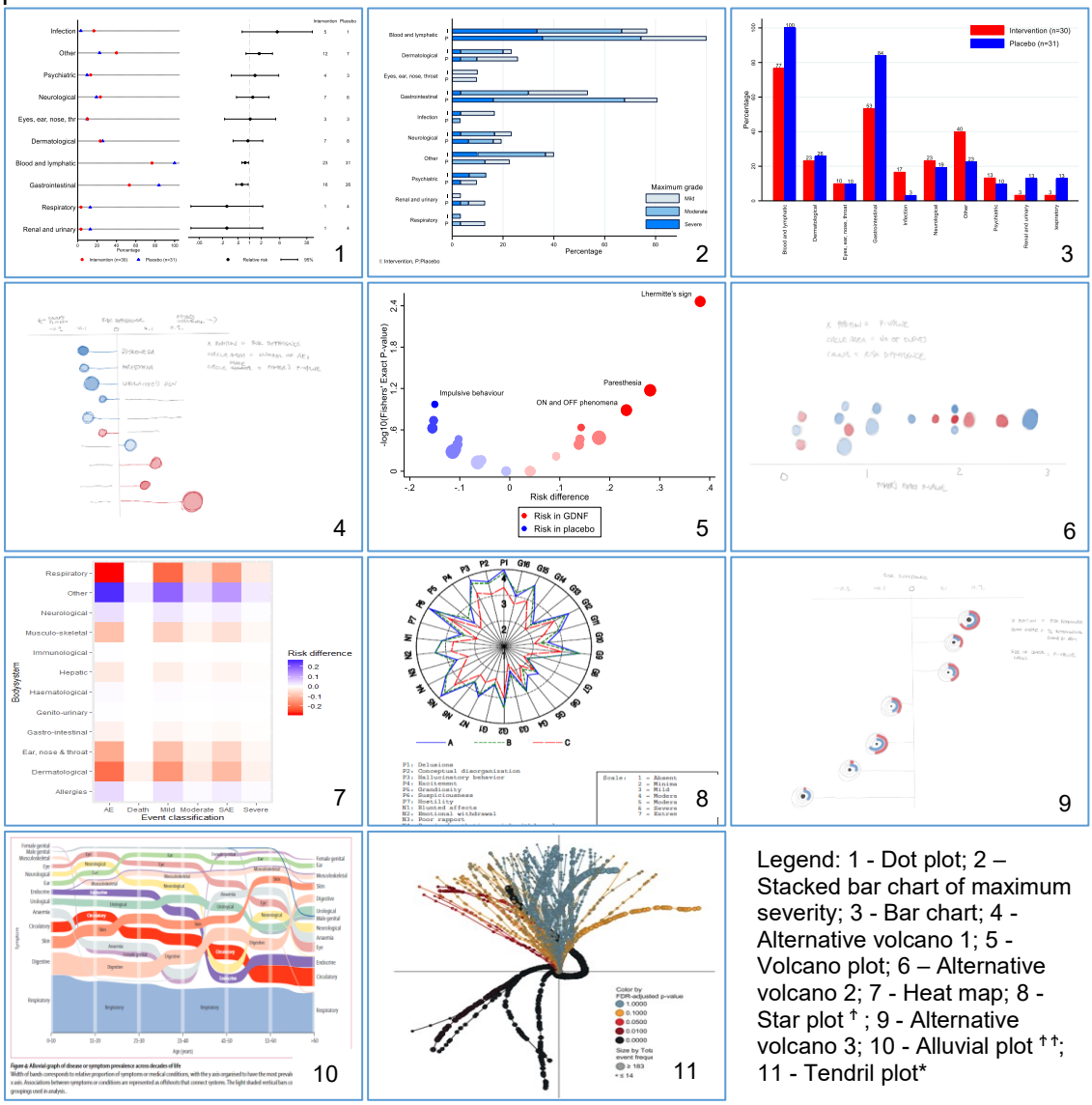
5.5.2 Multiple binary outcomes

Summaries of appraisals from consensus meeting

Plots considered for presentation of multiple binary outcomes are displayed in order of consensus attendees' ranked preference for recommended use in figure 5.1. In terms of overall scores (sum of items 1-7) the best performing plots were the dot plot (mean score 28.1 (SD 5.0)), the stacked bar chart (mean score 25.1 (SD 2.8)) and the bar chart (mean score 24.6 (SD 3.2)) (thumbnails 1, 2 and 3 in figure 5.1 reflecting rankings of first to third in terms of preference). The worst performers were the more complex alluvial plot (mean score 10.9 (SD 3.5)), the alternative volcano 3 (mean score 9.6 (SD 7.6)) and the tendril plot (mean score 9.3 (SD 2.4)) (thumbnails 10, 9 and 11 which reflect preference ranking positions in figure 5.1).

The alternative volcano 1 and the volcano plot (thumbnails 4 and 5 in figure 5.1) scored very similarly in terms of overall scores (mean 20.3 (SD 4.0) and 19.2 (SD 4.5), respectively), with alternative volcano 2 (thumbnail 6 in figure 5.1) scoring a little lower (mean 17.3 (SD 5.0)). Examining the individual items of the overall score both the volcano and alternative volcano 1 plots scored well in terms of clearly displaying an effect size and the direction of effect. However, they performed less well on other items, scoring a mean of two for their ability to display a measure of uncertainty and the need for supplementary data. They scored similarly for participants own understanding (mean grade approximately 3) but alternative volcano 1 performed better at being understandable to non-statisticians. Whilst alternative volcano 2 scored consistently worse compared to the volcano plot and alternative volcano 1, it did perform similarly to the volcano plot in its ability to be understood by a non-statistician, and scored comparably to alternative volcano 1 and the volcano plot due to its suitability to present data from multi-arm studies. The heat map (thumbnail 7 in figure 5.1) scored comparably to the volcano plot and alternative volcano 1 in terms of ease of participants own understanding and a little better than the volcano plot for ease of understanding to non-statisticians but scored poorly in terms of presenting clear effect sizes, direction of effects and uncertainty. Both the star and alluvial plot (thumbnails 8 and 10 in figure 5.1) performed poorly across items but the star plot outperformed the alluvial and scored comparably to other plots due to its ability to incorporate information on multi-arm studies. Overall scores and rankings are summarised in figure 5.2 and further details are provided in table A5.1 in appendix A5.9.

Figure 5.1: Thumbnails of considered plots for multiple binary outcomes in order of preference



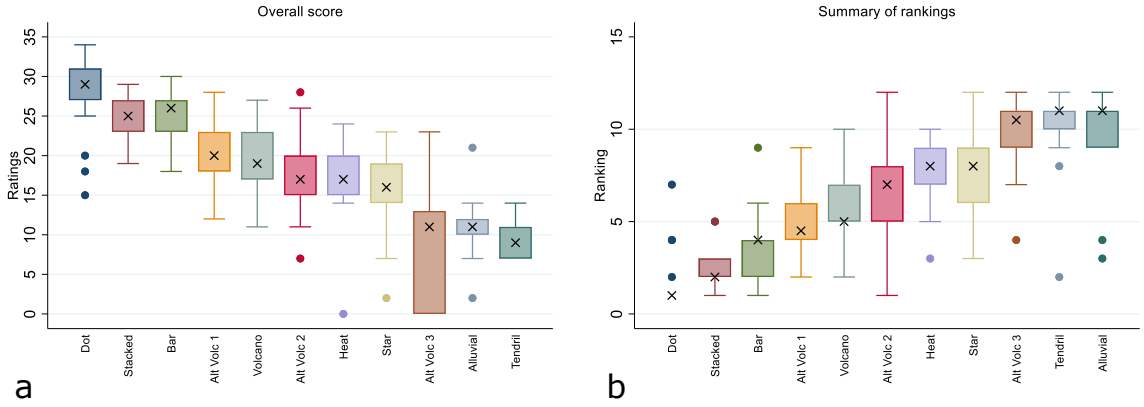
Legend: 1 - Dot plot; 2 – Stacked bar chart of maximum severity; 3 - Bar chart; 4 - Alternative volcano 1; 5 - Volcano plot; 6 – Alternative volcano 2; 7 - Heat map; 8 - Star plot † ; 9 - Alternative volcano 3; 10 - Alluvial plot † †; 11 - Tendril plot*

** Reprinted from: Karpefors, M. and J. Weatherall (2018). "The Tendril Plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials." Journal of the American Medical Informatics Association 25(8): 1069-1073 with permission of Oxford University Press.136*

† Reprinted from Squassante, L., et al. (2006). "Simple graphical methods of displaying multiple clinical results." Pharm Stat 5(1): 51-60 with permission from John Wiley & Son.238

† † Reprinted from Salvi S, Apte K, Madas S, et al. Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: a 1-day point prevalence study (the POSEIDON study). Lancet Glob Health. 2015;3(12):e776-e784. doi:10.1016/S2214-109X(15)00152-7 under the terms of the Creative Commons CC BY NC ND License.239

Figure 5.2: Summaries of overall scores and rankings for multiple binary outcome plots



a. Box plot of overall scores ordered by highest to lowest mean values (higher scores indicate better performance). **b. Box plot of rankings** ordered by best to worst mean rank (lower ranking indicates preferred plot).
Note: X indicates median values. Excludes summary for stacked bar chart of counts as only limited numbers scored this plot

Summary of decisions from the consensus meeting

Participants voted unanimously to recommend the dot plot (100% endorsement) (thumbnail 1 of figure 5.1), in addition there was strong endorsement for both the stacked bar chart and bar chart with 95% and 81% voting to recommend these plots, respectively (thumbnails 2 and 3 of figure 5.1). Further discussions highlighted the redundancy of presenting both a dot plot and bar chart; therefore, it was decided to omit the bar chart to present the frequency of events. One hundred percent of participants voted not to recommend use of the tendril plot or alternative volcano 3 (thumbnails 11 and 9 of figure 5.1), and the majority of participants did not wish to recommend the alluvial plot (94%), star plot (90%) or heat map (84%) (thumbnails 10, 8 and 7 of figure 5.1). Support for the volcano plot (thumbnail 5 of figure 5.1) was moderate with 65% wishing to recommend it and participants were split over whether to recommend alternative volcano 2 and alternative volcano 1, with 50% and 47% of participants voting to recommend them, respectively (thumbnails 6 and 4 of figure 5.1). Results of the Mentimeter votes are presented in table A5.7 appendix A5.11.

With regards to the proposed amendments (summarised in appendix A5.10), 67% of participants were happy with the dot plot as originally proposed but further discussions followed by a vote revealed that 60% wished to include both counts of events and number of participants in the data table and hence this information has been included in final version of the plot. Eighty percent of participants were happy to recommend the stacked bar chart as proposed which uses horizontal bars instead of vertical or pyramid style (which were also shown to participants, images A5.10a and A5.10c in appendix A5.2), displays the percentage of participants with the event using the bar height and labels the bars with frequencies. Results are summarised in table A5.8 appendix A5.11.

After discussions about potential amendments that could be made to alternative volcano 2 and the volcano plot that would address some of the issues raised about the original volcano plot, such as comments relating to the repetition of information and inefficient use of space (full comments summarised in appendix A5.10), participants were asked to consider if there was now a place in the recommendations for either plot. Only 35% of participants supported recommending either of these plots so neither were given further consideration (table A5.8 appendix A5.11).

5.5.3 Single binary outcomes

Summaries of appraisals from consensus meeting

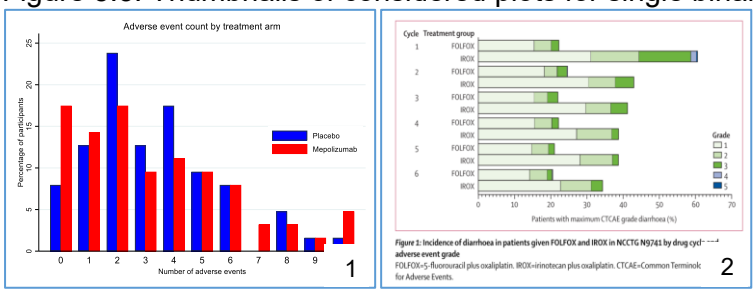
The bar chart of counts (thumbnail 1 in figure 5.3) was the only plot originally consider in this category with a mean overall score of 19.8 (SD 3.3). The score was largely driven by its ability to be understood by both consensus participants and non-statisticians and its ability to be adapted to multi-arm studies (mean scores 4.0 (SD 1.2), 3.9 (SD 0.9) and 3.9 (SD 0.9), respectively), scoring poorly in terms of presenting clear effect sizes, direction of effects and uncertainty. Summary statistics are presented in table A5.2 and figure A5.39 of appendix A5.9. Later discussions revealed

that the stacked bar chart of events over time (thumbnail 2 in figure 5.3) should be consider in this setting and not the single time-to-event category.

Summary of decisions from the consensus meeting

Discussions highlighted that some participants question the need for the bar chart of counts (thumbnail 1 in figure 5.3), with one participant commenting, "*is aggregation of data like this helpful?*", and others felt there could be difficulty in interpreting these plots (comments summarised in appendix A5.10). Discussions concluded that this plot might only be useful for summaries of serious events or pre-specified events. Taking into consideration these discussions participants were asked to vote whether this was a helpful plot in the harm setting. Sixty-seven percent of participants agreed that it was useful and preferred this information to be displayed in bars compared to the alternatives that were proposed in discussions such as using box-plots or using a variation of a dot plot (79%). Further details on the specific format according to context are provided in section 5.6.2. Summary statistics for each vote undertaken are presented in table A5.9 appendix A5.11.

Figure 5.3: Thumbnails of considered plots for single binary outcomes



Legend: 1 - Bar chart; 2 - Stacked bar chart*

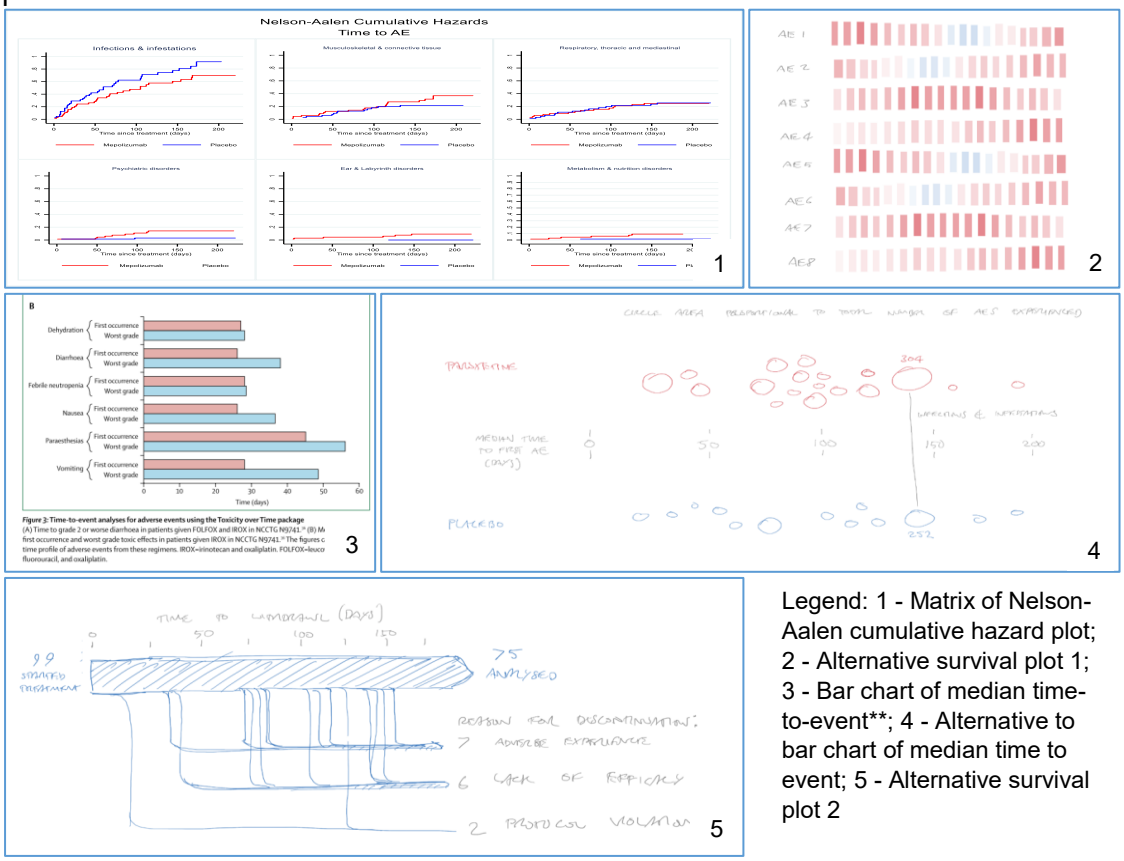
Reprinted from Thanarajasingam, G., et al. (2016). "Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254." Lancet Oncology **17(5): 663-670 with permission from Elsevier.*[241]

5.5.4 Multiple time-to-event outcomes

Summaries of appraisals from consensus meeting

Based on overall scores and the preference rankings the matrix of survival plots (thumbnail 1 in figure 5.4) outperformed all other plots considered in this setting (mean overall score 24.0 (SD 4.8) ranking in first place in terms of participant preference). This plot comprised of multiple cumulative hazard plots but participants discussed that this could be replaced with the optimal plot from the single time-to-event setting. Whilst the bar chart (thumbnail 3 in figure 5.4) performed well in terms of overall score (mean 19.0 (SD 5.4)), the alternative survival plot 1 (thumbnail 2 in figure 5.4) ranked ahead of it in second place. Summary statistics are presented in figure 5.5 and table A5.3 of appendix A5.9.

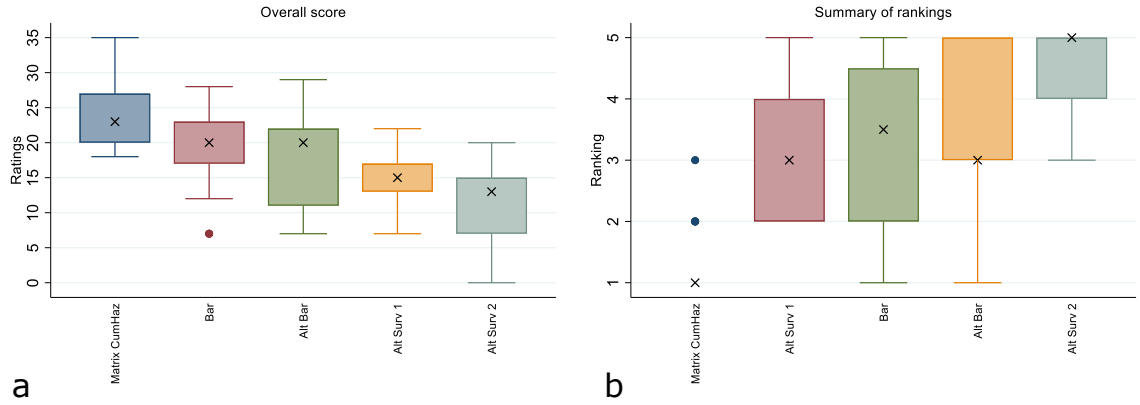
Figure 5.4: Thumbnails of considered plots for multiple time-to-event outcomes in order of preference



Legend: 1 - Matrix of Nelson-Aalen cumulative hazard plot; 2 - Alternative survival plot 1; 3 - Bar chart of median time-to-event**; 4 - Alternative to bar chart of median time to event; 5 - Alternative survival plot 2

***Reprinted from Thanarajasingam, G., et al. (2018). "Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies." The Lancet Haematology 5(11): e563-e598 with permission from Elsevier.[242]*

Figure 5.5: Summaries of overall scores and rankings for multiple time-to-event outcome plots



a. Box plot of overall scores ordered by highest to lowest mean values (higher scores indicate better performance). **b. Box plot of rankings** ordered by best to worst mean rank (lower ranking indicates preferred plot).
Note: X indicates median values.

Summary of decisions from the consensus meeting

There were limited graphical options in this setting and after appraisals and initial discussions 50% of participants voted not to endorse any of the available options. The strongest endorsement was for the matrix of survival plots (thumbnail 1 in figure 5.4), which was initially taken forward for consideration but after further discussions only 40% of participants voted to recommend it therefore this plot is not endorsed (table A5.10 in appendix A5.11).

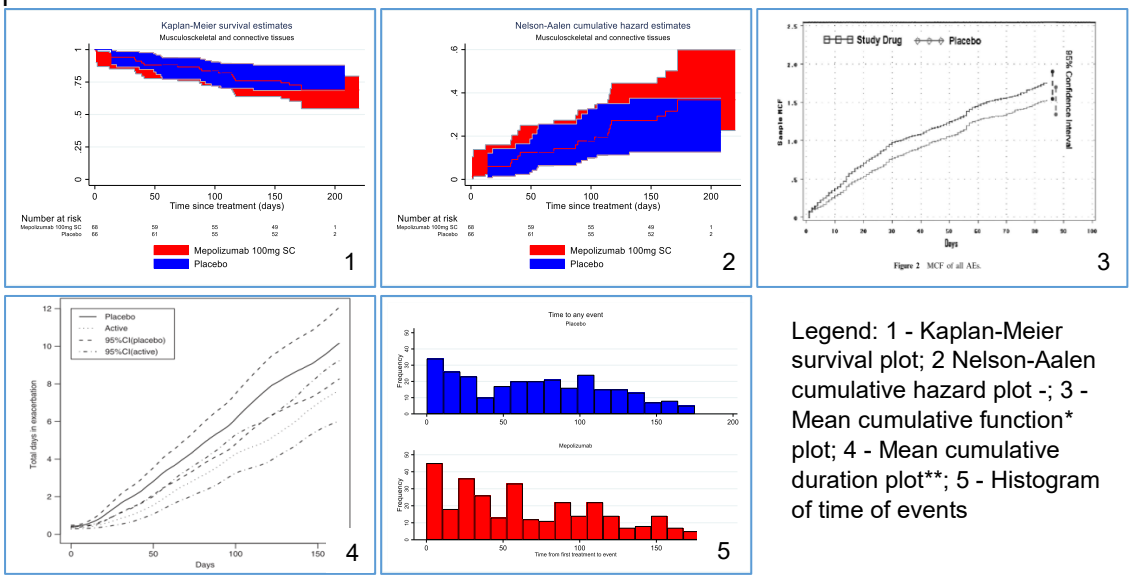
5.5.5 Single time-to-event outcomes

Summaries of appraisals from consensus meeting

The cumulative hazard and Kaplan-Meier plots (thumbnails 2 and 1 in figure 5.6 reflecting preference rankings positions of second and first) performed comparably with mean overall scores of 26.7 (SD 3.6) and 26.1 (SD 3.8). The Kaplan-Meier plot outperformed the cumulative hazard plot in its ability to display a clear effect size and being understandable to non-statisticians. The mean cumulative function plot (thumbnail 3 in figure 5.6) had a mean overall score of 23.2 (SD 5.6) and ranked third in terms of preference. It did not perform as well as the Kaplan-Meier and cumulative

hazards plots in terms of overall scores as it was deemed to more likely need supplementary data presentations and was judged not to be as easy to understand by both consensus participants and non-statisticians. At the lower end of the scale for overall score and rankings were the mean cumulative duration plot (thumbnail 4 in figure 5.6) (mean overall score of 21.0 (SD 5.1)) and the histogram of time of events (thumbnail 5 in figure 5.6) (mean overall score of 19.1 (SD 5.2)). The stacked bar chart of events over time (thumbnail 2 in figure 5.3) was originally appraised in this category but subsequent discussions revealed it should be grouped with the single binary plots. Summary statistics are presented in figure 5.7 and table A5.4 of appendix A5.9.

Figure 5.6: Thumbnails of considered plots for single time-to-event outcomes in order of preference

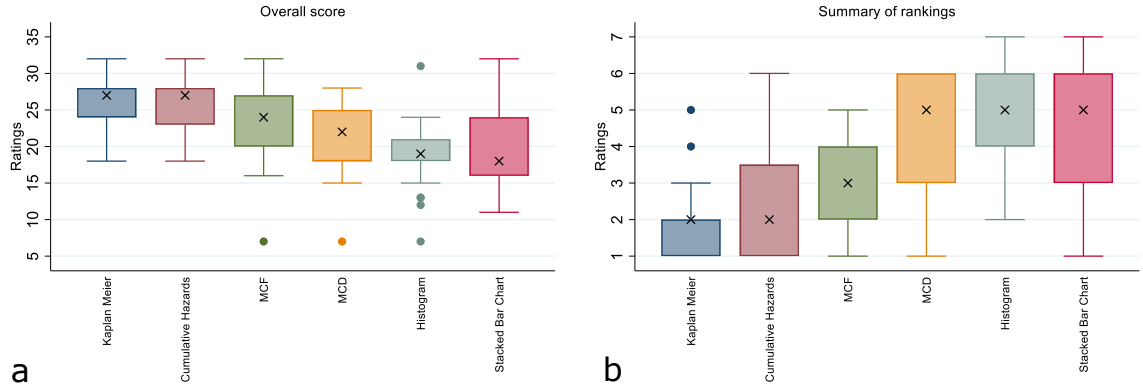


Legend: 1 - Kaplan-Meier survival plot; 2 Nelson-Aalen cumulative hazard plot -; 3 - Mean cumulative function* plot; 4 - Mean cumulative duration plot**; 5 - Histogram of time of events

Reprinted from Siddiqui, O. (2009). "Statistical methods to analyze adverse events data of randomized clinical trials." Journal of Biopharmaceutical Statistics 19(5): 889-899 with permission from Taylor & Francis.[44]

***Reprinted from Wang, J. and G. Quartey (2012). "Nonparametric estimation for cumulative duration of adverse events." Biometrical Journal 54(1): 61-74 with permission from John Wiley & Sons.[166]*

Figure 5.7: Summaries of overall scores and rankings for single time-to-event outcome plots



a. Box plot of overall scores ordered by highest to lowest mean values (higher scores indicate better performance). **b. Box plot of rankings** ordered by best to worst mean rank (lower ranking indicates preferred plot). Note: X indicates median values.

Summary of decisions from the consensus meeting

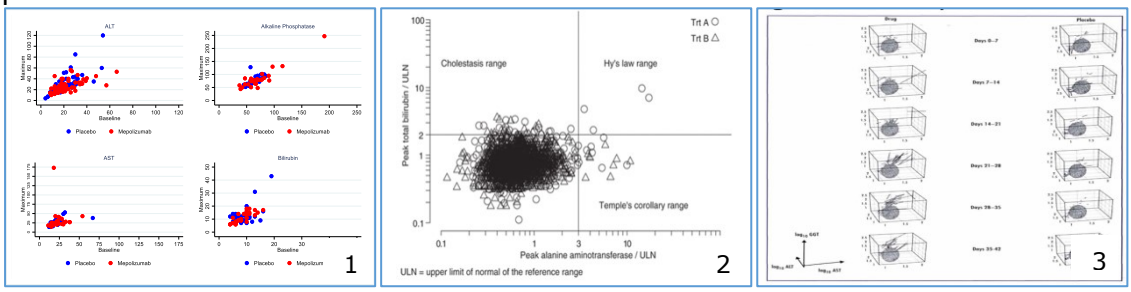
The majority of participants favoured the Kaplan-Meier plot (thumbnail 1 of figure 5.6) over the cumulative hazard plot (thumbnail 2 of figure 5.6) in this setting (83% versus 17%), with extended risk tables containing information on the numbers at risk, censored and experiencing an event per group (71%) as opposed to no tables (6%) or tables only with the numbers at risk (24%). Discussions highlighted that a limitation of the Kaplan-Meier plot was its lack of between treatment group comparison and that there was a desire to incorporate or present an alternative plot with a between group comparison. Participants suggested consideration of the survival ratio plot proposed by Newell et al. which are referred to in this thesis as event free ratio plots (comments summarised in appendix A5.10).[243] Following presentation of this plot and discussions sixty-seven percent of participants were in favour of recommending event-free ratio plots as an addition to the Kaplan-Meier plot to provide a between arm comparison. Eighty-eight percent of participants wished to recommend using plots of the mean cumulative function (thumbnail 3 in figure 5.6) to display information on repeated events, including a table of numbers at risk over time (94%) (table A5.11 appendix A5.11).

5.5.6 Multiple continuous outcomes

Summaries of appraisals from consensus meeting

The only plausible candidate in this category was the scatterplot matrix (thumbnail 1 in figure 5.8) after both the e-dish plot (thumbnail 2 in figure 5.8) and the vector plot (thumbnail 3 in figure 5.8) were deemed unsuitable and excluded from consideration. Appraisals were completed for the e-dish plot before it was decided to exclude it. The scatterplot matrix achieved a mean overall score of 22.8 (SD 4.3). This was driven by its perceived ability to be understood by both consensus participants and non-statisticians (mean scores of 4.5 (SD 0.8) and 4.3 (SD 0.6)), with all other items scoring a mean of three or less. Summary statistics are presented in table A5.5 and figures A5.40 and A5.41 of the appendix A5.9.

Figure 5.8: Thumbnails of considered plots for multiple continuous outcomes in order of preference



Legend: A - Matrix of scatterplots; B - e-Dish plot*; C - Vector plot**

**Reprinted from Xia HA, Crowe BJ, Schriver RC, Oster M, Hall DB. Planning and core analyses for periodic aggregate safety data reviews. Clin Trials. 2011;8(2):175-182. doi:10.1177/1740774510395635 with permission from Sage Publishing.*[182]

***Reprinted from : Trost, D. C. and J. W. Freston (2008). "Vector Analysis to Detect Hepatotoxicity Signals in Drug Development." Therapeutic Innovation & Regulatory Science 42(1): 27-34 under the terms of the Creative Commons CC BY License.*[142]

Summary of decisions from the consensus meeting

After initial discussions excluded the e-dish and vector plot there was only one option for consideration in this setting, the matrix of scatterplots, (thumbnail 1 in figure 5.8) and 94% of participants thought it should be recommended (table A5.12 appendix A5.11) despite the limitations

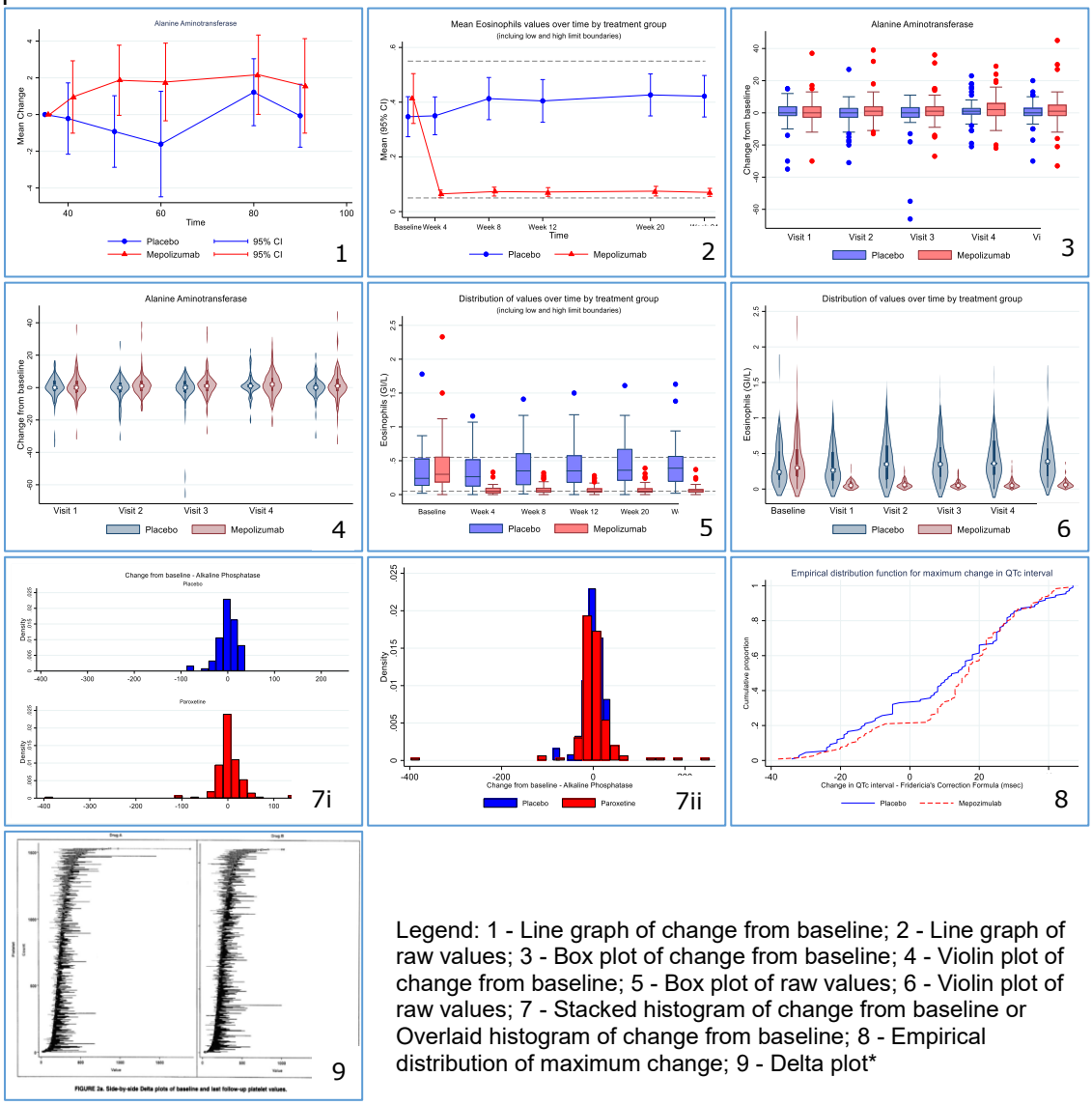
highlighted in discussions such as overlapping points obscuring data (comments from discussion summarised in appendix A5.10).

5.5.7 Single continuous outcomes

Summaries of appraisals from consensus meeting

The line chart of change (mean overall score 26.3 (SD 7.3)) and raw scores (mean overall score 25.3 (SD 7.4)) (thumbnails 1 and 2 in figure 5.9) performed the best in terms of overall scores. This was followed by the box plot of change (mean overall score 23.7 (SD 6.3)) and raw scores (mean overall score 23.1 (SD 6.4)) (thumbnails 3 and 5 in figure 5.9), the histogram of maximum change (mean overall score 21.9 (SD 5.8)) (thumbnail 7 in figure 5.9) and the violin plot of change (mean overall score 20.3 (SD 5.4)) and raw scores (mean overall score 20.1 (SD 5.4)) (thumbnails in 4 and 6 in figures 5.9). However, rankings revealed a preference for the line chart, box plot and violin plot over the histogram. The worst performers were the empirical distribution plot (mean overall score 17.7 (SD 5.9)) (thumbnail 8 in figure 5.9) and the delta plot (mean overall score 12.0 (SD 3.8)) (thumbnail 9 in figure 5.9), ranking eighth and ninth in terms of preference. Summary statistics are presented in figure 5.10 and table A5.6 in appendix A5.9.

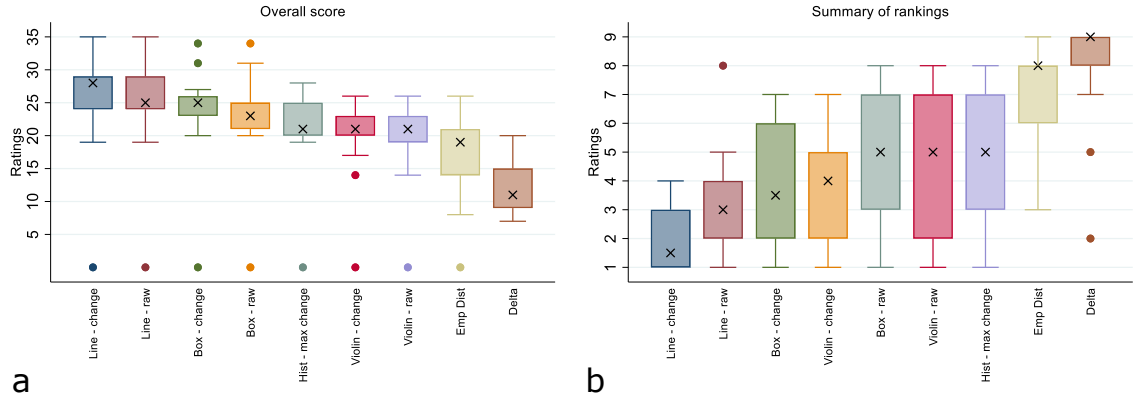
Figure 5.9: Thumbnails of considered plots for single continuous outcomes in order of preference



Legend: 1 - Line graph of change from baseline; 2 - Line graph of raw values; 3 - Box plot of change from baseline; 4 - Violin plot of change from baseline; 5 - Box plot of raw values; 6 - Violin plot of raw values; 7 - Stacked histogram of change from baseline or Overlaid histogram of change from baseline; 8 - Empirical distribution of maximum change; 9 - Delta plot*

**Reprinted from Chuang-Stein, C., et al. (2001). "Recent Advancements in the Analysis and Presentation of Safety Data." Drug Information Journal 35(2): 377-397 under the terms of the Creative Commons CC BY License.*[138]

Figure 5.10: Summaries of overall scores and rankings for single continuous outcomes



a b

a. Box plot of overall scores ordered by highest to lowest mean values (higher scores indicate better performance). **b. Box plot of rankings ordered** by best to worst mean rank (lower ranking indicates preferred plot).
Note: X indicates median values.

Summary of decisions from the consensus meeting

The majority of participants wished to recommend a version of the line chart (94%) (thumbnail 1 or 2 of figure 5.9) with discussions revealing that participants thought that it was best for individuals to decide whether they should display a summary of change scores, a summary of raw values or estimates from more advanced modelling approaches (comments from discussion summarised in appendix A5.10). Sixty-seven percent of participants wished to recommend the violin plot (thumbnail 4 or 6 of figure 5.9) as an alternative to the line graph, with only 53% favouring the box plot (thumbnail 3 or 5 of figure 5.9). Discussions indicated that whilst participants were interested in recommending a plot that visually displayed an informal comparison of the distribution of continuous outcomes they did not think the histogram was a good visual representation of such data. The kernel density plot was proposed as an alternative and sixty-one percent voted to recommend it over the histogram (thumbnail 7i or 7ii of figure 5.9) as a means to compare distributions (61% versus 0%). Thirty-nine percent of participants did not wish to recommend either the histogram or the kernel density plot (table A5.13 appendix A5.11).

5.6 Final recommendations

Recommended visualisations that consensus participants agreed on are displayed in figures 5.11-5.21 according to outcome type (binary, time-to-event or continuous) and number of events displayed (single or multiple). These are presented alongside the accompanying plot description, the specifics of the recommendations and cautions and limitations the consensus group and clinicians wished to raise in the published recommendations. Summaries of the plots by recommended scenario to be used in are presented in tables 5.3 and 5.4, and a decision tree to help trialists decide which plot to use is provided in figure 5.22. Plots that were considered but not recommended are included for information in appendices with descriptions.

An example of each recommended plot was produced using data from four pharmacological RCTs obtained via the ClinicalStudyDataRequest.com initiative from GlaxoSmithKline (GSK). The first was a two-arm (randomised in a 1:1 allocation) study that evaluated the efficacy of mepolizumab compared to placebo in patients with severe eosinophilic asthma (n=135). The second study investigated mepolizumab in patients with severe uncontrolled refractory asthma comparing two doses of mepolizumab to placebo (randomised in a 1:1:1 allocation) (n=576). The third study was a two-arm (randomised in a 2:1 allocation) trial examining the efficacy, safety and tolerability of paroxetine compared to placebo in adolescents with unipolar major depression (n=286). The fourth was a two-arm (randomised in a 1:1 allocation) trial examining the efficacy and tolerability of paroxetine compared to placebo in paediatric major depression (n=206).[244-247] In addition, a synthetic dataset was created based on a clinical trial of a novel active treatment for eczema compared to placebo (randomised in a 1:1 allocation) in adolescents unresponsive to standard care (n=61). The synthetic dataset is available for download in the Stata `aedot` and `aevolcano` command packages described in chapter three section 3.4.4.[204, 205]

These datasets were selected because they were from completed trials of pharmacological products with varying sample sizes and would potentially contain both short and long-term effects. In addition, each of these trials produced an abundance of data on emerging harms and are representative of 'typical' pharmacological trials.

5.6.1 Multiple binary outcome

Dot plot

Plot description: The dot plot provides a way to evaluate simultaneously both the relative and absolute risk for multiple events (figure 5.11). The dot plot displays the percentage of participants experiencing an event (each event labelled on the y-axis) in each treatment group on the left-hand side of the plot, and a relative measure, such as the relative risk for binary harm outcomes, with corresponding 95% confidence interval in the central panel on the log10 scale. Events are ordered from the bottom to the top by increasing relative risk. The 95% confidence interval can be used to assess precision of the relative estimate (which is particularly important when the event rate is low), and the strength of evidence against a null hypothesis of no difference. This can be done through examining the position of the lower or upper confidence limit in comparison to the value of no difference and the plot incorporates a line to show the value of no difference (for relative risks this is 1). However, interpretation of confidence intervals fixated only on whether they cross the summary statistic value of no difference is discouraged. This plot has been adapted to include a data table on the far right of the plot to contain information on number of participants with at least one event and the number of events by treatment group.

Recommendation: The group unanimously endorsed the dot plot for presenting data on multiple binary outcomes. Suggesting that the plot provides such a comprehensive presentation of the data that it could be presented instead of the traditional frequency table of events.

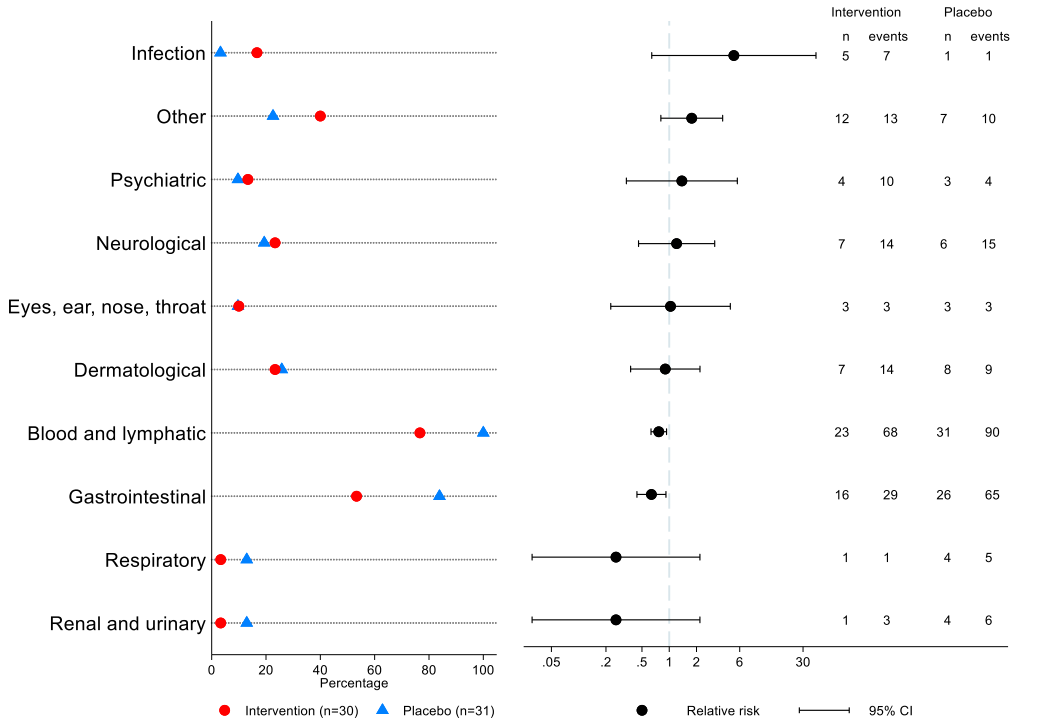
Potential amendments: The relative risk, odds ratio or incident rate ratios can be plotted as the relative measure in the central panel of this plot. Some may also prefer to present the data table in the central panel so that it appears alongside the absolute summary. It is also possible to create this plot in black and white without loss of meaning. Extension to multi-arm studies is possible in the situation of a common comparator group but consideration should be given to clarity of presentation.

Limitations/cautions: Confidence intervals around the relative differences are useful to raise potential signals of harm but the aim is not to encourage hypothesis-testing which will inflate both type I errors (concluding a false finding for a chance imbalance between arms), and type II errors (incorrectly concluding there is no imbalance in harm between arms).[248] Clinician feedback indicated that users should give careful consideration to the x-axis range for the absolute summary and scale for the relative measure to ensure clarity without exaggerating effects. Whilst the dot plot gives a comprehensive overview, some potentially important pieces of information are not included such as information on severity. In scenarios where it is important to display information on severity, researchers can plot the stacked bar chart (see details below).

Software: The dot plot can be produced in Stata using the `aedot` or `aedots` command dependent on the structure of the data and in SAS and R using code available from the CTSpedia Wiki page (https://www.ctspedia.org/do/view/CTSpedia/ClinAEGraph000) but modifications to the R and SAS code are needed to incorporate the data table.[249, 250]

Implementation and interpretation: In the dot plot presented in figure 5.11 the point estimates are evenly distributed on either side of the vertical line of 'no difference' (relative risk = 1) with great uncertainty in many of the estimates. The relative risk furthest from one communicates increased risk of infection in the intervention group but the absolute risk on the left and frequencies in the data table on the far right indicate small numbers of participants experiencing this event. There is also evidence of a reduced risk of respiratory events, and renal and urinary events in the intervention group; the absolute risks on the left and the raw numbers in the data table indicate only small numbers experiencing these events. Also of note are the estimates for blood and lymphatic disorders and gastrointestinal events where the relative risks indicate a small but reduced risk in the intervention group, with confidence intervals that do not cross one and absolute differences and raw numbers indicating large numbers experiencing these events. This suggests a potential beneficial effect of the intervention on these outcomes.

Figure 5.11: Dot plot of events - data taken from the two-arm example dataset with 1:1 allocation ratio



Dot Plot for emerging harm outcomes between two treatment groups for the simulated dataset. The left panel of the figure displays the percentage of participants experiencing an event (labelled on the y-axis) in the intervention group with a red circle and placebo group with a blue triangle. The central panel of the figure displays the relative risk and corresponding 95% confidence interval on the log10 scale and a line to show the value of no difference (for relative risks this is 1). The right panel displays the 'number of participants experiencing the event at least once' (*n*) and 'the number of events' (*events*) (accounting for recurrent events within participants) by treatment group. The dot plot provides a comprehensive visual representation of the entire harm profile.

Stacked bar chart

Plot description: The horizontal stacked bar chart presents the percentage of participants with an event by maximum severity grade i.e. if a participant had the same event twice, once classified as mild and once as moderate this participant would be counted once as experiencing a moderate event (figure 5.12). The bars are labelled with the corresponding number of participants. Bars are split by colour gradient to indicate different severity groups and the total bar height indicates the proportion of participants experiencing that event at least once. The most severe category is displayed closest to the y-axis to allow ease of comparison for the most harmful or burdensome events.

Recommendations: The stacked bar chart is easy to understand and can be used when it is important to present information on severity of multiple events. It can be used to informally compare severe or severe plus moderate events or overall events between groups. It is recommended that treatment groups are displayed directly adjacent to each other for each event and that horizontal labelling is used for ease of reading.

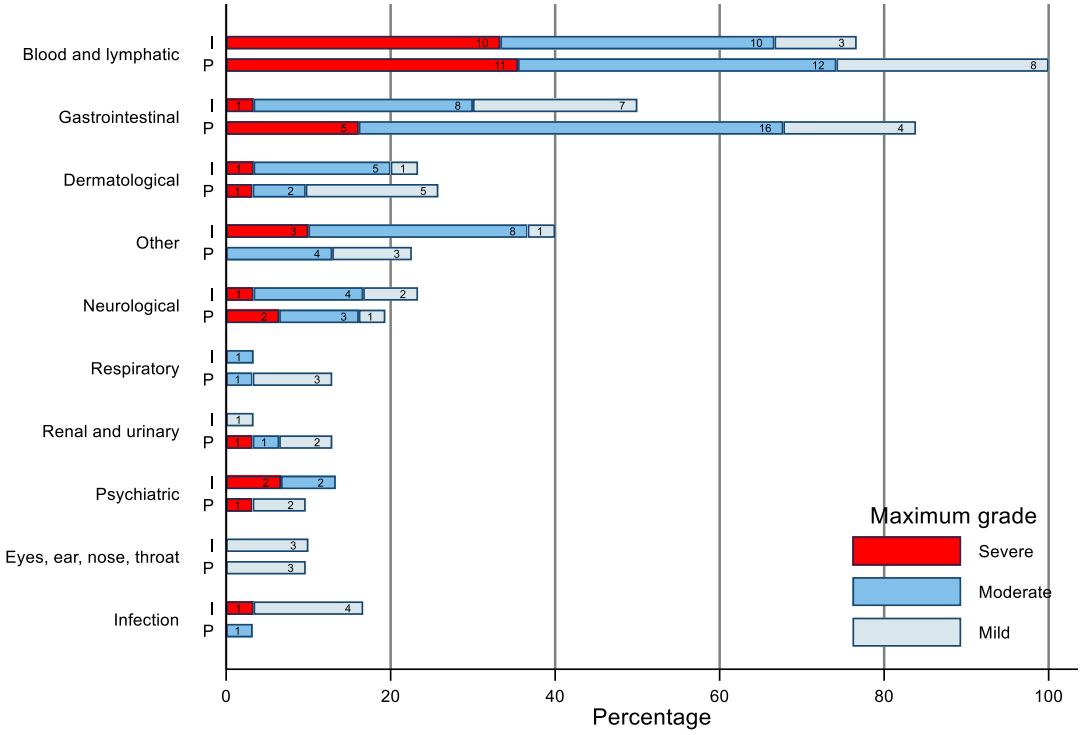
Potential amendments: This plot can be easily adapted to multi-arm studies and graduation in colour from black to white is possible to avoid use of colour. In addition, it could be adapted to the single event setting by replacing events on the y-axis with some representation of time e.g. visits or treatment cycle, an example of which can be found in Thanarajasingam et al.[241]

Limitations/cautions: Direct comparisons within stacked bars are not possible beyond the segment closest to the y-axis. It promotes presenting information on 'participants with at least one event' rather than 'number of events' and it is important that information on repeated events is still presented. In addition, there is no explicit display of effect sizes for differences between groups; trialists may wish to consider alternative visualisations such as the dot plot to include these.

Software: Stacked bar charts are easily implemented as standard plots across the variety of statistical packages. For example, using `graph hbar` in Stata or the R command `barplot` or the `ggplot2` package with `geom_bar` or SAS `proc gchart`.

Implementation and interpretation: In the stacked bar chart presented in figure 5.12 it is clear that the most common events are blood and lymphatic events and gastrointestinal disorders. It also shows that while more blood and lymphatic events occurred in the placebo group, there were similar numbers between groups in the most severe categories (severe plus moderate) and the difference in numbers between groups was because of the difference in numbers experiencing mild events. For gastrointestinal disorders, the stacked bar chart revealed that there were fewer events in the intervention group across each of the severity grades in comparison to the placebo group. The plot also revealed that events classified as 'other' were dominated by severe and moderate events in the intervention group compared to the control group, which could warrant closer inspection of what these events were.

Figure 5.12: Horizontal stacked bar chart of events by maximum severity – data taken from the two-arm example dataset with 1:1 allocation ratio



I: Intervention (n=30), P:Placebo (n=31)
Bars are labelled with the corresponding number of participants

Horizontal stacked bar chart for emerging harm outcomes by maximum severity and treatment group for the simulated dataset. Total bar height represents the proportion of participants with that event at least once and each bar is split into segments to indicate numbers by severity grading. Bar segments are labelled with the corresponding number of participants. The stacked bar chart used in this way is helpful when it is important to present information on the severity of multiple events.

5.6.2 Single binary outcomes

Bar chart – for counts

Plot description: A bar chart to present information on the number of events (event counts) experienced per participant (figures 5.13 and 5.14). Each bar represents the percentage of participants with 0, 1, 2 etc. events for each treatment group.

Recommendations: The bar chart is recommended to present information on the number of events experienced. This is a simple plot that can be useful to illustrate differences in counts of events between treatment groups and is potentially useful to highlight differences in the burden of harm amongst participants. It can be used to present information on an overall summary of events such as the total number of serious adverse events a participant experiences or for a limited number of specific events of interest. It can also be used in an exploratory setting to show the distribution of repeated events and has been used in the literature to help justify not considering recurrent events in subsequent analysis.[251, 252] Vertical bars with treatment groups alongside each other are the recommended format (figure 5.13) when comparing two treatment groups. When there are more than two treatment groups, separate plots stacked above each other for each group (figure 5.14) is the recommended alternative.

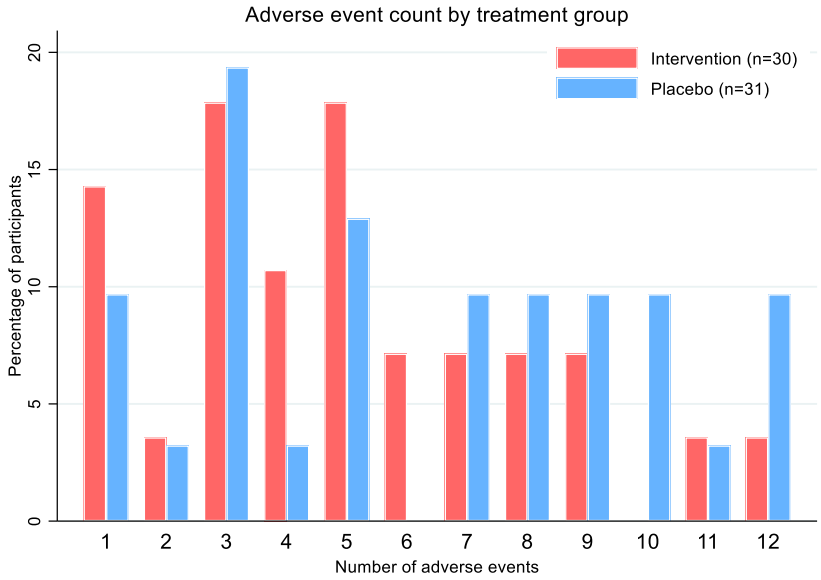
Potential amendments: This plot can be easily adapted to multi-arm studies and can be produced in black and white if necessary. Bars could also be labelled with number of participants to ensure accurate communication of total number of events if not listed elsewhere.

Limitations/cautions: Whilst this plot is helpful for summarising and comparing the overall burden of different treatments, it does not make a distinction between the types of events contributing to it. Therefore, it is still vitally important that trialists explore and report the individual event data, giving careful consideration as to whether such a plot for overall events could be misleading. In addition, whilst it could potentially reveal patterns in the data, clinician feedback indicated that they felt subtle differences would be less obvious and careful consideration of when to use this plot and the accompanying message it supports would be needed.

Software: Bar charts are easily implemented as standard plots across the variety of statistical packages. For example, using `graph bar` in Stata or using the R command `barplot` or `ggplot2` with `geom_bar` or SAS `proc gchart`.

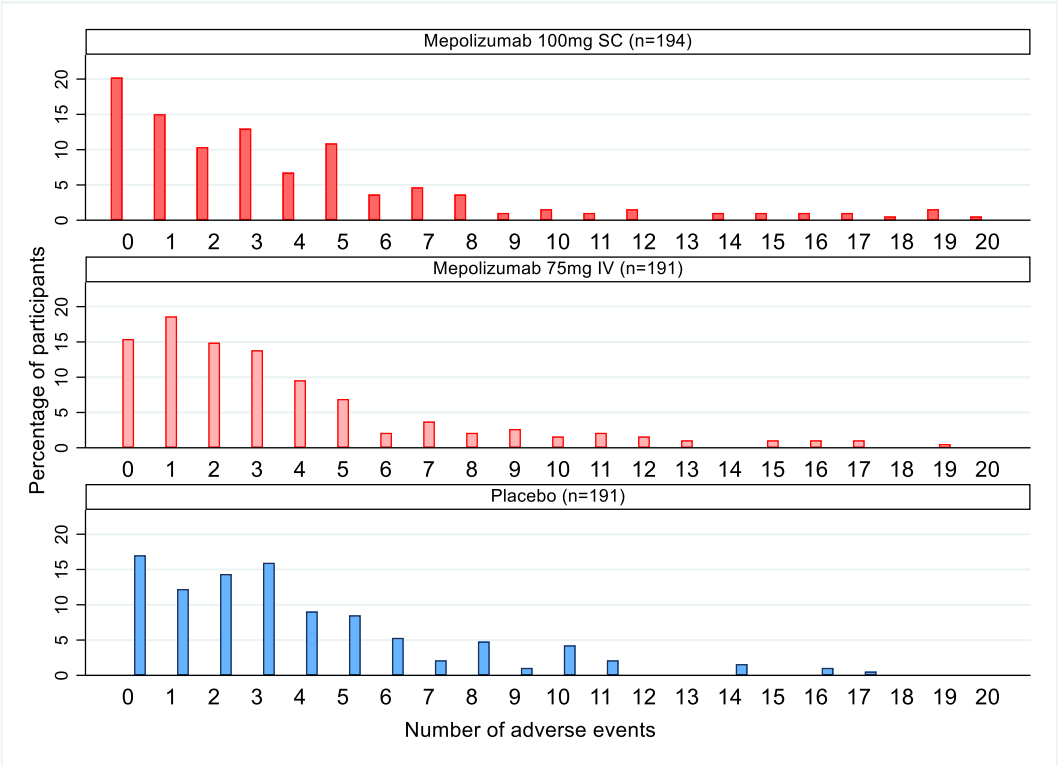
Implementation and interpretation: Figure 5.13 displays the distribution for the multiple events experienced by participants, with placebo participants experiencing higher numbers of multiple events more often. In figure 5.14, the distributions indicate that participants in either of the intervention groups experience multiple events more often compared to the placebo group.

Figure 5.13: Bar chart of event counts – data taken from the two-arm example dataset with 1:1 allocation ratio



Bar chart of counts of harm outcomes by treatment group for the simulated dataset. Each bar represents the proportion of participants with 0, 1, 2 etc. events for each treatment group. This plot groups all events together. Alternatively, it can be used to summarise this information for specific events of interest. Using the bar chart to present this information can help highlight between group differences in the burden of harm experienced.

Figure 5.14: Bar chart of event counts – data taken from the three-arm Mepolizumab dataset with 1:1 allocation ratio



Bar chart of counts of harm outcomes by treatment group (when >2 treatment groups). Each bar represents the proportion of participants with 0, 1, 2 etc. events for each treatment group. Separate stacked plots like this are recommended for trials with more than two treatment groups. Using the bar chart to present this information can help highlight between group differences in the burden of harm experienced.

5.6.3 Multiple time-to-event outcomes

Recommendation: The group did not endorse any plot in this setting.

5.6.4 Single time-to-event outcomes

Kaplan Meier plot

Plot description: The Kaplan-Meier plot shows the cumulative proportion of participants remaining event free over time by treatment group (figure 5.15). The 95% confidence interval bands indicate the precision of the within group estimates of being event free. The extended risk table below the plot shows the number of participants that remain 'at risk', the cumulative number that have been censored and the cumulative number that have experienced an event at discrete time points.

Recommendations: The Kaplan-Meier plot with within group confidence bands and extended risk tables is recommended to detect either a large between treatment group difference or a potential disproportionality over time, as frequently ADRs are time-dependent, in the occurrence of a specific event of interest between treatment groups.

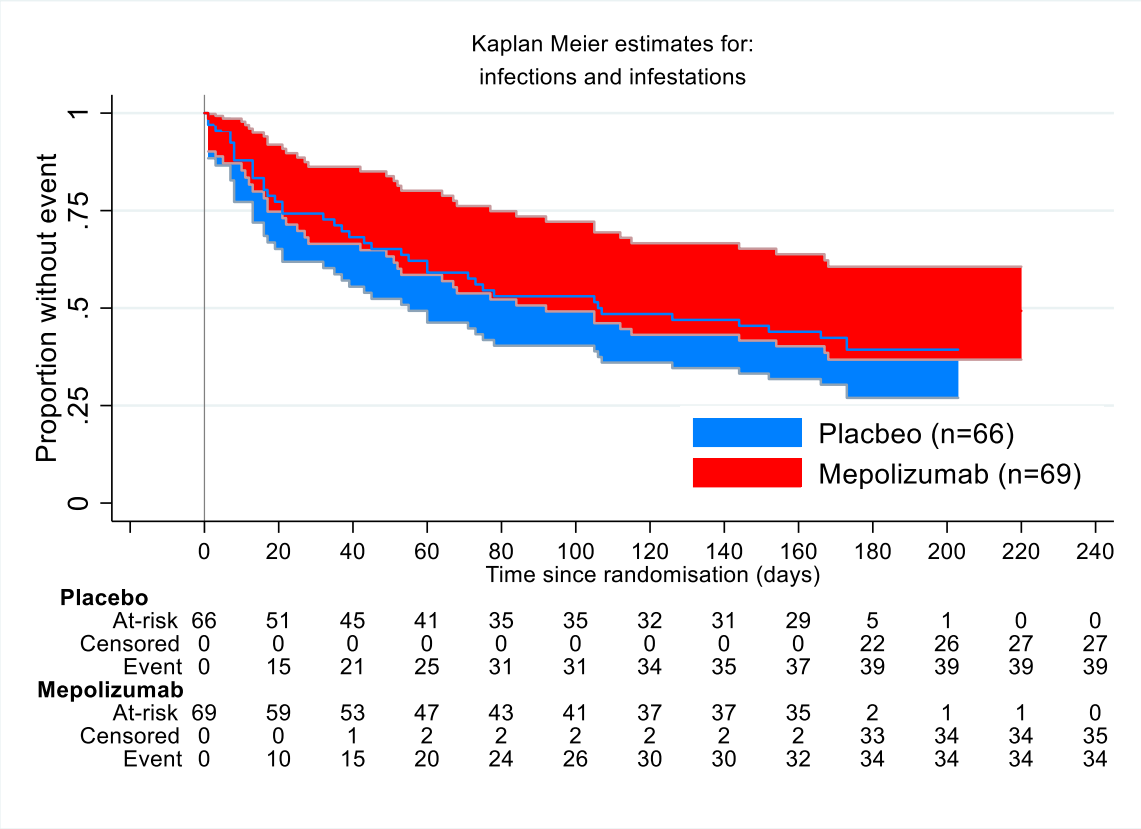
Potential amendments: For rare events, reversing the y-axis to display the cumulative proportion with the event adds clarity and ease of interpretation. It is also possible to create this plot in black and white and use different line styles to differentiate between groups. Extension to multi-arm studies is potentially feasible but consideration should be given to the clarity when displaying multiple overlying confidence bands, therefore trialists should consider only plotting the survival estimates with extended risk tables or present separate plots for comparison of each intervention group to a common comparator.

Limitations/cautions: Kaplan-Meier plots are typically limited to displaying information on one type of event at a time and only depict time-to-first event, failing to consider recurrent events. To present information on recurrent events a plot of the mean cumulative function (MCF) (see section: *Mean cumulative function*) is recommended. Some generic limitations of time-to-event plots in the harm setting are provided in section: *Limitations applicable to both Kaplan-Meier plots and plots of the MCF.*

Software: Kaplan-Meier plots are easily implemented as standard plots across the variety of statistical packages. To incorporate the extended risk tables there is an R package `KMunicate` and script for implementation in Stata and SAS are available here https://github.com/tpmorris/kmunicate.[253]

Implementation and interpretation: The extended risk tables indicate that by the end of follow-up there was little difference in the number of participants experiencing an infection or infestations disorder. However, the event curves indicate that 50% of the placebo group experienced this event within approximately 100 days of randomisation, but it took until 160 days post randomisation for 50% of participants in the mepolizumab group to experience the event.

Figure 5.15: Kaplan-Meier plot for an event of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio



Kaplan-Meier plot with an extended at risk table for specific harm outcome of interest by treatment group for the two arm Mepolizumab study. Plots the survival estimates by treatment group. Each line indicates the cumulative proportion of participants remaining event free over time by treatment group and includes 95% confidence intervals within groups. The extended risk table includes information on the number of participants that remain 'at risk', the cumulative number that have been censored and the cumulative number that have experienced an event at discrete time points. In the harm setting, Kaplan-Meier plots can be used to present information for specific events of interest as a useful way to detect a large between treatment group difference or a potential disproportionality between treatment groups, which is useful when trying to identify signals for ADRs.

Mean cumulative function plot

Plot description: The MCF plot is a non-parametric estimate of the mean cumulative number of events per participant (displayed on the y-axis) as a function of time (x-axis) by treatment group (figure 5.16). The 95% confidence interval bands show the precision of the within group estimates. The risk table includes information on the number of participants that remain at risk at discrete time points.

Recommendations: This plot is recommended to display information on recurrent events, providing a visual summary of the expected time until '*x number of an event*' will be experienced per participant by group. This can be provided as a summary to demonstrate the burden of 'any event' as demonstrated in figure 5.16, or the recurrence of events of special interest. As highlighted in clinical feedback these plots are potentially very useful when investigating long-term therapies for chronic conditions and can provide insight on periods the therapy might be considered 'safe' or 'well-tolerated'.

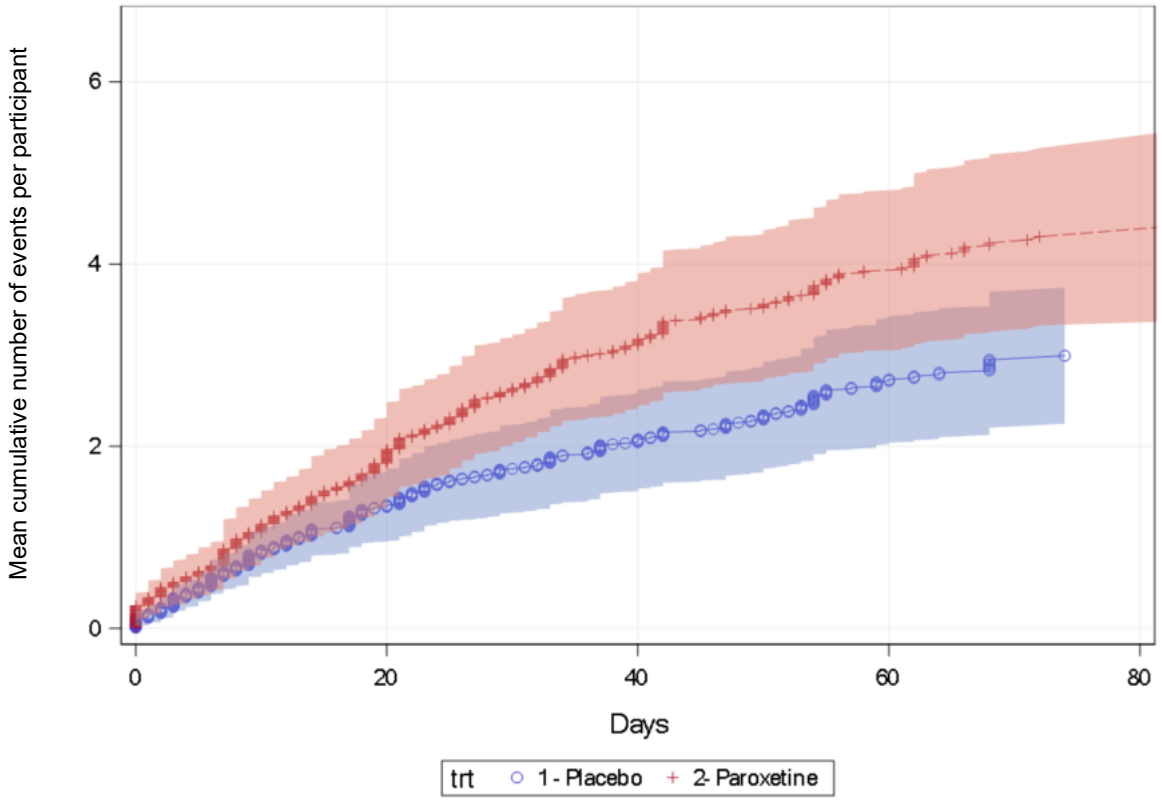
Potential amendments: As per the Kaplan-Meier plot, it is possible to create this plot in black and white without loss of meaning. Extension to multi-arm studies is potentially feasible but displaying multiple overlying confidence bands could make it unclear, therefore as per the recommendation for the Kaplan-Meier plot trialists should consider only plotting the MCF and risk table (without confidence bands) or present separate plots for comparison of each intervention group to a common comparator.

Limitations/cautions: MCF plots are limited to one type of event at a time. More generic limitations of time-to-event plots in the harm setting are provided in section: *Limitations applicable to both Kaplan-Meier and MCF.*

Software: The MCF with confidence interval bands using the SAS `proc reliability` procedure and `mcfplot` command.

Implementation and interpretation: The plot of the MCF shows the participant burden of recurrent events. Over the first week of follow-up, the mean number of events is similar across treatment groups, but by day 20, a divergence becomes apparent. In the paroxetine group, a mean of two events per participant was observed by day 20, but in the placebo group, this occurred nearer to day 40. By day 60, participants in the paroxetine group experienced a mean of four events, but the placebo group participants experienced a mean of less than three events by the same point.

Figure 5.16: Mean cumulative function plot for all events – data taken from the two-arm Paroxetine dataset with 1:1 allocation ratio



Placebo					
At risk	102	98	91	62	12
Paroxetine					
At risk	101	95	82	60	13

Mean cumulative function plot for harm outcomes by treatment group for the Paroxetine study with 1:1 treatment allocation. Plots the mean number of events per participant over time by treatment group and includes 95% confidence intervals within groups. The risk table includes information on the number of participants that remain 'at risk' at discrete time points throughout the study. In the harm setting, MCF plots can be used to demonstrate a comparison of the burden of experiencing 'any event' or the recurrence of events of special interest.

Limitations applicable to both Kaplan-Meier plots and plots of the MCF

The measure of uncertainty (confidence interval bands) in these plots is within treatment groups and not between treatment groups, which is the inference of interest in comparative clinical trials. The event-free ratio plot (originally referred to as the survival ratio plot) should be used to incorporate an estimate of the between group difference with a confidence interval (see section below). In addition, when using time-to-event methods for harm data, trialists must be cautious of the limitations around competing risks and consider these when performing the underlying time-to-event analysis. More information on alternative strategies to account for competing risks can be found in Proctor et al. and include using appropriate estimates such as the Aalen Johnson estimator or Fine and Grey method to plot the cumulative incident function.[197]

Event-free (survival) ratio plot

Plot description: This plot displays the ratio of event-free probabilities between treatment groups over time with a 95% confidence band of feasible values, allowing a direct comparison between treatment groups (figure 5.17). Departures from unity indicate potential differences in survival probabilities between treatments and includes a horizontal bar at the bottom of the plot that changes colour to indicate when the confidence band excludes unity.[243]

Recommendation: This plot is recommended for use alongside the Kaplan-Meier plot to incorporate a direct estimate of the between group difference for a specific event of interest. As it provides a between group comparison it can be used to detect departures from unity and help identify the time that such divergences occur, which can help detect potential signals for ADRs.

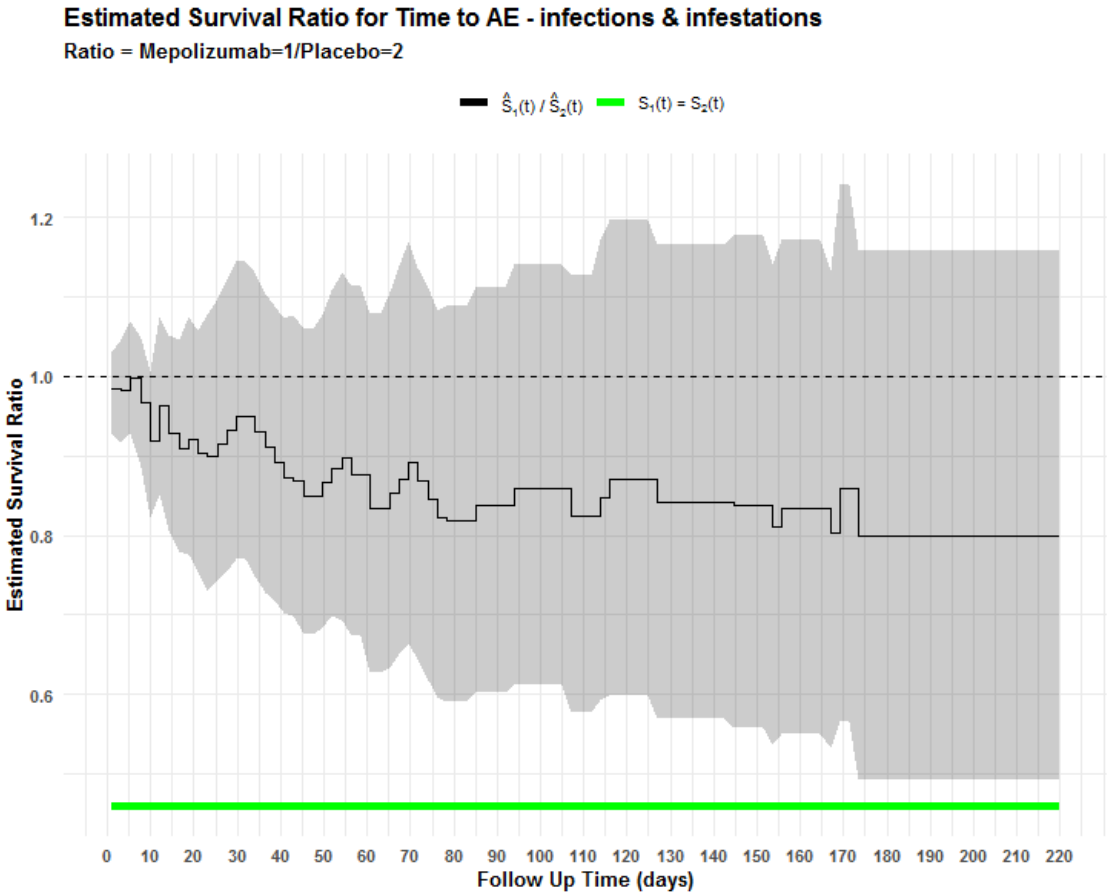
Potential amendments: The example displays the ratio of event-free probabilities estimated from the Kaplan-Meier method; alternatively, it could be used to display the difference in survival probabilities.

Limitations/cautions: As with Kaplan-Meier plots, the event-free ratio plot is limited to one type of event and only allows for time-to-first event, therefore it is not suitable for events that recur. As with other time-to-event plots it is important to consider competing risks when performing the underlying time-to-event analysis, further details of which are discussed in the section: *Limitations applicable to both Kaplan-Meier plots and plots of the MCF*. It also only provides a relative comparison. Therefore, it should be presented alongside the Kaplan-Meier plot to give an absolute comparison and this was deemed vital by clinicians. Confidence intervals around the relative differences are useful to detect potential signals of harm but again the aim is not to encourage hypothesis-testing.[248] Despite event-free ratio plots first being proposed in 2006 there is little evidence of application and as such any implementation of this plot will need to be accompanied with a detailed explanation, at least until the trials community become more familiar with the plot and how to interpret it.[243] This was confirmed in discussions with clinicians who initially struggled interpreting this plot but indicated strong endorsement once clarified.

Software: The event-free ratio plot can be implemented in R using the `survRatio` package with the `drsurv` function taking time, censoring indicator and treatment indicator as inputs. This returns Kaplan-Meier time-to-event estimates and corresponding confidence limits to create an object of the event-free ratio, event-free difference and pointwise (bootstrap) confidence bands and then the `ggsurv` function to create the plot of the event-free ratio and pointwise confidence bands.

Implementation and interpretation: Interpretation of the event-free ratio plot depicts a point estimate indicating a greater risk of infection and infestation disorders in the placebo group compared to the intervention group with a value between 0.9 and 1 until day 30 dropping to between 0.8 and 0.9 thereafter. Compared to the Kaplan-Meier plot, we can now see the confidence band for the between group comparison (rather than the within group confidence intervals in the Kaplan-Meier plot). The confidence band includes the point of unity (event-free ratio = 1) across all time periods and therefore would not provide sufficient evidence to raise a signal for this event as a potential ADR to undergo further investigation.

Figure 5.17: Event-free ratio plot for an event of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio



Event-free (survival) ratio plot for specific harms of interest for the two group Mepolizumab study. Plots the ratio of event-free estimates with the 95% pointwise confidence bands. Departures from unity are indicated using the horizontal band at the bottom of the plot, which is green when the confidence band includes one and red when it excludes one. In the harm setting, Kaplan-Meier plots can be used to present information for specific events of interest as a useful way to detect a large between treatment group difference or a potential disproportionality between treatment groups, which is useful when trying to identify signals for ADRs.

5.6.5 Multiple continuous outcomes

Matrix of scatter plots

Plot description: Multiple scatterplots of continuous outcomes. Each plot displays the relationship between values at two different time points e.g. baseline values along the x-axis and the maximum on treatment value along the y-axis (figure 5.18).

Recommendation: This plot is recommended in an exploratory setting to identify any outliers or patterns of interest and it is suggested that outlying values are labelled with a participant identifier to assess if one or more participants have abnormal measurements across outcomes. This could be useful to monitor participants in ongoing studies but may also help raise signals for potential ADRs in the final analyses.

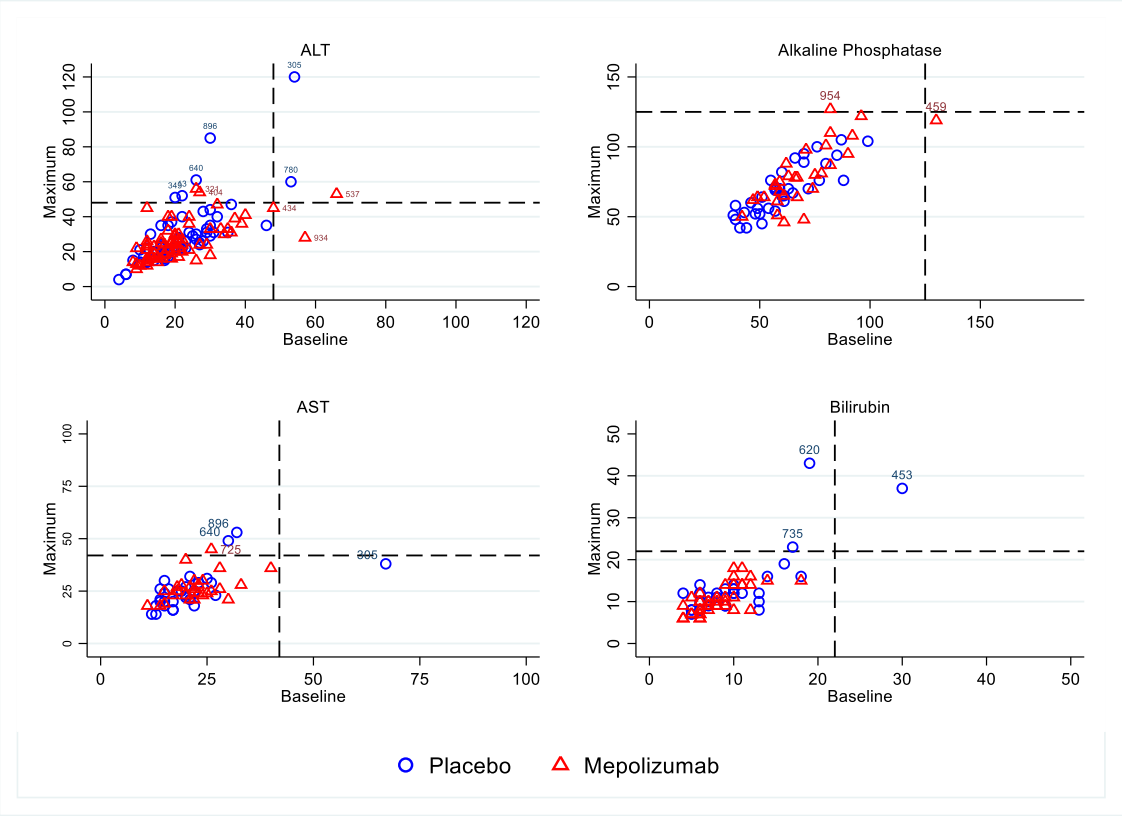
Possible adaptations: This plot could be used to explore two continuous measures post baseline. Variations in symbol style and colours should be used to help separate overlapping measurements between groups. Reference lines could be included to indicate both upper and lower limits of normal for each outcome.

Cautions/limitations: This plot presents several visual problems such as the use of solid colours results in occlusion making it impossible to distinguish individual points but transparency options could help with this. Trialists should use this plot whilst remaining mindful of its limitations.

Software: Scatterplots are easily implemented as standard plots across the variety of statistical packages. For example, in Stata, using `twoway scatter` to produce the individual plots and the `graph combine` or `grc1leg` command to produce the matrix of plots.

Implementation and interpretation: Figure 5.18 shows little change in the maximum on treatment values relative to baseline values for participants in the mepolizumab group. However, there are several placebo participants with ALT and bilirubin values of concern, with maximum on treatment values exceeding upper limits of normal.

Figure 5.18: Scatterplot matrix for continuous harm outcomes – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio



Scatterplot matrix for multiple continuous harm outcomes by treatment group. Plots each participant's baseline value against their maximum on treatment value. Outlying observations are labelled with participant identification numbers. This plot can be used in an exploratory setting to identify any outlying observations and to help identify any patterns within participants.

5.6.6 Single continuous outcomes

Line graph

Plot description: In this plot, the markers display mean values and the vertical lines indicate the standard deviation of raw values at each discrete time point, connected with a line over time for each treatment group (figure 5.19). Horizontal reference lines are included to indicate the upper and lower limits of normal values and a table of numbers at risk at each discrete time point is included.

Recommendations: This plot can be used to describe continuous harm outcomes such as laboratory or clinical outcomes of interest over time, using an appropriate summary statistic including an indication of variability. This plot can be helpful to identify shifts in distributions between treatment groups and highlight any potential trends.

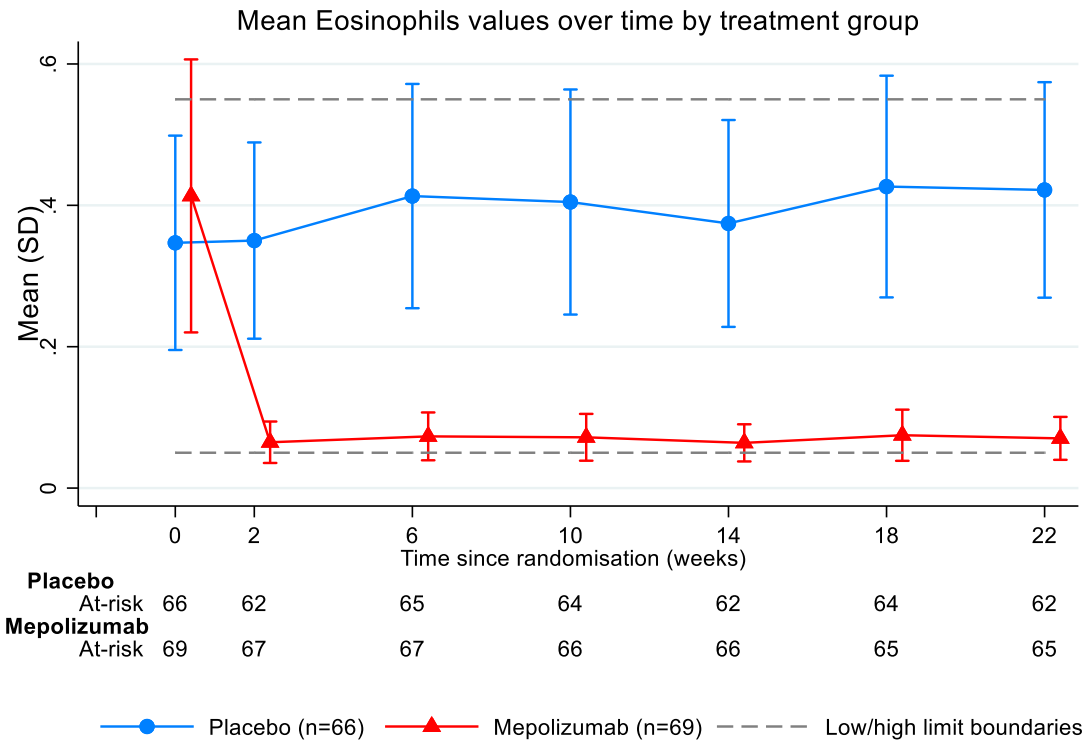
Potential adaptations: The summary statistic displayed in this plot should be chosen to reflect each individual dataset and the purpose of the plot e.g. when interest is in presenting descriptions of the distributions either means and SDs or medians and IQRs can be plotted, and if interest is in drawing inferences of between group comparisons then estimates from mixed effects models for repeated measures with 95% confidence intervals can be presented. This plot can easily incorporate multiple groups and can be modified not to use colour.

Cautions/limitations: Changes in the tails of the distributions are of interest when monitoring blood markers for harm and it may be difficult to see such changes using this plot. It is also unsuitable for skewed distributions so is better suited to present clinical outcomes rather than blood markers. Alternative plots for such data are presented below. Appropriate colour choices and line styles should be considered, particularly when adapting line graphs to multi-arm trials.

Software: Line graphs are easily implemented as standard plots across the variety of statistical packages. For example using `twoway connected` in Stata or using the R command `plot` and `lines` or the `ggplot2` package with `geom_line` and `geom_errorbar` or SAS `proc gplot`.

Implementation and interpretation: In figure 5.19, there is an immediate drop in the mean eosinophil count after randomisation in participants receiving mepolizumab and this is maintained across follow-up. The mean values for the placebo group fluctuate around the baseline value and the error bars exceed the upper limit of normal during follow-up.

Figure 5.19: Line graph of a summary statistic over time for a continuous harm outcome of interest – data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio



Line graph with risk table for specific continuous outcome of interest by treatment group over time. The markers display an appropriate summary statistic (in this example means) and the vertical lines indicate a measure of variability (in this example the standard deviation) of raw values at each discrete point connected with a line for each treatment group. This plot can be used to describe continuous outcomes over time for continuous harm outcomes of interest and can help identify shifts in distributions between treatment groups.

Violin plot

Plot description: The hollow circle marker on the violin plot indicates the median value, the narrow rectangular boxes around this marker indicate the inter-quartile range and the lines extend from the box to the minimum and maximum points. This is overlaid with vertical kernel density plots, which summarise the distribution of the raw values (figure 5.20).

Recommendation: This is an alternative plot to the line graph to describe continuous data and can be used even if the outcome of interest is not normally distributed. It depicts outlying values, and these can be labelled to highlight participants that are persistently showing values of concern.

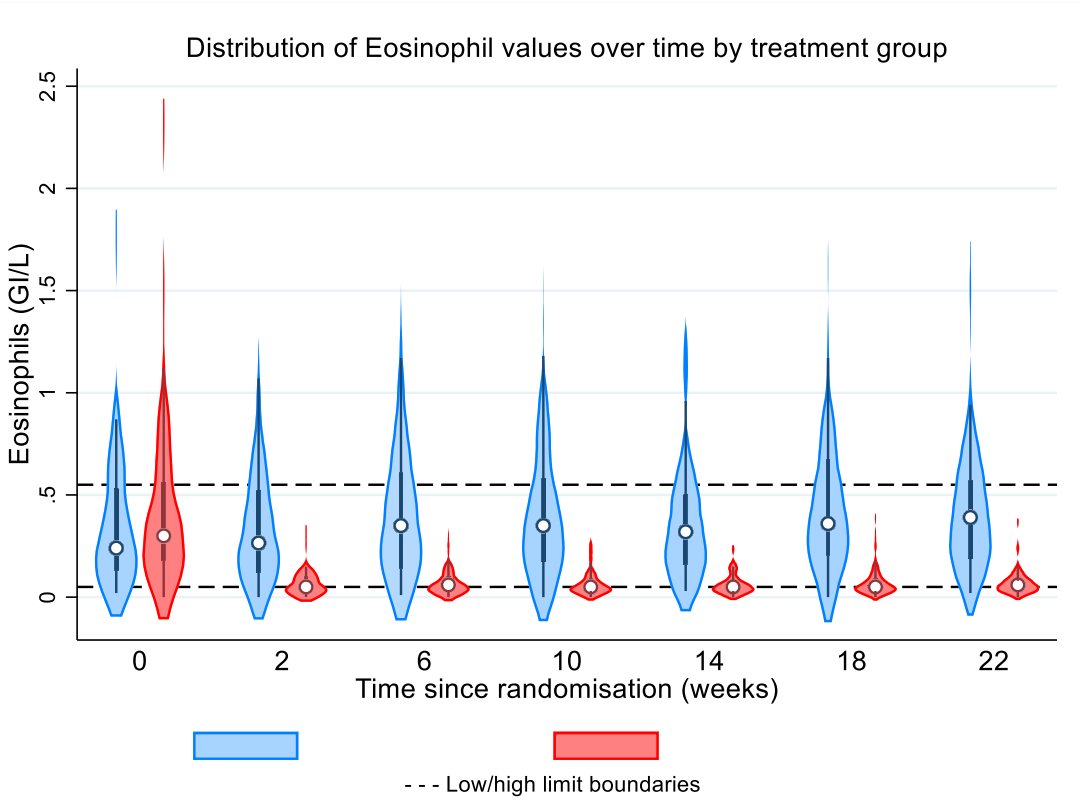
Possible adaptations: In the current format, there is duplication of information in the mirrored kernel density plot. Presenting only one kernel density would improve clarity and produce a more space efficient plot.

Cautions/limitations: The violin plot only allows for informal between group comparisons of distributions and does not allow for presentation of formal between group inferences such as the estimates from mixed effects models, which can be presented in a line graph. In addition, adaptations to multi-arm trials is not as space efficient as for the line graph. It is also possible for the kernel density estimates to extend to values outside the plausible range.

Software: The violin plot can be implemented in Stata using `vioplot` or using the `ggplot2` package in R with `geom_violin` or SAS `proc sgpanel`.

Implementation and interpretation: In this example, the violin plot shows that at randomisation, the distributions were similar across treatment groups, but by the first post-randomisation visit the distribution of the mepolizumab values was much narrower than the placebo group. The distribution of the placebo group values remained largely unchanged over time and indicated that a proportion of the participants remained in the upper tail exceeding the upper boundary of normal throughout follow-up.

Figure 5.20: Violin plot summarising the distribution of a continuous harm outcome of interest over time - data taken from the two-arm Mepolizumab dataset with 1:1 allocation ratio



Violin plot for specific continuous event of interest by treatment group over time. The markers indicate the median, the narrow rectangular boxes indicate the inter-quartile range and the lines extend to minimum and maximum points, overlaid with kernel density plots. The violin plot is a useful alternative to the line graph when presenting a continuous outcome that is far from a normal distribution and there is interest in exploring the distribution. It can also help identify outliers and to identify participants that are persistently showing values of concern.

Kernel density plot

Plot description: The kernel density plot displays the distribution of a continuous outcome. This can be at a single time point or a derived change score (e.g. the difference between the baseline value and maximum on treatment value) (figure 5.21). Vertical reference lines can be included to indicate the upper and lower limits of normal values for the outcome.

Recommendations: The kernel-density plot is recommended to explore an outcome of interest at a specific time-point or a change score e.g. the change from baseline to a specific point in time or maximum change over the entire trial. When plotting raw scores, vertical reference lines can be included to indicate the upper and lower limits of normal as per figure 5.21. The kernel-density plot should be used to informally compare whole distributions between treatment groups and can highlight important differences in distributions.

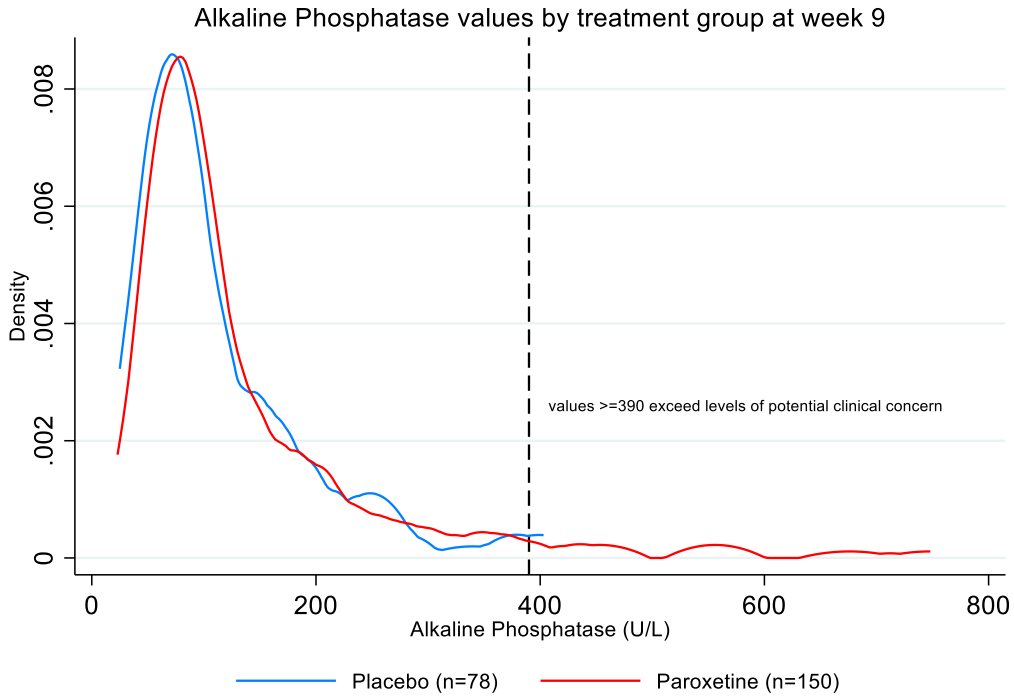
Potential adaptations: This plot can easily incorporate multiple groups and can be modified not to use colour.

Cautions/limitations: The kernel density plot only allows for informal between group comparisons of distributions and it loses the information on repeated measures, only displaying information for one time point.

Software: The kernel density plot can be implemented in Stata using `twoway kdensity` or using the `ggplot2` package in R with `geom_density` or SAS `densityplot`.

Implementation and interpretation: Figure 5.21 highlights a long right tail for the paroxetine group indicating that some participants have week 9 alkaline phosphatase levels exceeding the upper limit of normal. This plot highlights the increased alkaline phosphatase levels in participants taking paroxetine as an important event for closer monitoring in future trials.

Figure 5.21: Kernel density plot for a continuous harm outcome of interest - data taken from the two-arm Paroxetine study with 2:1 allocation ratio



Kernel density plot for a specific continuous outcome of interest by treatment group, at a single time point, with a reference line to indicate values above which are of clinical concern. It can be helpful to identify shifts in distributions between treatment groups.

Table 5.3: Visualisations for summarising the harm profile

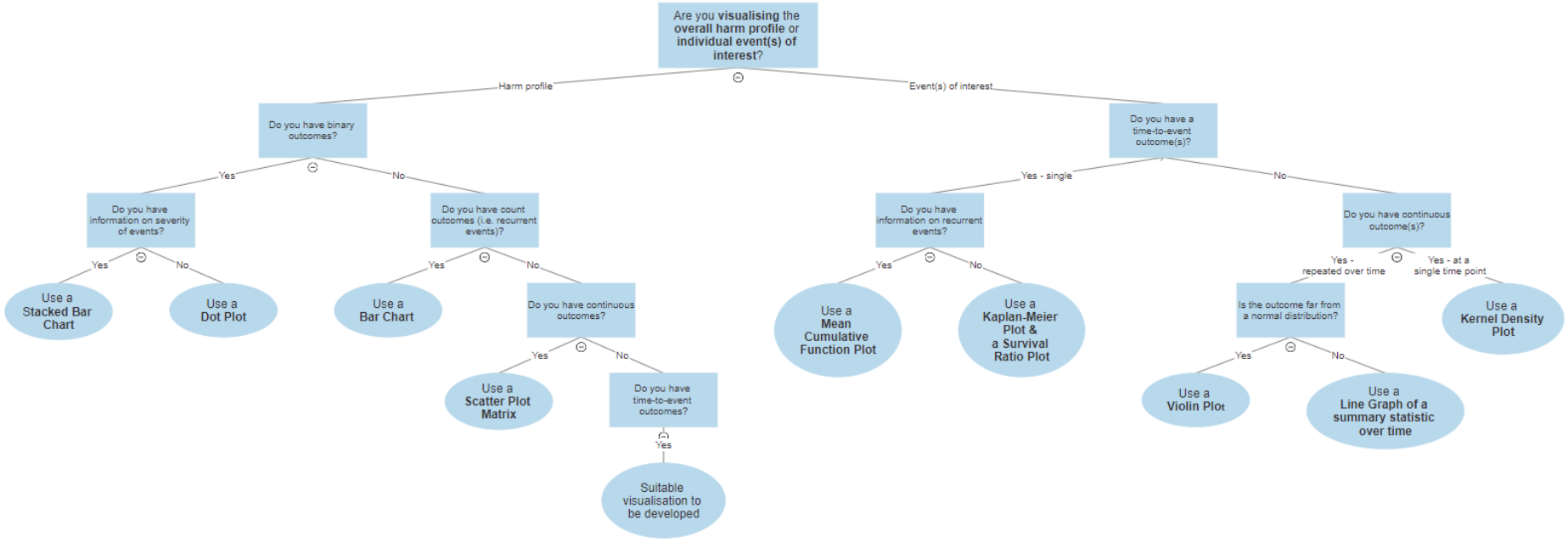
Outcome type	Plot	Recommendation
Binary	Dot	Use to present a comprehensive summary of multiple binary events
	Stacked bar chart	Use to present information on severity for multiple binary events
Count	Bar chart	Use to present information on event counts
Continuous	Matrix scatter	Use in an exploratory setting to help identify any outliers or patterns of interest across multiple continuous outcomes
Time-to-event	To be developed	No plot endorsed

Table 5.4: Visualisations to summarise individual event(s) of interest*

Outcome type	Plot	Recommendation
Time-to-event	Kaplan-Meier with extended at risk tables	Use to present information for specific events of interest as a way to detect large between treatment group differences and potential disproportionalities between treatment groups
	Event-free (survival) ratio	Use alongside the Kaplan-Meier plot to incorporate a direct estimate of the between group difference for time-to-event outcomes
	Mean Cumulative Function	Use to display time-to-event information for recurrent events. Provides a visual summary of expected time until '*x number of an event*' is experienced per participant by group
Continuous	Line	Use to describe continuous harm outcomes of interest over time, using an appropriate summary statistic including an indication of variability
	Violin	Use as an alternative plot to the line graph to present a description of continuous data if for example, the outcome of interest is far from a normal distribution and there is interest in exploring the distribution
	Kernel density	Use to explore and compare an outcome of interest at a specific time-point or to investigate how an outcome of interest changes from baseline to either a specific point in time or maximum change over the entire trial period

*Where an event may be a single adverse event e.g. a headache or a single category of events that have been grouped together e.g. neurological body-system or an aggregated summary such as the number of serious adverse events

Figure 5.22: Decision tree to help researchers decide which plot(s) to use to visualise data on harm outcomes



* Harm profile: a summary of all harm outcomes. Individual events: includes individual emerging events (including AEs and laboratory or vital sign data indicative of harm) and prespecified events of interest.

5.6.7 Areas for further development

Amongst the plots considered for displaying multiple time-to-event outcomes, consensus participants felt that the options available were poor. Whilst multiple Kaplan-Meier plots could be used to display information on a limited number of prespecified events of interest, there is still a gap in how to visualise multiple time-to-event outcomes simultaneously on the same plot. There were discussions about development of novel plots in this setting and this will be pursued in future work.

5.7 Discussion

5.7.1 Summary

The CONSORT extension to harm outcomes aimed to help improve reporting and the recommendations from the SPERT working group and Lineberry et al. provided detailed examples to sit alongside the CONSORT harms extension.[7, 28, 234] Each called for use of visualisations when reporting harm outcomes but did not give clear guidance on what visualisations would be helpful. Results of chapter four showed that researchers want guidance on appropriate methods for analysis of harm outcomes and case studies detailing examples of use, and informal feedback from journal editors indicated that specific guidance on which visualisations to use in journal publications would be useful.[216] Therefore informed by this feedback, in 2020, I led a collaboration to develop consensus recommendations on the use of visualisations for harm outcomes in clinical trial manuscripts. Recommendations were developed over a series of virtual meetings with researchers responsible for producing such reports, including clinical trial statisticians and researchers from UKCRC CTUs and industry, and clinicians.

In this chapter, I describe the final recommendations with examples of use in the RCT setting. The work in this chapter demonstrates visualisations as an alternative way to communicate risk of harm in contrast to tables of events identified as common practice in chapter two. It demonstrates

examples for a variety of different outcome types (e.g. binary, time-to-event and continuous), as well as for emerging outcomes and prespecified events of interest. Ongoing dissemination work such as a workshop at the 2021 NIHR statistics group conference and a presentation at the 2021 Society of Clinical Trials annual conference aimed to promote these recommendations to the clinical trial community with the aim of increasing the use of visualisations in clinical trial manuscripts and reports. Ultimately promoting presentation of clearer and more informative information on harm outcomes to aid interpretation. Each plot is accompanied with signposting to accessible software code to produce each of the graphics with the aim of supporting adoption and to ensure efficient implementation of the recommendations. Trialists can implement the recommendations alongside the CONSORT extension to harms and the Lineberry et al. recommendations for harm outcomes, as well as the more general guidance on the content of statistical analysis plans from Gamble et al.[28, 87, 234]

5.7.2 Application of recommendations in practice

Ultimately, the choice of visualisation will depend on the outcome type, scenario e.g. summarising multiple emerging events or one event of interest, the design of the trial (trials with more than two treatment groups require more care) and the purpose of the plot e.g. communicate information about the entire harm profile or convey a direct message about a particular event of interest. Therefore, it is for the statistician and clinical trial team to decide the most appropriate visualisation(s) for their data and objectives. It is likely that a combination of plots will be necessary, for example presenting both the traditional Kaplan-Meier plot alongside the event-free ratio plot for prespecified harm outcomes to explore the temporal relationship, in addition to the dot plot to summarise the overall harm profile.[237]

Whilst these recommendations give a clear steer on the type of plots to use with some guiding principles on format, there are still many aspects of plot design that users can vary and that the consensus group did not discuss and still needs careful consideration. For example, the colour and symbols used, the axis scales and limits, appropriate use of labels, and the number of groups compared can all impact interpretation and understanding. Much has been written on these aspects, such as the recent blog posts by Unwin and Rost, as well as lists of key principles for a good graphic.[237, 254-256] For direction on these aspects resources such as the Adobe Colour Wheel can be used to generate complementary colour palettes and Colour Oracle can help to take into consideration colour blindness when choosing colours.[257, 258]

5.7.3 Adoption and endorsement of recommendations

In chapter four, I discussed the lack of adoption of existing guidelines and recommendations for harms, which is supported by the recent review findings of Junqueira et al.[80] Involvement of the UKCRC CTU statistics group in the development of these recommendations is just one step to help instigate change. Results in chapter four also indicated that members of the trials community believed case studies and tutorial papers would go some way to support change. Therefore, in addition to the planned publication of the recommendations developed in this chapter, a case study demonstrating the practical use of visualisations and the impact on inferences drawn has been published, and includes code for implementation.[259] Journal endorsement has also been shown to increase use of guidelines and recommendations, and the ongoing support from the BMJ could have a huge impact but adoption by the wider journal community still needs to be achieved.[239] Whilst endorsement is helpful to increase awareness, authors will also need clear instructions on how they are expected to use any guidelines or implement recommendations. While visualisations can be helpful there is no expectation that they will become mandatory, instead it would be beneficial for

editors and reviewers to signpost authors to resources and case studies that demonstrate the usefulness of visualisations to encourage and inspire use.

5.7.4 Strengths and limitations

Black et al. highlighted that output from any consensus is dependent on a number of factors including: the participants, the selection and presentation of information, the structure of interactions and the method of synthesising individual judgments.[238] The predominance of statisticians over other researchers in the consensus group could be deemed a limitation of this work. However, given statisticians will ultimately be responsible for implementation of these recommendations their inputs were deemed highly relevant and their opinions of utmost importance, which is in line with the thoughts of Cleveland, who concluded in 1984 that "*statisticians can play ... the leading role, in effecting an improvement of graphical communication in science*".[260] In addition to statisticians, the BMJ's graphic designer was present across all meetings and his feedback sought continually throughout the project. Thus, staying in line with Black et al.'s suggestion that "*groups should be composed of people who are expert in the appropriate area and who have credibility with the target audience*."[238] Also, to ensure breadth of input, opinions of clinicians with experience in clinical trials was sought to seek their feedback on the recommendations to ensure understanding of each of the plots and their endorsements; their specific feedback has been incorporated into the recommendations where necessary. Choosing clinicians who are active trialists has the added potential to assist with dissemination and increases the likelihood of these plots being used in practice. Although this was limited to only two clinicians and whilst a larger number would have been advantageous this was impractical due to the time required to contribute.

The review described in chapter three plus feedback from the trials community helped identify a broad range of specific visualisations for harm outcomes, as well as a number of alternatives that could be easily adapted. Whilst participants were encouraged to put their ideas forward for adaptations there are potentially many possibilities that were not considered. In addition, in the ever-expanding field of data visualisations new ideas are constantly emerging and have potentially been omitted from consideration in these recommendations. However, an initiative run by the PSI data visualisation special interest group, of which I am a core member, called for novel ideas for visualising data from a typical AE dataset in September 2020, but no ideas that had not been considered by the consensus group were submitted, excluding interactive tools. Interactive visualisations were not considered in these recommendations as they are consider to fall into their own separate domain and require different considerations for appraisal, which are discussed in Wang et al., though, the multifaceted and complex nature of harm data lends itself to interactivity and should be considered in future work. [228]

The work in this chapter focused on the presentation of the final analysis in journal articles but some of the plots have been recommended for use in a more exploratory setting. In addition, many could be utilised for interim analysis and incorporated into reports presented to DMCs. I have not examined DMC reports in this thesis but it is clear that graphics could be of huge benefit when there is a need to review large amounts of data on harms in a short time-frame. This is supported by tools such as the template DMC reports shared by both Harrell and statisticians at the Department of Biostatistics and Medical Informatics at the University of Wisconsin–Madison and guidelines in the literature on plots to use in DMC reports.[52-54, 261] Interactive tools are also likely to have much to offer in this setting.

Due to external factors (i.e. the COVID-19 pandemic) the preferred format of a face-to-face (as proposed in the guidance from the CONSORT executive group) meeting was not possible and had to be redesigned as a series of online meetings via the Microsoft Teams platform. This presented a number of unique challenges not least because the majority of attendees were working from home with at times intermittent internet connections and competing priorities (e.g. child-care as schools and child minding facilities were closed). However, the online format offered many advantages over the originally planned face-to-face format. It facilitated the attendance of a greater number of participants and participants from a wider geographical area. It also allowed those with competing commitments to attend as suited. Conducting the meetings, over three half days instead of the one day originally planned gave us more time and helped keep participants engaged and energised throughout. The online format also enabled a wider use of measures to ensure equitable participation such as encouraging attendees to use both audio and chat functions to engage, seeking individual feedback from every participant, which they could return in their own time, and using an anonymous voting system to make decisions, which ultimately helped avoid dominant voices taking over. A threshold of 60% or more was used to indicate group endorsement as it represented a majority vote. Proportion of agreement in the 50-60% range were revisited for further discussions and votes retaken until a consensus could be reached. Providing multiple opportunities to vote on decisions allowed opinions to change and be updated as new information and ideas were presented, and reflections made. Ultimately, helping the group reach an agreement on many decisions.

5.7.5 Future work

A key next step is to get these recommendations adopted into practice. Potential strategies include development of good practice examples that incorporate visualisations into the analysis section for harm outcomes in statistical analysis plans and development of standard operating procedures detailing good practice for the analysis of harm outcomes that could be implemented across CTUs.

Anecdotal evidence suggests that some CTUs have already started to put these ideas into practice locally but a national strategy would ensure a faster path to change.

Several novel plots were considered for endorsement in this work, for example, the volcano and tendril plot shown in thumbnails 5 and 11 in figure 5.1, but ultimately the appraisals revealed their inadequacies and voting revealed an overriding preference for more traditional plots, which tended to be simpler. This could be due to an underlying bias as a result of participants being more familiar with the selected plots, as well as a perception that such plots would be more familiar to clinicians and thus more likely to be accepted by the clinical trials community. There was, however, endorsement for two unfamiliar plots, the plot of the mean cumulative frequency (figure 5.16) and the survival ratio plot, which is referred to in this setting as the event-free ratio plot (figure 5.17), and use of these is encouraged with clear explanations to ease interpretation. Clinical interviews highlighted a need for more in-depth explanation of both of these plots but both were felt to be acceptable and of great potential use once understood. Arguably, given the current lack of visualisations for displaying harm outcomes in the RCT setting as identified in chapters two and four, use of any plot to communicate information on harms in clinical trial publications can be considered novel and as such development of new plots was not pursued. In this first instance, we propose more of a gentle push towards the use of visualisations for harm data, as Unwin describes making "*the best use of know and well-understood graphics*".[237] Once use of visualisations for harm data is more common in the scientific press there will, perhaps, be an appetite for more innovative plots.

Whilst amendments to existing plots are proposed, the purpose of this work was not to develop new plots and would have been beyond the scope of a consensus, which as Black et al. described is "*a process for making policy decisions, not a scientific method for creating new knowledge*."[238] However, it was clear that there is a need for new approaches, particularly for presenting multiple

time-to-event plots or multiple continuous outcomes. Development of new plots will be undertaken in future work and recommendations will be updated in a timely manner to reflect any future progress. With a high likelihood of future updates being required, development of a website that can be more readily updated over time without need for new publications is one further avenue to explore and has previously been advocated by Chuang-Stein and Xia.[139] This would also serve as a readily available resource for dissemination. The CTSpedia Wiki page created by the FDA, industry and academics goes some way towards this, serving as a repository of potential graphics but provides limited direction on benefits of each, cautions of use and possible inferences to be drawn, it has also not been updated since 2014.[250]

5.7.6 Conclusion

Visualisations provide a powerful tool to communicate harm offering alternative perspectives to the traditional frequency tables. Implementation of these recommendations has the potential to help improve communication of harm outcomes in clinical trial manuscripts and reports, enabling clearer summaries of harm profiles to be presented. They could also help to identify the potential burden of harm participants experience and help identify potential signals for ADRs for further monitoring in future clinical trials, as well post-marketing surveillance studies. This work endorses the use of several visualisations but highlights the limitations and potential pitfalls of each, demonstrating the importance of continuing to examine crude numbers alongside visualisations.[206]

6. Utilising time-to-event methodology to detect signals for adverse drug reactions

6.1 Introduction

Visualisations explored in chapter five provide a range of ways to summarise the overall harm profile and help identify any potential signals for ADRs. However, they each rely on a subjective interpretation of either a between treatment group statistic or comparison of a statistic presented by treatment group. A more objective means to analyse emerging harm outcomes would provide a standardised approach leading to greater transparency and consistency in the results presented. Whilst statistical methods under a hypothesis-testing framework offer an objective analytical approach this can be problematic for the analysis of emerging harm outcomes if interpreted inappropriately i.e. p-values>0.05 interpreted as indicating that interventions are safe. An objective approach that sought to detect signals for potential ADRs which triggers closer monitoring in ongoing or future studies rather than making definitive conclusions would not be constrained by the same issues. Such a signal detection approach has been advocated for in the literature by Drago and colleagues.[25] The idea is that instead of thinking of the outputs of statistical tests within a hypothesis-testing framework, interpreting results as significant or not to confirm a difference between treatment groups, to instead use the output to detect signals for potential harm to prompt further investigation in line with practice in the pharmacovigilance setting. Such an approach is also supported by ICH E9 guidelines on statistical principles for clinical trials which advocates p-values as a useful "*'flagging' device applied to a large number of safety and tolerability variables to highlight differences worth further attention*."[6]

The body of emerging harm outcomes collected in clinical trials will be comprised of events not associated with the intervention and events that are, which we refer to as ADRs. Events not associated with the intervention will be expected to occur at a constant rate over time as they will

be events that occur to participants regardless of treatment start, but the causal mechanism of an ADR can mean that the occurrence of events associated to the drug is time dependent e.g. allergy related or organ function damage. The aim of the analysis of harm outcomes is to establish which events are likely caused by the intervention and which are not. Given a likely temporal relationship, analysis that incorporates the time an event occurs could help to discern which of the many emerging harms recorded are possible ADRs and warrant further exploration.[124] Exploring the value of incorporating time into the analysis of harm outcomes to help identify ADRs is supported by the results of the survey of academic and industry clinical trial statisticians reported in chapter four, and has been highlighted by several authors in the scientific literature and regulatory guidelines for wider use in this setting.[43, 118, 183, 216, 262, 263, 11-15] Therefore in this chapter I explore the value of time-to-event methodology to detect signals for ADRs.

6.1.1 What is time-to-event analysis?

Time-to-event analysis is a field of statistics that is concerned with the analysis of the time until an event of interest occurs.[264] There is a collection of mathematical expressions, analytical techniques and accompanying terminology that is unique to the field of time-to-event analysis.

There are two pieces of key data – the presence (or absence) of the event of interest, and the time to the event occurring or end of follow-up for the participant if the event does not occur.[265] This is measured from a common reference point, in RCTs this would typically be randomisation but it could be from treatment start.

Analysis methods for time-to-event data are well established and first came to prominence in the biomedical field to examine survival times in cancer studies, and thus is often referred to as 'survival analysis'.[266] Its use extends beyond the event of mortality and the techniques can be used to

analyse the time to any event of interest and thus is alternatively referred to as time-to-event analysis, which is the term I will use in this thesis. The event of interest is also often referred to as a failure event in the literature, however I will use the former throughout this chapter. Instead of thinking in terms of probability density functions and cumulative distribution functions for time-to-event data the concepts of survival functions, hazard functions, cumulative hazards and other related terms are used. These key concepts are defined below.

Key time-to-event terminology and concepts

Let T be a non-negative random variable that denotes the time to the occurrence of the event of interest. The survival function is defined as:

$$S(t) = \Pr(T > t)$$

which is the probability of surviving beyond time *t*, more generally it is the probability of the event of interest not having occurred by time t.[264] $S(t) = 1\ at\ t = 0$ and decreases toward zero as t increases.

Both the probability density function and the cumulative distribution functions can be obtained from S(t). The probability density function $f(t)$:

$$f(t) = dF(t)/dt = \frac{d}{dt}\left(1 - S(t)\right) = -S'(t)$$

and the cumulative distribution of T is:

$$F(t) = 1 - S(t)$$

Another important concept is that of the hazard function, also referred to as the hazard rate or simply the hazard. The hazard function, $h(t)$ is the instantaneous rate of failure, which is defined as the limiting probability (as the time interval tends to zero) that the event of interest occurs in a given interval, conditional upon the individual having not experienced the event (or survived) prior to the beginning of that interval:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t + \Delta t > T > t \mid T > t)}{\Delta t}$$

In the above formula, the numerator represents the conditional probability that the event will occur in the interval $[t, t + \Delta t]$ given that it has not already occurred and the denominator is the width of the interval, thus giving a rate of event occurrence per unit time. The hazard function, $h(t)$ varies in the range from $[0, \infty)$ and can be increasing, decreasing, constant or varying such that it is non-monotonic. If the hazard function is constant then the 'risk' of experiencing the event, if it has not yet happened, is equal at all time points over follow-up but if the hazard function varies then the 'risk' of experiencing the event is time-dependent.

A related function is the cumulative hazard function, which is the total amount of 'risk' that has been accumulated by time t and is defined as:

$$H(t) = \int_0^t h(u)\, du$$

and has an inverse relationship with the survival function such that $H(t) = -\ln(S(t))$. The cumulative hazard function also allows us to describe the probability density function and the cumulative distribution functions:

$$F(t) = 1 - \exp\big(-H(t)\big)$$

$$f(t) = h(t)\exp(-H(t))$$

Methods to analyse time-to-event data and model covariate effects are well established and can be grouped into parametric models, semi-parametric models, and non-parametric approaches.

Non-parametric approaches

The Kaplan-Meier method estimates the survival function, and the Nelson and Aalen method estimates the cumulative hazard function. The latter can be transformed to estimate the survival function and plotted over time to provide a visual summary of the survival function.[267-269] They are based on raw data and make no distributional assumptions about the survival times. There is no provision to adjust for model covariate effects and therefore these approaches are most useful when interest is in the overall sample estimates, or when groups are comparable such as in clinical trials where estimates are made for each treatment group. In the trial setting time-to-event curves can be compared visually using graphs such as the Kaplan-Meier plot (e.g. figure 5.15) and the equality of the survival functions can be formally tested using non-parametric tests such as the log-rank test. Whilst non-parametric approaches will not be given further consideration in this chapter, visual approaches to display the outputs were examined and recommendations are presented in chapter five.

Semi-parametric models

Semi-parametric models make no assumptions about the distributional form of the event times but instead utilise the ordering of the times. Use of a modelling approach allows covariates to be incorporated. The Cox proportional hazards model is the most common semi-parametric model.[266] Whilst no assumption is made about the distribution of event times i.e. the baseline hazard is left un-parametrised, the Cox model does assume that model covariates multiplicatively shift the baseline hazard function such that the hazard function for the j^{th} participant is:

$$h_j(t|\boldsymbol{x_j}) = h_0(t)\exp(\boldsymbol{x_j}\boldsymbol{\beta}_x)$$

where $h_0(t)$ represents the baseline hazard function which is the risk for participants when $\boldsymbol{x_j} = \mathbf{0}$, and $\boldsymbol{\beta}_x$ represents the vector of regression coefficients (i.e. each of the log hazard rate ratios) to be

estimated from the data for each of the $\boldsymbol{x_j}$ covariates. In addition, by keeping the effect of time and the effect of each of the covariates separate, the effect of each of the covariates, $\boldsymbol{x_j}$, is the same across time, t.

The Cox model makes no assumptions about the shape of the baseline hazard function over time i.e. it is not parametrised and the only constraint is that participant's hazard are proportional to each other's i.e. "*one participant's hazard is a multiplicative replica of another's*".[264] It is postulated that the absence of assumptions regarding the parametric form of the baseline hazards makes it the most commonly used framework for time-to-event analysis.[270]

Parametric time-to-event models

Parametric models make a distributional assumption about the time of events. The most common parametrisation are parametric proportional hazards models. Parametrisations such as the accelerated failure time metric are less common in the clinical trial literature and so are not considered further.

Parametric proportional hazards models can be written in terms of hazards in line with the semi-parametric Cox proportional hazards models such as:

$$h_j\left(t|\boldsymbol{x_j}\right) = h_0(t)\exp(\boldsymbol{x_j}\boldsymbol{\beta}_x)$$

but now with a functional form specified for the baseline hazard $h_0(t)$. The normality assumption made in linear regression models is unsuitable here because while event times may be constant over time, they are highly likely to be asymmetrical, they can be bimodal and they will always be positive, therefore, alternative functional forms such as an exponential, Weibull or Gompertz distribution are

used. Parametric proportional hazards models are directly comparable to the Cox regression model. In each, $\boldsymbol{x_j \beta}_x$ is the log relative hazard and the individual $\exp(\beta_x)$ are the hazard ratios for the x^{th} coefficient. It is this analogy to the widely used Cox models that make the proportional hazards framework popular. However, how each approach utilises time-to-event data differ. Whereas the semi-parametric approach compares participants at times when events happen (as do non-parametric approaches), the parametric approach utilises participants for the entire period they are under observation (i.e. prior to censoring). Therefore, parametric models are considered more efficient than semi-parametric models if a realistic distribution for the baseline hazard can be specified and these models will be more powerful. However, if the assumed parametric form is incorrect then the model estimates may be biased.

Estimates for each of the parameters of interest are obtained via maximum likelihood estimation where the likelihood of the data is:

$$L(\boldsymbol{\beta}|t_1, t_2, \ldots.) = f(t_1|\boldsymbol{\beta}, \boldsymbol{x_1}) f(t_2|\boldsymbol{\beta}, \boldsymbol{x_2}) \ldots$$

for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta_x})$. Which is equivalent to:

$$L(\boldsymbol{\beta}|t_1, t_2, \ldots.) = S(t_1|\boldsymbol{\beta}, \boldsymbol{x_1}) h(t_1|\boldsymbol{\beta}, \boldsymbol{x_1})\, S(t_2|\boldsymbol{\beta}, \boldsymbol{x_2}) h(t_2|\boldsymbol{\beta}, \boldsymbol{x_2}) \ldots$$

because $f(t) = S(t)h(t)$.[264]

Where maximum likelihood estimation is the analytical approach that given a set of observations $(t_1, t_2, \ldots . t_j)$ gives the value for each β that maximises the probability, or likelihood of observing those particular data.[271]

Whilst the proportional hazard ratio metric is popular and a key component of both the semi-parametric and parametric approaches, the fundamental proportionality assumption may not always

be valid.[270] This is potentially likely when examining the time to a potential adverse reaction where a temporal relationship between the event and the intervention is probable, I will discuss the potential implications of this in the following.[124]

Key features of time-to-event methods are that they take into account, at all time points, the number of participants at risk, thus, they accommodate censored information. Censoring occurs when the participant is no longer observed for the event of interest. For example, the outcome might be occurrence of an event of interest over 12 months, each participant is followed for 12 months at which point if they haven't experienced the event we stop observing them and they are consider as censored observations. This is more accurately called right-censoring and might also occur within a study if participants withdraw before follow-up is complete, or they become lost to follow-up or experience an event prior to the end of follow-up which precludes the event of interest from happening e.g. death. Methods typically assume censoring is random and uninformative i.e. not related to the reason for failure (occurrence of the event of interest). This assumption means both semi-parametric and parametric models can account for random, uninformative censoring through the shared contribution of the survival function to the likelihood function of both censored observations and participants with the event i.e. in a participant that experiences the event their contribution to the likelihood function is:

$$f(t_j) = h(t_j)S(t_j)$$

and in a censored observation the contribution to the likelihood function is:

$$f(t_j) = S(t_j)$$

The likelihood function can be re-expressed in a single expression such that it can be modelled and parameters estimated using maximum likelihood.[272] In the following, I have assumed complete

follow-up and further work is required to assess the impact of different patterns of censoring on each of the approaches.

6.1.2 Why are time-to-event methods potentially useful in the context of analysis of harms?

Proportional hazards is a popular metric for time-to-event analysis and has been shown to be valid in a high proportion of investigated trial scenarios in the context of efficacy outcomes.[270, 273] However, when examining harms, events unrelated to the intervention might be expected to occur at a constant rate over time but with ADRs we might expect to observe a clustering of events due to a time dependent mechanism. Therefore, in the presence of a time-dependent ADR we might expect a constant event rate in the control group and a non-constant event rate in the intervention group, in which case it is likely that the proportional hazards assumption would be violated. Whilst such a violation of the proportional hazards assumption may render the estimates of treatment effects insensitive or biased, being able to detect these disproportionalities could offer advantages when aiming to detect time-dependent ADRs.[124] Table 6.1 provides examples of typical ADRs and the times at which they typically occur relative to exposure.[274]

Table 6.1: Examples of common ADRs and time at which they typically occur relative to exposure

Time of reaction relative to exposure	Reaction
Immediate (within hours)	Anaphylactic shock
Very early reactions (within first/second week)	Extrapyramidal disorders (e.g. muscle spasms); acute renal failure; suicidal ideation
Early (within one to three months)	Stevens John syndrome; rhabdomyolysis; acute liver injury
Intermediate (up to 6 months from exposure)	Acute myocardial infarction; neutropenia
Later reactions	Peripheral neuropathy; unpredictable immune-mediated hypersensitivity

6.1.3 Prior use of time-to-event methods to raise signals of harm in exposure only cohort studies

Work in the observational setting has proposed detecting a non-constant hazard over time could be a powerful means to detect signals for harm where there is no control group or reference population for comparison i.e. in exposure only cohorts.[124, 265] The Weibull Shape Parameter (WSP) test is based on the idea that if an ADR occurs in a specific period of follow-up then the hazard will be non-constant, if the event is not an ADR then the hazard function will be constant over time. The WSP test uses the Weibull distribution to detect a non-constant rate. The baseline hazard for the Weibull model is:

$$h_0(t) = \lambda t^{\lambda-1} \exp(\beta_0)$$

The Weibull distribution requires two parameters to describe the distribution, the scale parameter parametrised here by $\exp(\beta_0)$ and the shape parameter, λ, which can be used to quantify how far the distribution is from a constant. When $\lambda = 1$ the hazard function is constant.

The WSP test fits a Weibull model to a single group dataset (i.e. no comparison group) and performs a statistical test on the hypothesis that $\lambda = 1$:

$$H_0: \lambda = 1 \; vs \; H_1: \lambda \neq 1$$

If the p-value from the Wald test, testing the null hypothesis that the estimated shape parameter equals one, is less than a prespecified significance level, e.g. 0.05, it is taken as evidence to reject the null hypothesis, indicating that $\lambda \neq 1$, and the hazard is considered non-constant and the test raises a signal for a possible ADR.

Based on initial results showing that the WSP test performed best at the extremes of follow-up the authors extended the test with the aim of improving performance across the entire follow-up

period. To do this the authors proposed censoring the data at regular intervals throughout follow-up and running the WSP test on each of the censored datasets, raising a signal if the p-value from testing $\lambda = 1$ is less than a prespecified significance level, e.g. 0.05, in any of the datasets. The aim being to improve detection of symmetrical, non-constant hazard functions. The authors described this as the WSP tool. Whilst the WSP test has been shown to be most powerful at detecting signals for ADRs shortly after treatment initiation the tool-based approach showed improved power overall. The WSP test and WSP tool were proposed for use on single arm exposed cohorts. In this chapter, I will adapt this idea to the RCT setting to analyse emerging harms utilising the control group to detect signals for potential ADRs.

6.1.4 Time-to-event methods in the RCT setting for the analysis of harms

In the RCT setting there is an ongoing industry/academic collaboration exploring the impact of different time-to-event methods in the presence of censoring and competing risks on the analysis of harm outcomes.[275] However the focus of that work is on quantifying the risk of the event and comparing different methods as predictors of the probability of an event, focusing on the analysis of prespecified events of interest.[197]

6.2 Aims

The aim of the work presented in this chapter is to examine a range of methods, including adaptations of existing methods, to assess and compare their utility at effectively identifying ADRs from the body of emerging harms reported during a trial. Specifically, the aims are:

1. To explore a range of statistical models under a time-to-event framework for use as signal detection tests for analysis of emerging harms. With the aim of detecting signals for ADRs based on the identification of a disproportional hazard between the treatment and control group.

2. To compare the performance of the signal detection tests to identify signals of ADRs in a range of RCT settings.

Novel approaches motivated by work in the observational setting that utilise the parametric Weibull time-to-event model and the semi-parametric Cox proportional hazards model, will be compared to widely used approaches that do not utilise information on the time events occur (e.g. the Fisher's exact test) and a selection of easily implementable Bayesian approaches (e.g. the beta-binomial model), as well as recently developed methods designed to identify treatment effects in the presence of non-proportional hazards (e.g. the combined test).

6.3 Development of a novel approach to detect signals for potential ADRs

In the following I modify work from the observational setting (described in section 6.1.3), which aimed to detect a non-constant hazard in the single arm setting to detect ADRs. I explore whether the control group can be incorporated in such a way as to detect a disproportionality in the hazards, which could be indicative of an ADR. The idea is to extend the principle of using a non-constant hazard to detect signals of ADRs in the observational setting to a RCT, where it may be reasonable to expect that the background rate of unrelated harm i.e. the AE rate in the control group, to be constant over time. Instead of aiming to detect a non-constant hazard in the single arm setting or a proportional difference in hazards between the treatment and control group, I propose that detecting a disproportionality in the hazard rates between treatment groups could be more relevant as this would be indicative of a time-dependent effect in the intervention group under the assumption of a constant hazard in the control group. None of the statistical methods developed to analyse harm outcomes identified in chapter three adopted such an approach.[130]

i) Weibull time-to-event model with ancillary parameter

The Weibull model is a parametric time-to-event model that assumes a functional form for the baseline hazard of:

$$h_0(t) = \lambda t^{\lambda-1}\exp(\beta_0)$$

Where λ is an ancillary shape parameter estimated from the data and the scale parameter is parametrised as $\exp(\beta_0)$. A Wald test can be undertaken (and is output as default when fitting such a model in Stata) to test whether the shape parameter significantly differs from one i.e. a non-constant hazard. If treatment group is included in the model as a covariate, under the proportional hazards assumption it is possible to determine if there is a statistically significant proportional difference in the estimated hazards between treatment groups, which can be used to detect a proportional treatment effect. For example, the hazard for the Weibull proportional hazards model is:

$$h(t_j|\boldsymbol{x_j}) = h_0(t)\exp(\boldsymbol{x_j}\boldsymbol{\beta}_x) = \lambda t_j^{\lambda-1}\exp(\beta_0 + \boldsymbol{x_j}\boldsymbol{\beta}_x)$$

If treatment group is included in the covariate list, $\boldsymbol{x_j}$ then an estimate for the treatment effect under the assumption of proportional hazards can be made i.e. allow the scale to change:

$$h(t_j|\boldsymbol{x}_j) = \lambda t_j^{\lambda-1}\exp(\beta_0 + TRT_j\beta_1)$$

where TRT_j indicates treatment group covariate and $\exp(\beta_1)$ is the estimated hazard ratio for the treatment covariate. However, testing for a significant hazard ratio will not detect a disproportionality between the hazard functions for each treatment group over time which would indicate a time-dependent ADR.[124]

Ancillary parameters can be used to specify linear predictors for the other parameters in the assumed distribution. For the Weibull model, there is one ancillary parameter, the shape parameter, λ. In the above specifications, λ has been assumed to be constant across covariates. Including

treatment group as an ancillary parameter in the model allows the shape to differ between treatment groups i.e. in a study with two treatment groups this would allow one constant value for the control group and another constant value for the intervention group. If the event is an ADR then it might be reasonable to expect hazards to be disproportional and the shape of the hazard, λ, to differ between treatment groups. This model could be utilised to detect a disproportionality, as this would allow a separate estimate of the baseline hazard for each treatment group. Relaxing the proportional hazards assumption and allowing the shape parameter to differ between treatment groups via an ancillary parameter gives the following hazard function:

$$h(t_j|\boldsymbol{x}_j) = \exp(\beta_0 + \boldsymbol{x_j}\boldsymbol{\beta_x})(\alpha_0 + \boldsymbol{y_j}\,\boldsymbol{\alpha_x})t_j^{(\alpha_0 + \boldsymbol{y_j}\,\boldsymbol{\alpha_x} - 1)}$$

The shape parameter, λ is parametrised as $\ln(\lambda) = \alpha_0 + \boldsymbol{y_j}\,\boldsymbol{\alpha_x}$**.** The treatment group can now be included in the covariate list $\boldsymbol{x_j}$ and $\boldsymbol{y}_j$ such that:

$$h(t_j|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0 + TRT_j\,\alpha_1 - 1)}$$

thus allowing treatment group to have an effect on both the scale and shape of the hazard. The Wald test can then be used to identify the presence of a significant shape parameter for the treatment covariate, which would be indicative of a disproportionality i.e. testing the following hypothesis:

$$H_0\colon \alpha_1 = 1 \; vs \; H_1\colon \alpha_1 \neq 1$$

Using $p\ value \leq 0.05$ to indicate a significant shape parameter for the treatment covariate, raising a signal for a potential ADR in the simulation work described below. It is not necessary to constrain $\boldsymbol{x_j}$ and $\boldsymbol{y}_j$ to contain the same covariates but in this work they will both only comprise of the treatment group covariate so that it is clear that any difference in the shape of the hazard between treatment groups is not a proportional effect being constrained.

A signal is raised if: $p\ value \leq 0.05\ for\ H_0\colon \alpha_1 = 1$

ii) Double-Weibull time-to-event model with ancillary parameter

Motivated by work in the observational setting extending the WSP test to the WSP tool (described in section 6.1.3) and initial simulation results that indicated good performance of the Weibull model with ancillary parameter (described in (i) directly above) at the extremes of time but reduced power toward the middle of the observation period, a simple modification of this model is explored, which will be referred to as the double-Weibull model.[124]

The double-Weibull model performs two tests, the first where the data is censored halfway through follow-up i.e. when $t = 0.5$ and the second on the full follow-up period i.e. when $t = 1$, with the aim of improving the test performance away from the extremes of time. This means for the first test that the observation period is constrained to the period $0 \leq t \leq 0.5$ such that each participant is followed until $t = 0.5,$ at which point if they have not experienced the event they are consider as censored observations. The Weibull time-to-event model with ancillary parameter is then fitted on this constrained dataset such that:

$$h(t_j|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)\,(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0 + TRT_j\,\alpha_1 - 1)}\ when\ t \leq 0.5$$

The model is then also fitted on the complete dataset such that:

$$h(t_j|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)\,(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0 + TRT_j\,\alpha_1 - 1)}\ when\ t \leq 1$$

This test will raise a signal for an ADR if either model indicates a significant treatment shape parameter with $p\ value \leq 0.025$ for the treatment covariate, α_1 i.e. a signal is raised if:

$$p\ value\ \leq 0.025\ for\ H_0\colon \alpha_1 = 1\ when\ t\ \leq 0.5\ \ OR\ \ p\ value\ \leq 0.025\ for\ H_0\colon \alpha_1 = 1\ when\ t\ \leq 1$$

6.4 Existing methods that could be used to detect signals for potential ADRs

In the above, I have proposed a novel signal detection approach for the analysis of emerging harm outcomes. It is important to understand how these new approaches perform and how they compare to other available approaches. In the following I outline alternative statistical methods that are already in use that could also be utilised as signal detection tools to identify treatment effects that could indicate an ADR, each is summarised in table 6.2.

i) Chi-squared and the Fisher's exact test

The chi-squared test and Fisher's exact test are widely used to compare the proportion of events experienced in each treatment group in RCTs (results of chapter two). The tests compare the proportion of those with at least one event over the entire follow-up period, and assume that all participants are followed-up for the entire study period i.e. no withdrawals, loss-to-follow-up etc. These tests do not account for time and others have commented that "*the use of naive proportions is inappropriate*" but they are included here for reference due to their continued prevalent use.[123, 276]

The chi-squared test is a hypothesis test that aims to identify differences in proportions between groups based on a two-by-two table of treatment and event frequencies.[277] The chi-squared test calculates the expected frequencies of events under the null hypothesis of 'no difference' between treatment groups and calculates the difference with observed frequencies to obtain the test-statistic. The test statistic is then compared to the chi-squared distribution, where the chi-squared test statistic, χ^2, is calculated as:

$$\chi 2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

where n_{ij} represents the number of observations in the i^{th} row and j^{th} column of the two-by-two table and m_{ij} represents the expected number of observations in the i^{th} row and j^{th} column under the assumption of no difference such that $m_{ij} = n_{i.}n_{.j}/n$.[278-280]

A signal is raised if: $p\ value = P(\mathrm{X}^2 \geq \chi^2) \leq 0.05$

The chi-squared test is based on a 'large sample' approximation, so when expected frequencies are small the Fisher's exact test is often used as a more robust alternative. Fisher's exact test is also a hypothesis test approach where the null hypothesis is that there is no difference in proportions between treatment groups. Fisher's exact test evaluates probabilities of all possible two-by-two tables with the same row and column totals as the observed data.[281] The probability of every possible table is computed as:

$$P = \sum_{T \in A} \Pr(T)$$

holding row and column marginals fixed where A is the set of all tables with the same marginals as the observed table, T^*, and in the same tail as T^*, such that $\Pr(T) \leq \Pr(T^*)$.[279, 282]

A signal is raised if: $p\ value = P = \sum_{T \in A} \Pr(T) \leq 0.05$

ii) Detecting disproportional hazards in a Cox proportional hazards model

The Cox proportional hazards model introduced in section 6.1.1 is a common semi-parametric approach to analyse time-to-event outcomes that assumes that the general shape of the hazard is the same for everyone and that covariate effects act multiplicatively i.e. proportional hazards. The hazard function for the Cox proportional hazard model is expressed as:

$$h(t_j|\boldsymbol{x_j}) = h_0(t) \exp(\boldsymbol{x_j}\boldsymbol{\beta}_x)$$

A common way to assess the validity of the proportional hazard assumption made by the Cox model is to analyse the residuals from the fitted model. This can be done in Stata via the generalisation proposed by Grambsch-Therneau.[283] After fitting a Cox model, the Grambsch-Therneau approach estimates the Schoenfeld residuals, fits a smooth function to these residuals and then tests whether there is non-zero slope.[284] The Schoenfeld residual for covariate $x_u, u = 1, \dots n$ and observation j, which has been observed to 'fail' is:

$$r_{uj} = x_{uj} - \frac{\sum_{i \in R_j} x_{ui} \exp(x_i \widehat{\beta_x})}{\sum_{i \in R_j} \exp(x_i \widehat{\beta_x})}$$

which is the difference between the covariate value for the observed failure, j and the weighted average of the covariate values over all participants at risk when participant j failed. If the covariate effect x_u varies with time then the coefficient β_u can be expressed as:

$$\beta_u(t) = \beta_u + q_j g(t)$$

where q_j is a coefficient and $g(t)$ is a function of time. Under proportional hazards $q_j = 0$. Grambsch-Therneau provide a way to scale r_{uj} that can then be plotted against time to give a visual assessment of the proportional hazard assumption with evidence of a non-zero slope indicative of a violation. This can also be formally tested such that:

$$H_0 : q_j = 0$$

Utilising Stata's `estat phtest` command and taking $\Pr(q_j = 0) \leq 0.05$ to indicate a violation in the proportional hazards assumption.

A signal is raised if: $\Pr(q_j = 0) \leq 0.05$

iii) Double-Cox - detecting disproportional hazards in a Cox proportional hazards model

To explore the influence a time-dependent effect has on the Cox proportional hazards model and the accompanying Grambsch-Therneau test to detect a deviation from the proportional hazards assumption, a simple modification of this model is explored. The rationale for this new approach was that it could potentially help to identify early effects that flatten out over time and which might be missed when examining the complete follow-up period. I refer to this novel approach as the double-Cox model.

Like the double-Weibull, the double-Cox model censors the data halfway through follow-up i.e. when $t = 0.5$. This simply means that the observation period is constrained to the period $0 \leq t \leq 0.5$ such that each participant is followed until $t = 0.5,$ at which point if they have not experienced the event they are consider as censored observations. The Cox model is fitted on this constrained dataset such that:

$$h(t_j|\boldsymbol{x_j}) = h_0(t)\exp(\boldsymbol{x_j}\boldsymbol{\beta}_x)\ when\ t \leq 0.5$$

Then the Grambsch-Therneau test estimates the Schoenfeld residuals, fits a smooth function to these residuals and then tests whether there is non-zero slope:

$$H_0 : q_j = 0$$

This is repeated on the complete dataset such that:

$$h(t_j|\boldsymbol{x_j}) = h_0(t)\exp(\boldsymbol{x_j}\boldsymbol{\beta}_x)\ when\ t \leq 1$$

and the Grambsch-Therneau test is performed. The approach raises a signal for a potential ADR when either of the tests indicates a non-zero slope.

A signal is raised if:

$$\Pr(q_j = 0) \leq 0.025\ when\ t \leq 0.5\ OR\ \Pr(q_j = 0) \leq 0.025\ when\ t \leq 1$$

iv) Combined test to detect treatment effects in presence of disproportional hazards

The combined test was first proposed by Royston and Palmer in 2016 to design and analyse trials with time-to-event outcomes in the presence of non-proportional hazards. It has been chosen for inclusion here as it aims to preserve power in the presence of disproportionality and therefore could be helpful to identify treatment effects indicative of ADRs.[285] The test is based on both a non-parametric permutation test of the restricted mean survival time (RMST) and the Cox proportional hazards model. In the presence of disproportionality, it has been shown that the power of the Cox proportional hazards model can be reduced, however a test based on the RMST would not be impacted and is a useful alternative in the presence of disproportionality.

The RMST is the mean survival time from randomisation to a specific point of interest, say t^*. The treatment effect is the estimated change in the RMST at t^* for the treatment group compared to control group. The RMST u at t^* can be defined as:

$$u(t^*) = \int_0^{t^*} S(t)dt$$

Then the treatment effect at t^* can be expressed as:

$$\Delta RMST = \int_0^{t^*} S_1(t)dt - \int_0^{t^*} S_0(t)dt$$

where $S_1(t)$ and $S_0(t)$ represent the survival functions in the treatment and control group respectively and a suitable test statistic to test the hypothesis:

$$H_0: S_0(t) = S_1(t) \; vs \; H_1: S_0(t) \neq S_1(t)$$

might be sought. However, specifying only one time, t^*may be problematic and could miss important differences at other values of t. A test that searches over a range of values of t^* is more likely to identify important differences in the time-to-event curves. To identify important differences

the authors proposed a test that searches over time to find the time, t^*, that maximises the chi-square statistic for testing the RMST difference i.e.

$$C_{max} = \max(Z^2) \; where \; Z = \Delta RMST / SE(\Delta RMST)$$

Given this search requires multiple tests to be performed, to avoid an inflated type I error a permutation test approach is adopted to correct the p-value.

For the permutation test the authors examined multiple values for n_t, i.e. the number of points that the RMST is evaluated at and comment that they found that $n_t = 10$ equally spaced times performed sufficiently and is thus set as the default value when undertaking this analysis in Stata using the user written package provided by the authors.[286] The test randomly permutes the treatment covariate a large (M) number of times, creating M values of C_{max} under the null hypothesis of no difference. Then in each dataset the C_{max} is calculated giving a sample of $C_1, \ldots . C_M$. Then $N = \sum_i^M I(C_i > C_{max})$ is the number of permuted samples in which C_i exceeds C_{max} and $I(.)$ is the indicator function. The larger the N the weaker the evidence that C_{max} is extreme and the larger the p-value. The p-value for the permutation test is:

$$P_{perm} = (N + 0.5)/(M + 1)$$

where 0.5 is a continuity correction.

The combined test undertakes the test described above plus it fits a Cox proportional hazards model to estimate the treatment effect. The combined test then takes the minimum of the p-values from the permutation test of the RMST and the Cox proportional hazards model:

$$P_{min} = \min\left(P_{cox}, P_{perm}\right)$$

where P_{cox} is the p-value for the treatment covariate after fitting the Cox proportional hazards model. However, since P_{cox} and P_{min} are positively correlated a correction is needed to obtain a 0.05 type I error probability, the authors propose a correction based on an incomplete inverse beta function. The p-value from the combined test can then be used to identify a treatment effect in the presence of disproportionality.

A signal is raised if: $p\ value = P_{min} \leq 0.05$

The combined test has been shown to have increased power in the presence of non-proportional hazards when there is an early treatment effect compared to a Cox model and retains similar power to a Cox model in other scenarios including in the presence of proportional hazards. Full details can be found in Royston and Parmar (2016).[285]

v) Beta-binomial model to estimate the probability that a threshold of risk is exceeded

In chapter 3, potentially useful Bayesian approaches to analyse prespecified events of interest were identified. For example, if the number of events experienced in each treatment group and the total number of participants per group are known (e.g. from a historical trial), then a beta-binomial model can be fitted. Alternatively, in the event of no prior information being available for prespecified events or emerging events non-informative priors can be used. This latter approach is of potential use to identify ADRs from the body of emerging harm outcomes.

The beta-binomial method assumes that the event rate e.g. proportion of participants with an event follows a binomial distribution:

$$Y^T \sim Binomial(n, \pi_T) \text{ and } Y^C \sim Binomial(n, \pi_C)$$

and assumes a beta prior for the event rate:

$$\pi_T \sim Beta(\alpha_T, \beta_T) \text{ and } \pi_C \sim Beta(\alpha_C, \beta_C)$$

where π_T and π_C are the event rates in the treatment and control group respectively. Parameter values (α, β) are based on the number of events and the total participants observed for each treatment group from, for example, historical data. In an ongoing study, at each analysis, say time t, the prior distributions can be updated using observed information on number of events experienced (y_t) and number of participants enrolled (n_t) to give a beta posterior distribution:

$$(\pi|y_t, n_t) \sim Beta(\alpha + y_t,\ \beta + n_t - y_t)$$

The posterior distribution is then used to calculate the probability that a predefined 'tolerable risk difference' (δ) is crossed:

$$P(\Delta(\pi_T,\ \pi_C) >\ \delta \mid\ y_t^T, y_t^C,\ n_t^T,\ n_t^C)$$

where $\Delta(\pi_T,\ \pi_C)$ is a function that measures the difference between π_T and π_C such as the risk difference or risk ratio.[287]

Applying this approach to the emerging harm setting (i.e. events not pre-specified) one needs to make an assumption about the form of the prior distribution. A commonly used 'non-informative' or minimally informative prior is Jeffreys prior, which assumes $\pi_T \sim Beta(0.5, 0.5)$ and $\pi_C \sim Beta(0.5, 0.5)$. In the simulation work to compare the performance of each of these methods (described in section 6.6) Jeffreys prior is assumed. Alternatives have been proposed such as the neutral prior, which assumes $Beta(1/3, 1/3)$ or Bayes Laplace prior, which assumes $Beta(1, 1)$ but have not been investigated in this work.

A signal is raised if: $P(\Delta(\pi_T/\pi_C) >\ 1 \mid\ y_t^T, y_t^C,\ n_t^T,\ n_t^C) \geq 0.9$

This indicates that a signal is raised if the probability that the risk ratio exceeds one is at least 90%. For completeness, I also examined $P(\Delta(\pi_T/\pi_C) > 1.25, 1.5, 2 \mid y_t^T, y_t^C, n_t^T, n_t^C) \geq 0.9$ and present these results in appendix A7.5. Like the chi-squared and Fisher's exact tests this approach does not utilise information on the time of event occurrence but is included here to explore its potential utility and for comparative purposes.

vi) Gamma-Poisson model to estimate the probability that a threshold of risk is exceeded

The gamma-Poisson model follows a similar approach to the beta-binomial approach described above but can incorporate information on exposure or follow-up time.[171] The event rate in each treatment group e.g. number of events per unit time is assumed to follow a Poisson distribution:

$$Y \sim Poisson\ (\lambda)$$

and the event rate per unit time, λ, for each treatment group has a Gamma prior:

$$\lambda \sim Gamma(\alpha, \beta)$$

Parameter values, (α, β), are based on number of events and total participant exposure time observed for each treatment group in, for example, historical data. In an ongoing study at each analysis, say time t, the prior distributions can be updated using observed information to give Gamma posterior distributions:

$$(\lambda | y_t, \tau_t) \sim Gamma(\alpha + y_t, \beta + \tau_t)$$

The posterior distribution is then used to the calculate the probability that a predefined 'tolerable risk difference' (δ) is crossed. In the simulation work, the corresponding Jeffreys prior for the event rate is used: $\lambda_t \sim Gamma(0.5, 0.00001)$ and $\lambda_c \sim Gamma(0.5, 0.00001)$.

A signal is raised if: $P(\Delta(\lambda_T/\lambda_C) > 1 \mid y_t^T, y_t^C, \tau_t^T, \tau_t^C) \geq 0.9$

A signal is raised if the probability that the incident rate ratio exceeds one is at least 90%. For completeness, I also examined $P(\Delta(\lambda_T/\lambda_C) > 1.25, 1.5, 2 \mid y_t^T, y_t^C, \tau_t^T, \tau_t^C) \geq 0.9$ and present these results in appendix A7.5. In the simulation work, time is taken as the duration between the start of treatment and the time of the event or participants are assumed to be followed to study end if no event occurs. Whilst this approach accounts for time of event it assumes a constant event rate over time, an assumption that has been shown to be valid in limited circumstances, and as such would be unable to detect time-dependent effects but it is included here to explore its potential utility.[288] It could also be used to model recurrent events by substituting time-to-event for overall exposure time or follow-up time.

6.5 Software for implementation of tests to detect signals for potential ADRs

The methods described in sections 6.3 and 6.4 were implemented in Stata version 15.1 using standard Stata commands, apart from for the combined test which utilised the user written command `stctest`.[286] For the Bayesian methods described in section 6.4 the suite of commands described in `help winbugs` which describes routines for running WinBUGS or OpenBUGS from within Stata were utilised.[289]

Table 6.2: Summaries of the methods considered as candidate signal detection tests to identify time-dependent treatment effects indicative of ADRs

	Method	Testing for:	Time	Prior	Model/Test statistic	Null hypothesis	Signal for ADR if
1	**Weibull time-to-event model with ancillary parameter**	Difference in baseline hazards (i.e. a non-proportional hazards)	Yes	NA	$h(t_j\|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0+TRT_j\,\alpha_1-1)}$	H_0: $\alpha_1 = 1$	$pvalue \leq 0.05\ for\ H_0$: $\alpha_1 = 1$
2	**Double-Weibull time-to-event model with ancillary parameter**	Difference in baseline hazards (i.e. a non-proportional hazards)	Yes	NA	$h(t_j\|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0+TRT_j\,\alpha_1-1)}$ $when\ t \leq 0.5$ AND $h(t_j\|\boldsymbol{x}_j) = \exp(\beta_0 + TRT_j\beta_1)(\alpha_0 + TRT_j\,\alpha_1)t_j^{(\alpha_0+TRT_j\,\alpha_1-1)}$ $when\ t \leq 1$	H_0: $\alpha_1 = 1$	$pvalue \leq 0.025\ for\ H_0$: $\alpha_1 = 1$ $when\ t \leq 0.5$ OR $pvalue \leq 0.025\ for\ H_0$: $\alpha_1 = 1$ $when\ t \leq 1$
3	**Chi-squared test**	Difference in proportions	No	NA	$\chi 2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$	$m_{ij} = n_{i.}n_{.j}/n$	$P(\mathrm{X}^2 \geq \chi^2) \leq 0.05$
4	**Fisher's exact test**	Difference in proportions	No	NA	$P = \sum_{T \in A} \Pr(T)$	No association between the rows and columns of the 2x2 table	$P \leq 0.05$
5	**Detecting disproportional hazards in a Cox proportional hazards model**	Non-zero slope between smoothed residuals and time (i.e. a non-proportional hazards)	Yes	NA	$\beta_u(t) = \beta_u + q_j g(t)$	$H_0 : q_j = 0$	$\Pr(q_j = 0) \leq 0.05$
6	**Double-Cox - detecting disproportional hazards in a Cox proportional hazards model**	Non-zero slope between smoothed residuals and time (i.e. a non-	Yes	NA	$\beta_u(t) = \beta_u + q_j g(t)\ when\ t \leq 0.5$ AND $\beta_u(t) = \beta_u + q_j g(t)\ when\ t \leq 1$	$H_0 : q_j = 0$	$\Pr(q_j = 0) \leq 0.025\ when\ t \leq 0.5$ OR $\Pr(q_j = 0) \leq 0.025\ when\ t \leq 1$

		proportional hazards)					
7	**Combined test to detect treatment effects in presence of disproportional hazards**	Hazard rate ratio in presence of disproportionality	Yes	NA	$P_{min} = \min(P_{cox}, P_{perm})$	$H_0: S_0(t) = S_1(t)$ vs $H_1: S_0(t) \neq S_1(t)$	$P_{min} \leq 0.05$
8	**Beta-binomial model to estimate the probability that a threshold of risk is exceeded**	Risk ratio > 1	No	Yes	Proportion of participants with an event follows a binomial distribution: $Y^T \sim Binomial(n, \pi_T)$ and $Y^C \sim Binomial(n, \pi_C)$ Assuming a beta prior for the event rate: $\pi_T \sim Beta(\alpha_T, \beta_T)$ and $\pi_C \sim Beta(\alpha_C, \beta_C)$	Posterior distribution $(\pi \mid y_t, n_t) \sim Beta(\alpha + y_t, \beta + n_t - y_t)$	$P(\Delta(\pi_T/\pi_C) > 1 \mid y_t^T, y_t^C, n_t^T, n_t^C) \geq 0.9$
9	**Gamma-Poisson model to estimate the probability that a threshold of risk is exceeded**	Incident rate ratio > 1	Yes	Yes	Number of events per unit time is assumed to follow a Poisson distribution: $Y \sim Poisson\ (\lambda)$ Event rate per unit time, λ, for each treatment group has a Gamma prior: $\lambda \sim Gamma(\alpha, \beta)$	Posterior distribution: $(\lambda \mid y_t, \tau_t) \sim Gamma(\alpha + y_t, \beta + \tau_t)$	$P(\Delta(\lambda_T/\lambda_C) > 1 \mid y_t^T, y_t^C, \tau_t^T, \tau_t^C) \geq 0.9$

Acronym: ADR – adverse drug reaction

6.6 Simulations to assess the performance of the described tests to detect signals for ADRs in RCTs

Datasets were simulated to mimic simple RCT designs with the aim of assessing and comparing the performance of tests 1 to 9 summarised in table 6.2 as a means to detect signals for ADRs. Where the criteria to flag a signal for an ADR is described in the final column of table 6.2 labelled – '*Signal for ADR if*'. The aim of the simulations was to identify if any of the aforementioned approaches adequately detected signals for ADRs when present and correctly did not detect signals when not present, as assessed by the corresponding type I (false-positives) and II errors (1 – power). Specifically, the aims were to:

1. To determine the performance of the proposed tests across a range of trial scenarios as measured by power, false positive rate and accuracy.
2. To compare test performance across varying trial scenarios to identify the 'best' overall test and 'best' test in specific trial scenarios.
3. To estimate sample sizes above which tests will have sufficient power to detect a true signal.

Simulation studies were used to create datasets by pseudo-random sampling from known distributions to examine the performance of statistical methods where some 'truth' is known. In this study the 'truth' under consideration is the presence or absence of a time-dependent treatment effect. The simulation study was designed and performed in line with the recommendations from Burton et al. and the Morris et al. tutorial for simulation studies.[229, 290]

6.6.1 Methods to generate the datasets - data generating mechanism (DGM)

Datasets were generated using assumed distributions and associated parameters such that they mimicked simple two-arm RCTs with equal treatment allocation (i.e. 1:1 allocation ratio) to a hypothetical intervention or control group. Datasets were generated as follows:

- Data were generated on:

$$Total\ trial\ sample\ size\ =\ n_{obs}\ =\ 200, 400, 800, 1000, 2000\ and\ 5000\ participants$$

- Let $X_j\ \epsilon\ (0,1)$ be an indicator variable denoting treatment assignment for the j^{th}participant, generated using:

$$X_j \sim Bernoulli(0.5)$$

  where 1 indicates intervention group and 0 control group.

- Time of occurrence of background events (AEs not associated to the treatment start) were generated over the time period 0 to 1 using:

$$AE_i \sim Uniform(0,1)$$

- The number of background events were set at the following proportions of total trial sample size:

$$0.01, 0.05\ and\ 0.10$$

- The time of ADRs in the intervention group were generated using different mechanisms dependent on the proposed timing of reaction. Reaction times considered reflected: immediate, early, intermediate or late reactions relative to study length. Times of ADRs were generated using:

$$ADR_i \sim Normal(\mu, \sigma)$$

  where assigned values for parameters mean, μ and standard deviation, σ are summarised in table 6.3.

- The number of ADRs was set as proportion increases above the background rate:

$$0.25, 0.5\ and\ 1$$

  These are equivalent to 25%, 50% and 100% increases.

If the time generated for the ADR was outside of the observation period i.e. $ADR_i > 1$ then the value was censored to ensure the time of occurrence was restricted to the interval $0 < ADR_i \leq 1$ such that all participants either had the event or completed follow-up before experiencing the event, so censoring only occurs for trial completion. Different patterns of censoring were not considered but should be explored in future work. Table 6.3 summarises each of the data generating mechanisms (DGMs) to be used.

Table 6.3: Scenarios to be simulated and data generating mechanisms (DGM) used to create them

	Scenario	**DGM for time of AE_i across treatment groups** (background event rate)	**DGM for time of ADR_i in the intervention group** (above the background event rate)
1	**Immediate disproportionality** - an immediate increase in number of events in intervention group - equates to ADRs happening on mean of day 1 ± 0.5 day in a 12 month trial	Uniform(0,1)	Normal(0.0027, 0.0014)
2	**Very early disproportionality** - an early increase in the number of events - equates to ADRs happening at mean 1 months ± 2 weeks in a 12 month trial	Uniform(0,1)	Normal(0.0833333, 0.0416)
3	**Early disproportionality** - an early increase in the number of events - equates to ADRs happening at mean 3 months ± 2 weeks in a 12 month trial	Uniform(0,1)	Normal(0.25, 0.0416)
4	**Intermediate disproportionality** - an intermediate increase in the number of events equivalent to the disproportionality occurring approximately half-way through the trial period - equates to ADRs happening at a mean of 6 months ± 2 weeks in a 12 month trial	Uniform(0,1)	Normal(0.5, 0.0416)
5	**Late disproportionality** - a late increase in the number of events equivalent to the disproportionality occurring approximately towards the end of the trial period -equates to ADRs happening at a mean of 11 months ± 2 weeks in a 12 month trial	Uniform(0,1)	Normal(0.916, 0.0416)
6	**No increase** – only background events occurring in both treatment groups	Uniform(0,1)	Uniform(0,1)

Acronym: ADR – adverse drug reaction; AE – adverse events (background events); DGM- data generating mechanisms

6.6.2 Scenarios to be investigated - simulated scenarios

Simulated datasets have been designed to reflect a range of typical trial scenarios. Characteristics of simulated scenarios are displayed in table 6.4 and will be considered in a fully factorial manner.

Combining these trial scenarios across the six different DGMs gives a total 288 scenarios consistent with a fully factorial design (see tables 6.5 and 6.6 for details).

The times for ADRs have been chosen to reflect the timing of common adverse reactions as summarised in table 6.1.[274] For example:

- An immediate reaction might be an anaphylactic shock, which typically occurs within an hour of drug exposure.
- Very early reactions can include extrapyramidal disorders that occur within 5 days of exposure, acute renal failure that occurs within 1-7 days of exposure, and suicidal ideation that generally occurs within two weeks of exposure.
- Early reactions can include Stevens John syndrome that usually occurs within 3 to 42 days of exposure, rhabdomyolysis with statin use that typically occurs between 1 and 60 days from exposure, and acute liver injury typically occurring between 5 and 90 days from exposure.
- Intermediate reactions include acute myocardial infarction when exposed to non-steroidal anti-inflammatory drugs typically occurring less than 100 days from exposure, and neutropenia, which presents within the first six months and usually within the first 3 months.
- Later reactions can include peripheral neuropathy often occurring months after first exposure and sometimes not occurring until after the end of treatment, or unpredictable immune-mediated hypersensitivity reactions with Isoniazid that are known to have a latency period of 12 months.

Table 6.4: Study characteristics of simulated scenarios

Characteristics of simulated datasets	Range of values
Trial total sample size (total) - n_{obs}	200, 400, 800, 1000, 2000, 5000
Treatment group – X_j	0, 1
Background AE rate across treatment groups	0.01, 0.05, 0.1
ADR rate increase above background rate in the intervention group	0.25, 0.5, 1
Time of increase relative to study length (normal distribution) - ADR_i	0.0027, 0.083, 0.25, 0.5, 0.916
Time of increase relative to study length (equivalent to a study with 12 months follow-up) - ADR_i	Day 1 ± 0.5 day, month 1, 3, 6 and 11 ± 2 weeks

Acronym: ADR – adverse drug reaction

Sample sizes ranging from 200 to 5000 were chosen to reflect small to large size phase III RCTs based on the results of the review of trials presented in chapter two. Background event rates (1%, 5% and 10%) were chosen to reflect a range of event rates that would be considered common AEs as per the EMA's summary of product characteristics (SmPC) guidelines.[291] The increases of 25%, 50% and 100% above the background event rate were chosen to reflect increases in event numbers in the smallest sample sizes and background event rate scenarios that were likely to be detectable. Figure 6.1 summarises the Kaplan-Meier survival estimates for the simulated scenarios where the sample size equals 2000 participants, the background event rate is 10% and there is a 100% increase in events in the intervention group at month 1, month 3, month 6 and month 11.

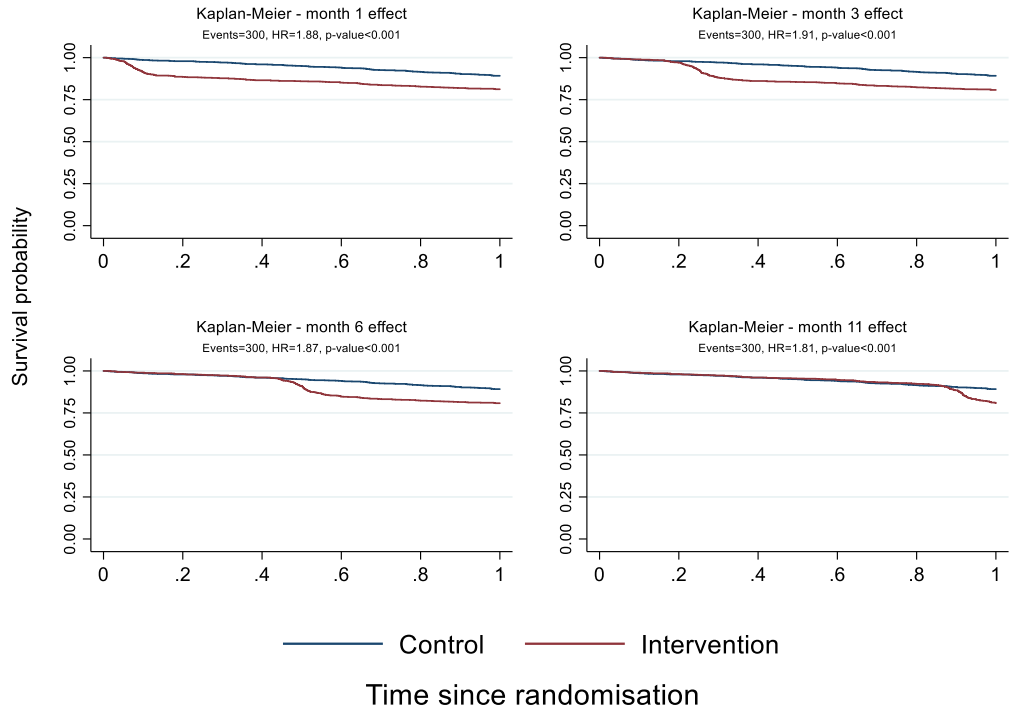


Figure 6.1: Kaplan-Meier plots at month 1, 3, 6 and 11 (relative to a 12 month trial) for the scenario where n=2000, AE background rate = 10% and the ADR increase= 100%

6.6.3 Simulation procedures - computational and coding considerations

Level of dependence between simulated datasets

As the aim of the simulations is to compare test performance across scenarios I allowed 'moderate' independence between simulated datasets.[290] This means for each scenario a set of datasets is simulated and to each dataset all methods are applied i.e. the same set of datasets are used to compare methods for the same scenario but a different set of datasets was generated for each scenario investigated. Comparing methods within a set of datasets for each scenario enables identification of any differences between methods.

Allowance for failures

Model parameters may fail to be estimated if, for example, algorithms fail to converge. Any such failures were discarded from the total number of simulations when calculating performance measures but the number of occurrences was retained and are reported. A high failure rate indicates that the method is likely to be impractical in similar real-world scenarios.

Software to perform simulations

Stata version 15.1 was used to generate all datasets.

Random number generator and specification of the starting seeds

The seed was set prior to simulating each set of datasets to ensure reproducibility of the random number list.

6.6.4 Criteria to evaluate the performance of the statistical methods across scenarios

Across scenarios the following performance measures were calculated for each method summarised in table 6.2:

i) Power (or sensitivity) was calculated as the proportion of simulations that detected a signal of the total number of simulated scenarios where a signal exists:

$$Power = \frac{1}{n_{sim\_sig}} \sum_{i=1}^{n_{sim\_sig}} (p_i \leq \alpha)$$

where n_{sim_sig} represents the number of total number of simulated scenarios where a signal exists and $(p_i \leq \alpha)$ represents the criteria to raise a signal for each test as outlined in table 6.3.

ii) The false positive rate was calculated as the proportion of simulations that detected a signal of the total number of simulated scenarios where no signal exists:

$$False\ postive = \frac{1}{n_{sim\_nosig}} \sum_{i=1}^{n_{sim\_nosig}} (p_i \leq \alpha)$$

where n_{sim_nosig} represents the number of total number of simulated scenarios where no signal exists and $(p_i \leq \alpha)$ represents the criteria to raise a signal for each test as outlined in table 6.3.

iii) Accuracy was calculated as the proportion of simulations that correctly detected (when an ADR was present) or did not detect a signal (when an ADR was not present) of all the simulated scenarios:

$$Accuracy = \frac{1}{n_{sim\_sig}} \sum_{i=1}^{n_{sim\_sig}} (p_i \leq \alpha) + \frac{1}{n_{sim\_nosig}} \sum_{i=1}^{n_{sim\_nosig}} (p_i > \alpha)$$

$$= Power + (1 - False\ Positives)$$

iv) A range for the required sample size was identified from the scenarios where power exceeded 80% i.e. the sample size when the proportion of significant results of the simulated scenarios exceeded 80%:

$$N = n\ such\ that\ \frac{1}{n_{sim\_sig}} \sum_{i=1}^{n_{sim\_sig}} (p_i \leq \alpha) \geq 0.80$$

where n_{sim_sig} represents the number of total number of simulated scenarios where a signal exists and $(p_i \leq \alpha)$ represents the criteria to raise a signal for each test as outlined in table 6.3.

6.6.5 Number of simulations

The number of simulations required was calculated to ensure that the performance measures outlined above would be estimated to sufficient precision. The number of simulations required to estimate precision for the power of tests is given by:

$$Standard\ Error = \sqrt{\frac{\widehat{power} * (1 - \widehat{power})}{n_{sim}}}$$

where n_{sim} represents the total number of simulated scenarios. For 90% power, 10000 simulations will give a standard error of 0.003, and 1000 simulations will give a standard error of 0.009. To minimise the running time 1000 simulations were used.

The number of simulations required to estimate the precision around significance is given by:

$$Standard\ Error = \sqrt{\frac{\hat{\alpha} * (1 - \hat{\alpha})}{n_{sim}}}$$

where n_{sim} represents the total number of simulated scenarios. With a significance level α = 0.05, 10000 simulations will give a standard error of 0.002, and 1000 simulations will give a standard error of 0.007. To minimise the running time 1000 simulations were used.

For the beta-binomial and gamma-Poisson models, 5000 random draws were taken to calculate the posterior probability for each simulated dataset.

6.6.6 Analysis

Results are presented through tabulations and graphical displays. Overall results for each performance measure across scenarios for prespecified sample size are presented, as well as by

background event rates, time of ADR and each ADR percentage increase. The number of missing estimates will be presented along with the performance measures as this has implications on the practicalities of implementation in similar real-world scenarios.[292]

The time description for the trial scenarios investigated are given relative to a 12 month study (i.e. 12 months between initial drug exposure and final follow-up) for ease of understanding but this is an arbitrary time frame and the results should be thought of as in fractions relative to the overall study length. For example, events simulated with a normal distribution such as ADR ~ Normal(0.25, 0.0416) are equivalent to events occurring at month 3 ± 2 weeks in a 12 month trial but month 6 ± 1 month in a 24 month trial.

Results and discussion are presented in chapter seven.

Table 6.5: Summary of simulated scenarios where signals truly exist (n = 270)

Total sample size	**Follow-up**	**Control**		**Intervention**				
Allocation: 1:1		**Model**	**Background rate**	**Model**	**Model parameters**	**% increase**	**ADR rate**	**Overall event rate**
					Mean (μ), Standard deviation (σ)			
WITH A SIGNAL								
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.1	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	25%	0.025	0.125
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.1	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	50%	0.05	0.15
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.1	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	100%	0.1	0.2
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.05	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	25%	0.0125	0.0625
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.05	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	50%	0.025	0.075
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.05	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	100%	0.05	0.1
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.01	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	25%	0.0025	0.0125
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.01	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	50%	0.005	0.015
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.01	Normal	μ= 0.0027, σ=0.0416 and μ= 0.083, 0.25, 0.5, 0.916, σ=0.0416	100%	0.01	0.02

Acronym: ADR – adverse drug reaction

Table 6.6: Summary of simulated scenarios without signals (n= 18)

Total sample size	**Follow-up**	**Control**		**Intervention**				
Allocation: 1:1		**Model**	**Background rate**	**Model**	**Parameters**	**% increase**	**ADR rate**	**Overall event rate**
					Mean (μ), Standard deviation (σ)			
WITHOUT A SIGNAL								
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.1	Uniform(0,1)	-	0%	0.0	0.1
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.05	Uniform(0,1)	-	0%	0.0	0.05
200, 400, 800, 1000, 2000, 5000	12 months	Uniform(0,1)	0.01	Uniform(0,1)	-	0%	0.0	0.01

Acronym: ADR – adverse drug reaction

7. Utilising time-to-event methodology to detect signals for ADRs: simulation results

7.1 Overall results across simulated trial scenarios

In practice, when screening emerging harm outcomes in clinical trials, the background event rates (i.e. the prevalence of events in the population not associated to the intervention), risk increase and times of increases will all be unknown and there will likely be a variety of these different characteristics. Therefore, performance measures of the signal detection tests have been aggregated across all simulated scenarios and presented by sample size to reflect a more realistic setting for the analysis of emerging events.

Table 7.1 and figure 7.1 display the mean power of each test across aggregated scenarios by sample size, where power is defined as the proportion of simulations that detected a signal of the total number of simulated scenarios where a signal exists. Table 7.2 and figure 7.2 show the mean rate of false positives for each test across aggregated scenarios by sample size, where the false positive rate is defined as the proportion of simulations that detected a signal of the total number of simulated scenarios where no signal exists. Given the almost identical results observed for the chi-squared and Fisher's exact tests and their inclusion only as a benchmark for comparison, the results for the chi-squared test are omitted for clarity in presentation. Appendix A7.1 table A7.1 and table A7.2 display the mean power and false positive rates across scenarios for comparative purposes of the chi-squared test and the Fisher's exact test.

Aggregating simulation results across scenarios (i.e. across all AE background rates, risk increases, and time of event) indicated that all of the methods under consideration lacked power to screen emerging events to detect time-dependent ADRs in trials of 5000 participants or smaller (n=2500 per

treatment group) (table 7.1 and figure 7.1) as power failed to exceed 80%. Therefore, none of the investigated methods can be recommended as a screening tool to detect signals of harm in trials smaller than 5000 participants. In samples of 5000 participants only the Bayesian tests, the beta-binomial and gamma-Poisson models achieved power of more than 80%, with mean values of 86% and 83%, respectively. These models also had negligible failure rates (1 out of 45000), where failures indicate that model parameters were not estimated due to, for example, non-convergence. The proportion of false positives for both the beta-binomial and gamma-Poisson ranged from 9% to 15% across sample sizes and again models very rarely failed to estimate the model parameters across these scenarios (table 7.2 and figure 7.2). The combined test had marginally lower power of 79% in samples of 5000 participants and the rate of false positives remained close to 5%.

The novel Weibull model with ancillary parameter and the double-Weibull model reached only 56% and 62% power with samples of 5000. Both approaches based on the Weibull distribution failed in just over 2% of scenarios where a signal truly existed. In addition, the Weibull model with ancillary parameter had the highest rates of false positives ranging from 15% to 18% across sample sizes. The double Weibull model had marginally smaller rates of false positives ranging from 12% to 15% across sample sizes, which could be due to the reduction in the p-value used to raise a signal from 0.05 to 0.025 to account for multiple tests. Both approaches had a high rate of failure in smaller sample sizes in scenarios where no signal existed (approximately 8% of tests failed to provide output with sample size of n=200).

The widely used approach of the Fisher's exact test had suboptimal power of 72% in sample sizes of 5000 but false positives remained around 5%. The novel double-Cox model approach showed similar results with power of 71% and 5% false positives but the simple Cox approach (i.e. conducting a Grambsch-Therneau test after fitting a Cox proportional hazards model) only achieved 60% power

whilst maintaining false positive rate of 5%. The double-Cox model also had an approximate 2% failure rate (995 out of 45000).

First recommendation*: the combined test can be used as a screening tool to analyse emerging harms in samples of 5000 or more with a minimum assumed background event rate of 1% but where there is no prior knowledge about background event rates in the study population (i.e. the prevalence of events in the population not associated to the intervention), the expected increase in the background rate due to ADRs or time ADRs are likely to occur.*

Table 7.1: Power of each test to detect a signal for an ADR by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
		N=45,000															
Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
200	Power	7,007	**0.16**	7,483	**0.18**	4,381	**0.10**	4,210	**0.10**	6,583	**0.17**	4,631	**0.10**	15,136	**0.34**	14,996	**0.33**
	Model fail	2,437		3,167		201		1,146		5,847		0		0		0	
400	Power	10,413	**0.23**	10,740	**0.25**	8,125	**0.18**	7,998	**0.18**	13,007	**0.29**	9,584	**0.21**	19,062	**0.42**	18,716	**0.42**
	Model fail	314		1,249		46		1,011		0		0		1		1	
800	Power	13,957	**0.31**	14,322	**0.33**	12,793	**0.28**	13,823	**0.31**	18,355	**0.41**	15,320	**0.34**	24,648	**0.55**	23,114	**0.51**
	Model fail	0		995		1		995		0		0		0		3	
1000	Power	14,964	**0.33**	15,471	**0.35**	14,257	**0.32**	15,746	**0.36**	20,214	**0.45**	17,179	**0.38**	25,954	**0.58**	25,823	**0.57**
	Model fail	0		995		0		995		0		0		0		3	
2000	Power	19,296	**0.43**	20,282	**0.46**	20,017	**0.44**	22,796	**0.52**	27,563	**0.61**	23,661	**0.53**	31,900	**0.71**	31,143	**0.69**
	Model fail	0		995		0		995		0		0		1		0	
5000	Power	25,186	**0.56**	27,334	**0.62**	26,897	**0.60**	31,259	**0.71**	35,357	**0.79**	32,487	**0.72**	38,560	**0.86**	37,389	**0.83**
	Model fail	0		995		0		995		0		0		1		1	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

Table 7.2: False positive rate for each test by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
		N=45,000															
Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
200	False positive	7,230	**0.18**	6,210	**0.15**	1,530	**0.03**	1,200	**0.03**	1,365	**0.03**	975	**0.02**	6,675	**0.15**	6,585	**0.15**
	Model fail	4,005		3,795		225		225		120		0		0		0	
400	False positive	7,500	**0.17**	6,360	**0.14**	1,740	**0.04**	1,395	**0.03**	3,165	**0.07**	1,605	**0.04**	3,900	**0.09**	3,945	**0.09**
	Model fail	960		915		30		15		0		0		0		0	
800	False positive	7,950	**0.18**	6,795	**0.15**	2,430	**0.05**	1,920	**0.04**	2,415	**0.05**	1,575	**0.04**	4,980	**0.11**	4,755	**0.11**
	Model fail	45		45		0		0		0		0		0		0	
1000	False positive	7,350	**0.16**	5,835	**0.13**	2,205	**0.05**	1,605	**0.04**	1,725	**0.04**	1,320	**0.03**	4,815	**0.11**	3,960	**0.09**
	Model fail	0		0		0		0		0		0		0		0	
2000	False positive	7,290	**0.16**	5,760	**0.13**	2,400	**0.05**	2,025	**0.05**	2,580	**0.06**	2,160	**0.05**	5,070	**0.11**	4,845	**0.11**
	Model fail	0		0		0		0		0		0		15		0	
5000	False positive	6,945	**0.15**	5,580	**0.12**	2,355	**0.05**	2,415	**0.05**	2,100	**0.05**	2,040	**0.05**	4,455	**0.10**	4,230	**0.09**
	Model fail	0		0		0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

Figure 7.1: Power of each test by sample size - summarised over varying AE background rates, time of increase and increases in background rate due to ADRs

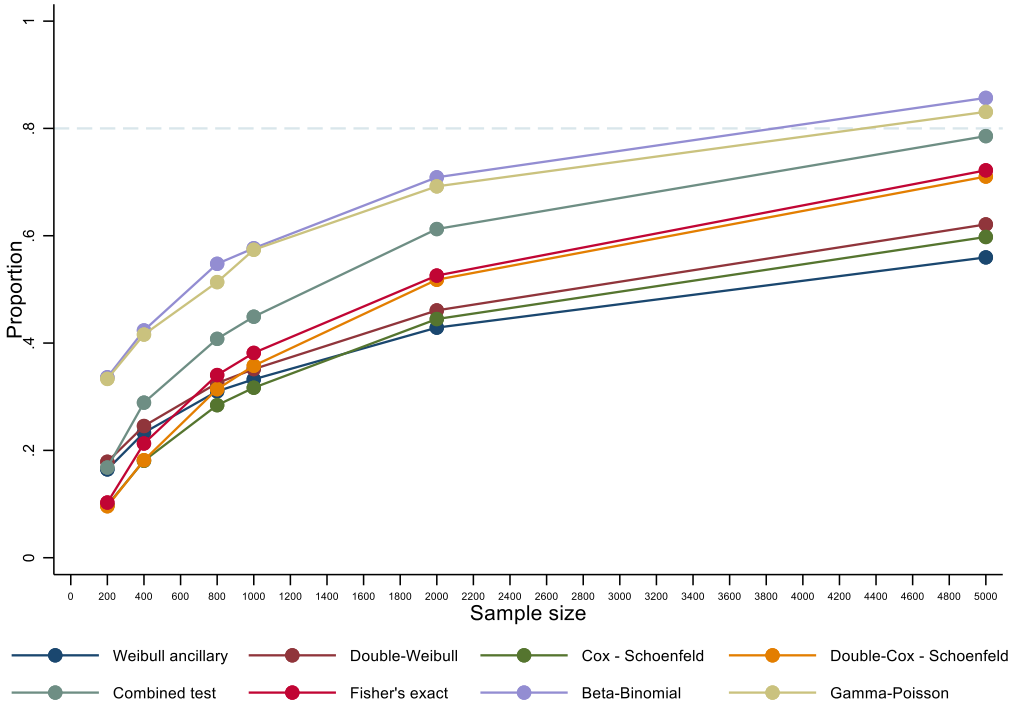
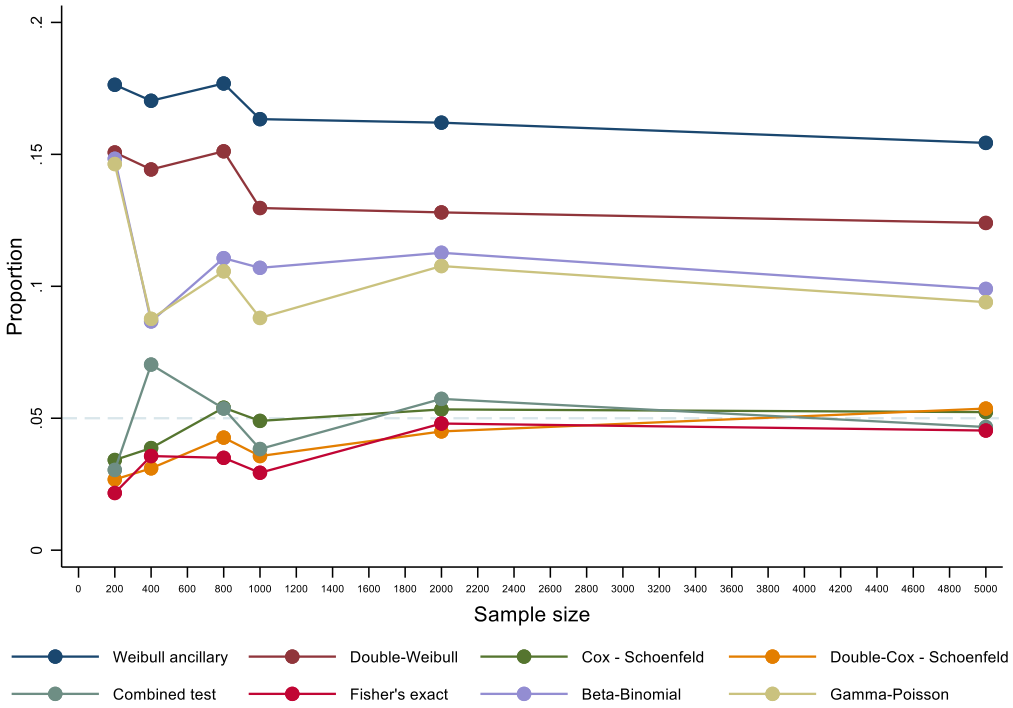


Figure 7.2: False positive rate for each test by sample size – summarised over varying AE background rates, time of increase and increases in background rate due to ADRs



7.2 Performance of the novel Weibull methods and the modified Cox proportional hazard model

The Weibull time-to-event model with ancillary parameter and the double-Weibull model described in section 6.3 (i) and (ii) were novel approaches I adapted from the observational setting to include a between arm comparison as potentially useful tools to detect signals for ADRs in RCTs.

The performance of both Weibull models across the range of simulated scenarios described in section 7.1 is clearly suboptimal in terms of both power and false positive rates. However, looking at specific simulated scenarios both models showed some potential value when the increase in events occurred immediately (day 1 ± 0.5 day) after exposure, with more than 80% power to detect a signal in many of the scenarios investigated (table 7.3). With an AE background rate of 10%, in samples of 800 participants or more the double-Weibull model could detect immediate increases of 25% with more than 80% power, and only required 400 and 200 participants to detect larger increases of 50% and 100% respectively, with false positives ranging from 11-15% across sample sizes (table 7.7). With an AE background rate of 5%, at least 2000 participants were required to detect immediate increases of 25% with more than 80% power, at least 800 participants were required to detect increases of 50% and at least 400 participants were required to detect increases of 100%, with false positives ranging between 12-15% across sample sizes. With AE background rates of 1%, the double-Weibull model failed to achieve sufficient power (≥ 80%) in the investigated sample sizes to detect immediate increases of 25%, and with sample of at least 5000 and 2000 participants the test had more than 80% power to detect 50% and 100% increases respectively, with false positive rates of 13% and 14% respectively. The power of the double-Weibull model struggled to rise above 80% in scenarios where increases occurred later in relation to exposure.

The Weibull model with ancillary parameter performed similarly to the double-Weibull model in detecting signals that occurred immediately (day 1 ± 0.5 day) after exposure but achieved at least 80% power in fewer of the later scenarios (table 7.3) and had slightly higher rates of false positives across background rates (table 7.7). The double-Weibull model was better able to detect later signals with fewer false positives in comparison to the simple Weibull but in only a limited number of scenarios and neither would be recommended for use as a screening tool across scenarios but could be used when interest is in detecting ADRs expected soon after starting treatment.

Comparing the investigated Weibull model (i.e. with treatment group included as both a treatment covariate and ancillary parameter, using the shape coefficient as the parameter to detect a signal) to various other Weibull approaches suggested that other approaches utilising the Weibull time-to-event model might outperform the proposed approach (appendix A7.2 table A7.3 and table A7.4). Alternatives included the Weibull model with treatment group included as a covariate only (i.e. no ancillary parameter), the Weibull model with treatment group only included as an ancillary parameter (i.e. no treatment group covariate), the same model investigated in this work but using the treatment group covariate parameter to raise a signal rather than the shape parameter. For example, using the treatment group covariate parameter to detect signals had power of 72% and false positive rate of 6% in samples of 5000 compared to 56% power and false positives of 15% using the shape parameter. However, these alternative approaches did not appear to outperform the other investigated tests such as the combined test and the beta-binomial and gamma-Poisson models.

I also explored a modification of the Cox proportional hazards model and the accompanying Grambsch-Therneau test to detect a deviation from the proportional hazards assumption. When increases occur immediately (day 1 ± 0.5 day) after exposure the double-Cox model had greater than

80% power to detect a signal in many of the scenarios investigated (table 7.3). With an AE background rate of 10%, samples of 2000 participants detected increases as low as 25%, and only required 400 and 200 participants respectively to detect 50% and 100% increases and false positive rates remained below 5% across scenarios (table 7.7). With an AE background rate of 5%, in samples of 5000 participants the double-Cox detected 25% increases, 2000 participants were required to detect 50% increases and 800 participants were required to detect 100% increases with more than 80% power, with false positives rates remaining below 6%. With AE background rates of 1%, none of the investigated sample sizes achieved sufficient power (≥ 80%) to detect increases of 25% or 50%, but with more than 5000 participants the test had more than 80% to detect 100% increases above the 1% background rate with 5% false positives.

The power of the double-Cox test struggled to rise above 80% in scenarios with later increases. Early (month 1 ± 2 weeks) increases of 100% on background rates of 10%, 5% and 1% could be detected with at least 80% power in samples of 800, 1000 and 5000 respectively, increases of 50% on AE background rates of 10% and 5% with at least 80% power in samples of 2000 and 5000, and increases of 25% on AE background rates of 10% with 80% power in samples of 5000. Performance at 3 and 6 months was similar, with sufficient power being reached in samples of 1000 or more to detect 100% increases on AE background rates of 10%, and 5000 or more on 50% increases on AE backgrounds of 10% or 100% increases on AE backgrounds of 5%. At month 11, the double-Cox was able to detect 100% increases on AE background rates of 10% and 5% with 800 and 2000 participants respectively; 50% increases on AE background rates of 10% and 5% with more 2000 and 5000 participants respectively; and 25% increases on an AE background rate of 10% in samples of more than 5000.

Sample sizes required by each of the other investigated tests to achieve 80% power and the specific power of each test are presented in tables A7.5 to A7.13 in appendix A7.3. None of the tests outperformed either of the investigated Weibull models when events occurred shortly (day 1 ± 0.5 day) following drug initiation. However, the combined test, and the beta-binomial and gamma-Poisson models outperformed both Weibull models and the double-Cox in many of the later scenarios. Test performance in specific scenarios will be described in detail in section 7.3.

Table 7.3: Sample sizes required for the proposed novel signal detection tests to achieve ≥ 80% power by simulated scenarios

Time		Day 1 ± 0.5 day			Month 1 ± 2 weeks			Month 3 ± 2 weeks			Month 6 ± 2 weeks			Month 11 ± 2 weeks		
		Increased AE rate (%)														
Model	AE background rate	25	50	100	25	50	100	25	50	100	25	50	100	25	50	100
		Sample size														
Weibull model with ancillary parameter	1%	-	5000	2000	-	-	-	-	-	-	-	-	-	-	-	-
	5%	2000	800	400	-	5000	2000	-	-	-	-	-	-	-	5000	2000
	10%	800	400	200	5000	2000	800	-	-	-	-	-	-	-	2000	800
Double- Weibull model with ancillary parameter	1%	-	5000	2000	-	-	-	-	-	-	-	-	-	-	-	-
	5%	2000	800	400	-	5000	2000	-	-	-	-	-	5000	-	5000	2000
	10%	800	400	200	5000	2000	1000	-	-	2000	-	5000	2000	-	5000	800
Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals	1%	-	-	5000	-	-	5000	-	-	-	-	-	-	-	-	-
	5%	5000	2000	800	-	5000	1000	-	-	5000	-	-	5000	-	5000	2000
	10%	2000	800	400	5000	2000	800	-	5000	2000	-	5000	1000	5000	2000	800

Note: Dash (-) indicates that 80% power not achieved across the sample sizes explored (n= 200, 400, 800, 1000, 2000, 5000)

7.3 Developing a signal detection strategy for screening emerging harm outcomes to detect ADRs based on simulation results

Examining the performance measures across tests by different trial characteristics revealed key insights into the potential utility of each of the tests as tools to detect signals for ADRs from the body of emerging harm outcomes and allowed the development of a signal detection strategy. The results that informed the signal detection strategy are summarised in detail in sections 7.3.1 to 7.3.3.

7.3.1 Signal detection strategy specifying the size of effect to detect (figure 7.3 & table 7.4)

Simulations looked at test performance in a range of scenarios, including varying increases in event rates in the intervention group compared to control. If a signal detection strategy for analysing emerging events specifies effect sizes to detect then more targeted strategies can be used to screen emerging events than the overall recommendation described in section 7.1. Performance measures for each of the tests according to increases of 100%, 50% and 25% are summarised below, the power of each test is displayed graphically in figure 7.3 and detailed results are presented in table 7.4. Guiding principles regarding a signal detection strategy utilising this information follow. The proportion of false positives do not change with varying increases in event rates and are presented for information only in table A7.14 in appendix A7.4.

AE background rate increase of 100%

The beta-binomial and gamma-Poisson models were able to detect a doubling (100% increase) in the event rate with nearly 80% power in sample sizes of 800 or more. The combined test detected a doubling with more than 80% power in samples of 2000 but only achieved 72% power to detect such differences in samples of 1000. Similarly, the Fisher's exact required a sample of at least 2000 to achieve 80% power to detect a doubling. Both the double-Weibull and double-Cox models required

samples of 5000 to detect a doubling in event rates with sufficient power. Neither the simple Weibull model with ancillary parameter nor the Cox proportional hazard model with the Grambsch-Therneau test achieved 80% power in samples of 5000.

AE background rate increase of 50%

None of the tests achieved 80% power to detect a 50% increase in event rates in samples smaller than 5000; and only the beta-binomial, gamma-Poisson and combined test achieved more than 80% power in samples of 5000. The beta-binomial and gamma-Poisson achieved 74% and 75% power to detect a 50% increase in samples of 2000 participants, which increased to 87% and 84% in samples of 5000; and the combined test only had 64% power in samples of 2000, which increased to 82% in samples of 5000.

AE background rate increase of 25%

None of the tests achieved sufficient power (≥ 80%) to detect a 25% increase in event rates in any of the simulated scenarios. Sample sizes would need to exceed 5000 participants for the tests to be powered sufficiently to detect such differences.

If a signal detection strategy for analysing emerging events specifies size of effects to detect then the following guiding principles can be used:

Second recommendation: *(a) In modest sample sizes (n≥800) the beta-binomial and gamma-Poisson models are recommended to detect at least a doubling in event rates where an excessive number of false signals is not of concern (i.e. false positive rate exceeds 5%). (b) The combined test or Fisher's exact test are recommended as alternatives in samples of 2000 or more to detect at least a doubling*

in event rates if an excessive number of false-positives is of concern (i.e. maintains false positive rate of 5%).

Third recommendation: *To detect 50% increases in event rates the beta-binomial or gamma-Poisson models are recommended in samples of 5000 or more but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate exceeds 5%).*

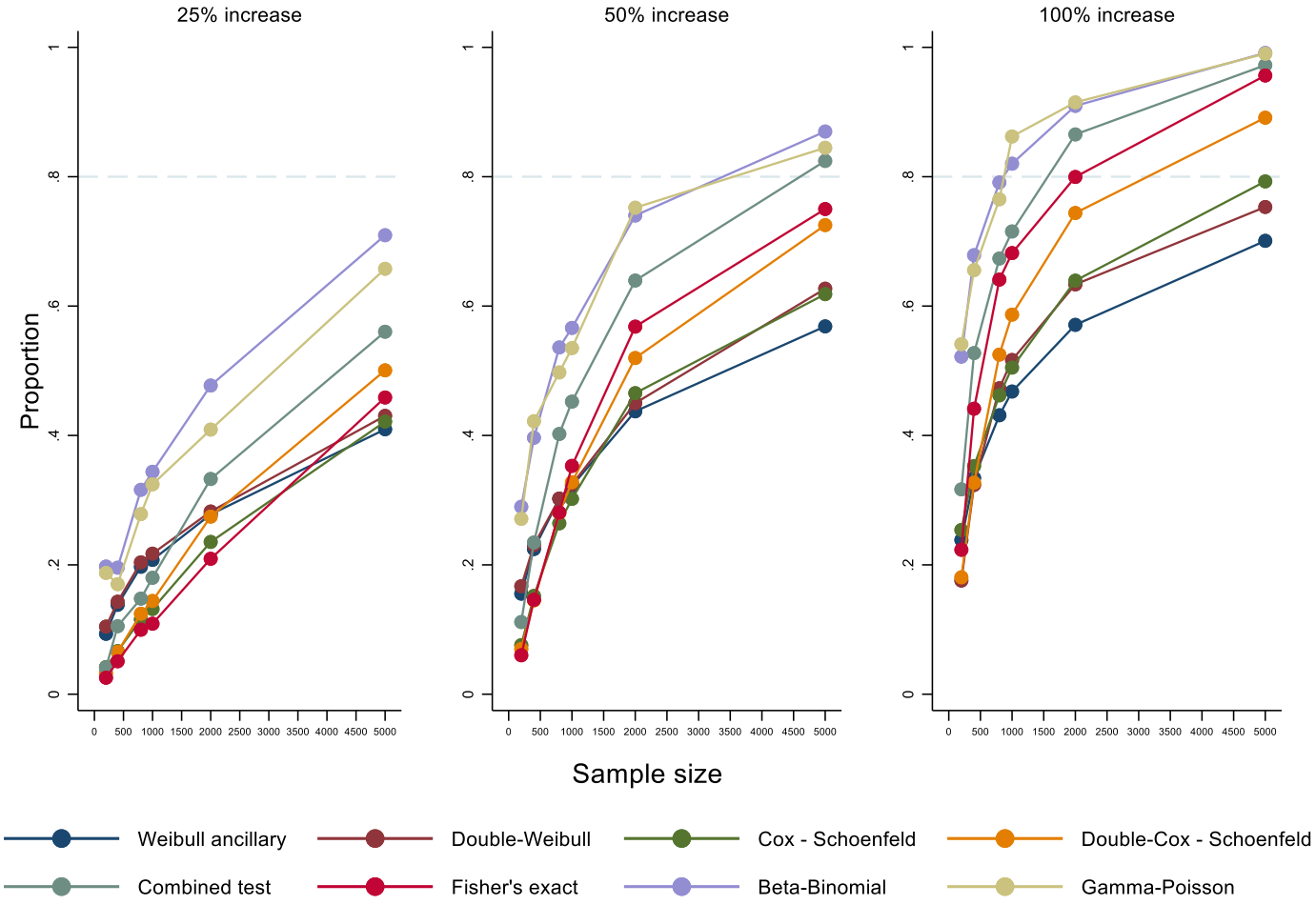
Table 7.4: Power of each test by sample size and increases in background event rates due to ADRs of 25%, 50% & 100% over background rates of 1%, 5% & 10% at day 1, month 1, 3, 6 & 11

			Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
			N=15,000															
Event increase	Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
25%	200	Power	1,286	**0.09**	1,371	**0.11**	628	**0.04**	463	**0.03**	495	**0.04**	379	**0.03**	2,961	**0.20**	2,813	**0.19**
		Model fail	1234		1967		50		1036		2900		0		0		0	
	400	Power	2,030	**0.14**	1,987	**0.14**	998	**0.07**	921	**0.07**	1,577	**0.11**	774	**0.05**	2,934	**0.20**	2,552	**0.17**
		Model fail	312		1247		18		1001		0		0		0		1	
	800	Power	2,956	**0.20**	2,856	**0.20**	1,737	**0.12**	1,742	**0.12**	2,217	**0.15**	1,494	**0.10**	4,740	**0.32**	4,177	**0.28**
		Model fail	0		995		0		995		0		0		0		1	
	1000	Power	3,115	**0.21**	3,034	**0.22**	1,977	**0.13**	2,028	**0.14**	2,700	**0.18**	1,655	**0.11**	5,158	**0.34**	4,867	**0.32**
		Model fail	0		995		0		995		0		0		0		0	
	2000	Power	4,171	**0.28**	3,942	**0.28**	3,532	**0.24**	3,841	**0.27**	4,991	**0.33**	3,143	**0.21**	7,158	**0.48**	6,137	**0.41**
		Model fail	0		995		0		995		0		0		0		0	
	5000	Power	6,143	**0.41**	6,039	**0.43**	6,322	**0.42**	7,012	**0.50**	8,405	**0.56**	6,889	**0.46**	10,643	**0.71**	9,864	**0.66**
		Model fail	0		995		0		995		0		0		0		1	
50%	200	Power	2,147	**0.16**	2,307	**0.17**	1,134	**0.08**	1,045	**0.07**	1,349	**0.11**	903	**0.06**	4,347	**0.29**	4,066	**0.27**
		Model fail	1189		1186		60		60		2910		0		0		0	
	400	Power	3,367	**0.22**	3,461	**0.23**	2,279	**0.15**	2,176	**0.17**	3,520	**0.23**	2,190	**0.15**	5,947	**0.40**	6,331	**0.42**
		Model fail	2		2		12		2		0		0		1		0	
	800	Power	4,532	**0.30**	4,536	**0.30**	3,962	**0.26**	4,206	**0.28**	6,035	**0.40**	4,212	**0.28**	8,041	**0.54**	7,462	**0.50**
		Model fail	0		0		0		0		0		0		0		0	
	1000	Power	4,830	**0.32**	4,863	**0.32**	4,529	**0.30**	4,916	**0.33**	6,784	**0.40**	5,295	**0.35**	8,494	**0.57**	8,026	**0.54**
		Model fail	0		0		0		0		0		0		0		0	
	2000	Power	6,560	**0.44**	6,750	**0.45**	6,982	**0.47**	7,796	**0.52**	9,592	**0.64**	8,524	**0.57**	11,099	**0.74**	11,281	**0.75**
		Model fail	0		0		0		0		0		0		1		0	
	5000	Power	8,530	**0.57**	9,404	**0.63**	9,277	**0.62**	10,878	**0.73**	12,366	**0.82**	11,251	**0.75**	13,047	**0.87**	12,672	**0.84**
		Model fail	0		0		0		0		0		0		1		0	
	200	Power	3,574	**0.24**	3,805	**0.25**	2,619	**0.18**	2,702	**0.18**	4,739	**0.32**	3,349	**0.22**	7,828	**0.52**	8,117	**0.54**

100%		Model fail	13		14		91		50		37		0		0		0	
	400	Power	5,016	**0.33**	5,292	**0.35**	4,848	**0.32**	4,901	**0.33**	7,910	**0.53**	6,620	**0.44**	10,181	**0.68**	9,833	**0.66**
		Model fail	0		0		16		8		0		0		0		0	
	800	Power	6,469	**0.43**	6,930	**0.46**	7,094	**0.47**	7,875	**0.53**	10,103	**0.67**	9,614	**0.64**	11,867	**0.79**	11,475	**0.77**
		Model fail	0		0		1		0		0		0		0		2	
	1000	Power	7,019	**0.47**	7,574	**0.50**	7,751	**0.52**	8,802	**0.59**	10,730	**0.72**	10,229	**0.68**	12,302	**0.82**	12,930	**0.86**
		Model fail	0		0		0		0		0		0		0		3	
	2000	Power	8,565	**0.57**	9,590	**0.64**	9,503	**0.63**	11,159	**0.74**	12,980	**0.87**	11,994	**0.80**	13,643	**0.91**	13,725	**0.92**
		Model fail	0		0		0		0		0		0		0		0	
	5000	Power	10,513	**0.70**	11,891	**0.79**	11,298	**0.75**	13,369	**0.89**	14,586	**0.97**	14,347	**0.96**	14,870	**0.99**	14,853	**0.99**
		Model fail	0		0		0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

Figure 7.3: Power of each test by sample size and percentage increase in background event rate due to ADRs - summarised over background event rates and time of increase



7.3.2 Signal detection strategy specifying a period of concern for detection (figure 7.4 & table 7.5)

Simulations also examined a range of scenarios with varying time of increases for ADRs. If there is a specific period of concern following drug exposure then a more targeted signal detection strategy can be employed to screen emerging events. Power for each of the tests according to a range of periods of concern are summarised below and recommendations regarding a signal detection strategy follow. The rates of false positives do not change with the time of increase so are only summarised in the first scenario for reference and are presented in table A7.15 in appendix A7.4.

Detecting immediate signals for ADRs (day 1 ± 0.5 day relative to a study with 12-month follow-up)

If interest is in detecting ADRs very early after treatment initiation (day 1 ± 0.5 day in reference to a trial with 12 months of follow-up), the Weibull model with ancillary parameter outperformed all other tests giving nearly 80% power for samples of 1000 or more, with near identical results shown for the double-Weibull model. Similar values for power were seen for the combined test with samples of 2000 or more. In small sample sizes (n=200) the combined test produced failure rates of 13% (n/N=1167/9000) and both Weibull models had failure rates of around 6%, but with power below 50% in this scenario these methods would not be recommended. Both the beta-binomial and gamma-Poisson needed at least 5000 participants in order to exceed 80% to detect immediate ADRs.

The rate of false positives for both Weibull models ranged from 12% to 16% in samples of at least 1000 participants, peaking at 18% for the smallest sample size of n=200. The rate of false positives for the combined test varied from 3% to 7% and for the Fisher's exact test rates remained below 5% across the range of sample sizes. The beta-binomial and gamma-Poisson models had failure rates of 10% and 9% respectively in samples of 5000 and remained around this point for smaller samples, peaking at 15% for a sample size of 200.

Detecting early signals for ADRs (1 month ± 2 weeks relative to a study with 12 month follow-up)

If interest is in detecting ADRs that occur early after treatment exposure (month 1 ± 2 weeks in reference to a trial with 12 months of follow-up) the combined test had the highest power (86%) of the examined tests followed by the beta-binomial and gamma-Poisson models (84% in both) in samples of 5000. The remaining models achieved power ranging from 70% to 78% in samples of 5000. None of the examined models achieved more than 80% power in samples smaller than 5000.

Detecting early to intermediate signals for ADRs (3 months ± 2 weeks and 6 months ± 2 weeks relative to a study with 12-month follow-up)

Similar results to those seen for early signals in samples of 5000 were seen when increases occurred at three and six-months post exposure (in reference to a trial with 12 months of follow-up) i.e. non-immediate reactions. In these scenarios, the beta-binomial (85% power at 3 months and 91% at 6 months) and gamma-Poisson (83% at both 3 months and 6 months) outperformed the combined test (77% at 3 months and 71% at 6 months). The Fisher's exact test also showed promise when aiming to detect ADRs occurring 6 months post treatment exposure, with 72% power in samples of 5000, outperforming all but the beta-binomial and gamma-Poisson models. Again, none of the examined models achieved more than 80% power in samples smaller than 5000.

Detecting late-onset signals for ADRs (11 months ± 2 weeks relative to a study with 12-month follow-up)

If the aim is to detect late onset ADRs (e.g. month 11 ± 2 weeks in reference to a trial with 12 months of follow-up) then the beta-binomial and gamma-Poisson models continued to perform well in samples of 5000, with power of 84% and 83%, respectively. However, the power of the combined test fell below 70% at this point and the double-Cox reached 80% for the first time in samples of

5000. Again, none of the examined models achieved more than 80% power in samples smaller than 5000.

In all of these scenarios failure rates are negligible across models in samples of 5000 participants.

If there is a specific period of concern following drug exposure then the following guiding principles can be used to develop an analysis strategy:

Fourth recommendation*: (a) if aiming to detect ADRs that occur immediately post treatment exposure the Weibull model with ancillary parameter can be used in samples of 1000 or more where an excessive number of false signals is not of concern (i.e. false positive rate exceeds 5%). (b) The combined test can be used in samples of 2000 or more when excessive numbers of false-positives is of concern (i.e. maintains false positive rate of 5%).*

Fifth recommendation*: to detect increases in events occurring early after exposure (one-month ± 2 weeks) the combined test is recommended to screen events on sample sizes of 5000 or more.*

Sixth recommendation*: if interest is to detect longer-term events (month 1- 11) the beta-binomial and gamma-Poisson tests are recommended in samples of 5000 or more but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate exceeds 5%).*

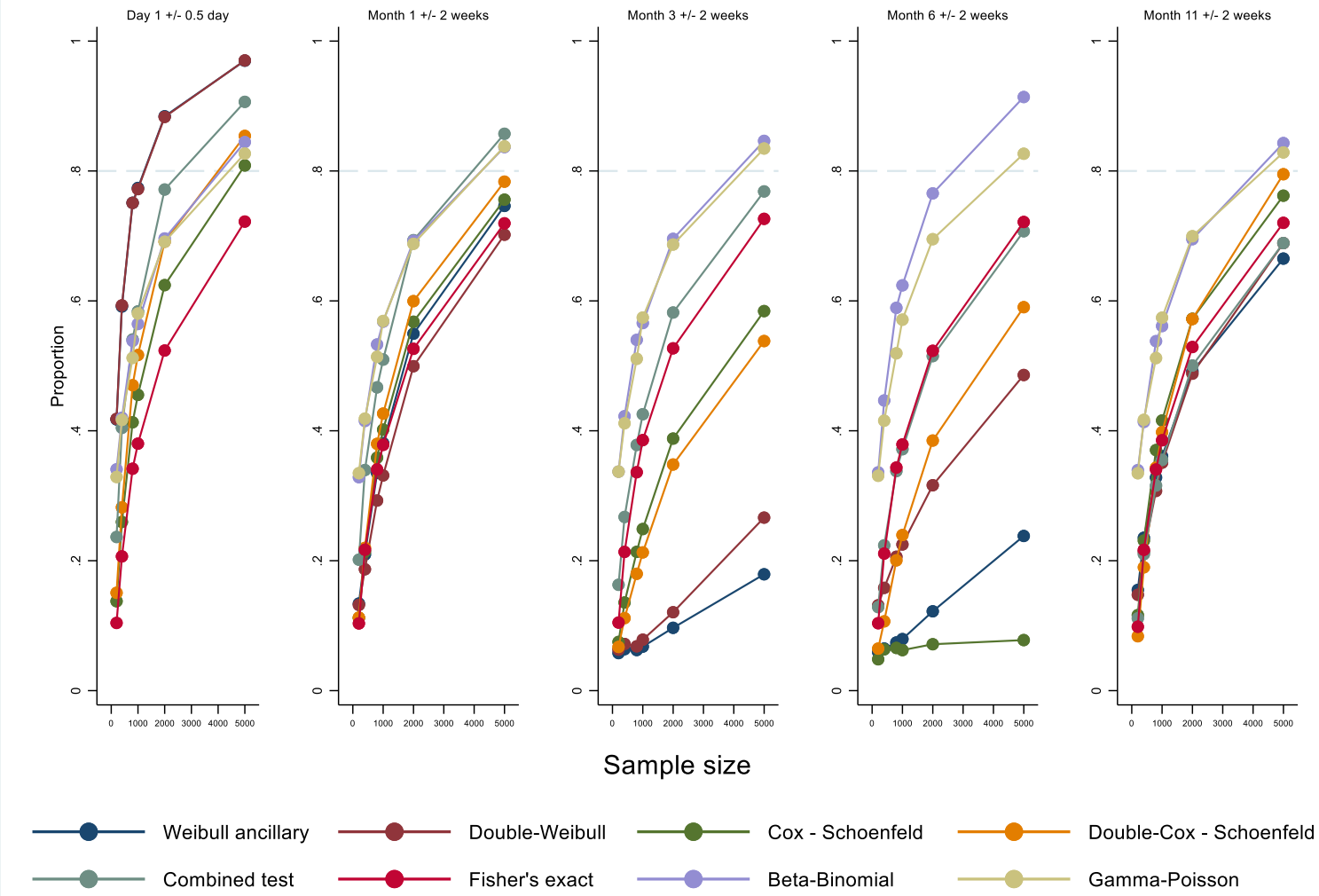
Table 7.5: Power of each test by sample size and time of increase over: background rates of 1%, 5% & 10% & increases in background rates due to ADRs of 25%, 50% & 100%

Time	Sample size		Weibull model with ancillary parameter		Double-Weibull model with ancillary parameter		Cox PH model with GT test for disproportionality using Schoenfeld residuals		Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals		Combined test		Fisher's exact test		Bayesian beta-binomial model		Bayesian gamma-Poisson model	
			n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
			N=9000															
Day 1	200	Power	3,547	**0.42**	3,555	**0.42**	1,234	**0.14**	1,351	**0.15**	1,853	**0.24**	938	**0.10**	3,066	**0.34**	2,960	**0.33**
		Model fail	505		502		35		26		1,167		0		0		0	
	400	Power	5,288	**0.59**	5,304	**0.59**	2,334	**0.26**	2,539	**0.28**	3,641	**0.40**	1,860	**0.21**	3,781	**0.42**	3,751	**0.42**
		Model fail	58		58		11		2		0		0		0		0	
	800	Power	6,759	**0.75**	6,758	**0.75**	3,718	**0.41**	4,232	**0.47**	4,865	**0.54**	3,075	**0.34**	4,846	**0.54**	4,610	**0.51**
		Model fail	0		0		0		0		0		0		0		0	
	1000	Power	6,965	**0.77**	6,949	**0.77**	4,098	**0.46**	4,650	**0.52**	5,255	**0.58**	3,422	**0.38**	5,080	**0.56**	5,225	**0.58**
		Model fail	0		0		0		0		0		0		0		0	
	2000	Power	7,958	**0.88**	7,949	**0.88**	5,619	**0.62**	6,221	**0.69**	6,943	**0.77**	4,712	**0.52**	6,265	**0.70**	6,220	**0.69**
		Model fail	0		0		0		0		0		0		0		0	
	5000	Power	8,729	**0.97**	8,732	**0.97**	7,277	**0.81**	7,686	**0.85**	8,158	**0.91**	6,499	**0.72**	7,600	**0.84**	7,441	**0.83**
		Model fail	0		0		0		0		0		0		1		0	
Month 1	200	Power	1,145	**0.13**	1,127	**0.13**	1,002	**0.11**	1,008	**0.11**	1,571	**0.20**	931	**0.10**	2,957	**0.33**	3,013	**0.33**
		Model fail	456		454		46		36		1,206		0		0		0	
	400	Power	1,871	**0.21**	1,668	**0.19**	1,924	**0.21**	1,977	**0.22**	3,056	**0.34**	1,952	**0.22**	3,736	**0.42**	3,769	**0.42**
		Model fail	71		71		5		2		0		0		0		0	
	800	Power	3,016	**0.34**	2,634	**0.29**	3,228	**0.36**	3,419	**0.38**	4,202	**0.47**	3,065	**0.34**	4,796	**0.53**	4,626	**0.51**
		Model fail	0		0		0		0		0		0		0		0	
	1000	Power	3,428	**0.38**	2,982	**0.33**	3,616	**0.40**	3,839	**0.43**	4,587	**0.51**	3,406	**0.38**	5,113	**0.57**	5,122	**0.57**
		Model fail	0		0		0		0		0		0		0		0	
	2000	Power	4,946	**0.55**	4,495	**0.50**	5,111	**0.57**	5,397	**0.60**	6,241	**0.69**	4,736	**0.53**	6,231	**0.69**	6,191	**0.69**
		Model fail	0		0		0		0		0		0		0		0	
	5000	Power	6,716	**0.75**	6,317	**0.70**	6,803	**0.76**	7,052	**0.78**	7,716	**0.86**	6,475	**0.72**	7,531	**0.84**	7,541	**0.84**
		Model fail	0		0		0		0		0		0		0		0	

Month 3	200	Power	492	**0.06**	538	**0.06**	672	**0.07**	603	**0.07**	1,278	**0.16**	943	**0.10**	3,034	**0.34**	3,034	**0.34**
		Model fail	484		480		35		28		1,158	**0**	0		0		0	
	400	Power	571	**0.06**	642	**0.07**	1,220	**0.14**	1,004	**0.11**	2,406	**0.27**	1,922	**0.21**	3,802	**0.42**	3,704	**0.41**
		Model fail	56		56		10		2		0		0		1		1	
	800	Power	562	**0.06**	614	**0.07**	1,923	**0.21**	1,619	**0.18**	3,404	**0.38**	3,026	**0.34**	4,860	**0.54**	4,596	**0.51**
		Model fail	0		0		1		0		0		0		0		3	
	1000	Power	610	**0.07**	706	**0.08**	2,239	**0.25**	1,915	**0.21**	3,830	**0.43**	3,471	**0.39**	5,093	**0.57**	5,168	**0.57**
		Model fail	0		0		0		0		0		0		0		3	
	2000	Power	870	**0.10**	1,086	**0.12**	3,492	**0.39**	3,133	**0.35**	5,239	**0.58**	4,743	**0.53**	6,260	**0.70**	6,181	**0.69**
		Model fail	0		0		0		0		0		0		0		0	
	5000	Power	1,612	**0.18**	2,397	**0.27**	5,259	**0.58**	4,844	**0.54**	6,917	**0.77**	6,537	**0.73**	7,617	**0.85**	7,511	**0.83**
		Model fail	0		0		0		0		0		0		0		1	
Month 6	200	Power	503	**0.06**	1,116	**0.13**	432	**0.05**	580	**0.06**	1,006	**0.13**	934	**0.10**	3,024	**0.34**	2,977	**0.33**
		Model fail	507		506		40		34		1,177		0		0		0	
	400	Power	579	**0.06**	1,413	**0.16**	571	**0.06**	959	**0.11**	2,013	**0.22**	1,899	**0.21**	4,020	**0.45**	3,740	**0.42**
		Model fail	68		68		10		7		0		0		0		0	
	800	Power	667	**0.07**	1,855	**0.21**	591	**0.07**	1,805	**0.20**	3,042	**0.34**	3,091	**0.34**	5,304	**0.59**	4,674	**0.52**
		Model fail	0		0		0		0		0		0		0		0	
	1000	Power	714	**0.08**	2,025	**0.23**	561	**0.06**	2,156	**0.24**	3,345	**0.37**	3,411	**0.38**	5,616	**0.62**	5,140	**0.57**
		Model fail	0		0		0		0		0		0		0		0	
	2000	Power	1,099	**0.12**	2,846	**0.32**	643	**0.07**	3,465	**0.39**	4,635	**0.52**	4,709	**0.52**	6,889	**0.77**	6,255	**0.70**
		Model fail	0		0		0		0		0		0		1		0	
	5000	Power	2,143	**0.24**	4,373	**0.49**	700	**0.08**	5,313	**0.59**	6,365	**0.71**	6,494	**0.72**	8,225	**0.91**	7,439	**0.83**
		Model fail	0		0		0		0		0		0		0		0	
Month 11	200	Power	1,320	**0.16**	1,147	**0.15**	1,041	**0.12**	668	**0.08**	875	**0.11**	885	**0.10**	3,055	**0.34**	3,012	**0.33**
		Model fail	485		1,225		45		1,022		1,139		0		0		0	
	400	Power	2,104	**0.24**	1,713	**0.21**	2,076	**0.23**	1,519	**0.19**	1,891	**0.21**	1,951	**0.22**	3,723	**0.41**	3,752	**0.42**
		Model fail	61		996		10		998		0		0		0		0	
	800	Power	2,953	**0.33**	2,461	**0.31**	3,333	**0.37**	2,748	**0.34**	2,842	**0.32**	3,063	**0.34**	4,842	**0.54**	4,608	**0.51**
		Model fail	0		995		0		995		0		0		0		0	
	1000	Power	3,247	**0.36**	2,809	**0.35**	3,743	**0.42**	3,186	**0.40**	3,197	**0.36**	3,469	**0.39**	5,052	**0.56**	5,168	**0.57**
		Model fail	0		995		0		995		0		0		0		0	
	2000	Power	4,423	**0.49**	3,906	**0.49**	5,152	**0.57**	4,580	**0.57**	4,505	**0.50**	4,761	**0.53**	6,255	**0.70**	6,296	**0.70**
		Model fail	0		995		0		995		0		0		0		0	
	5000	Power	5,986	**0.67**	5,515	**0.69**	6,858	**0.76**	6,364	**0.80**	6,201	**0.69**	6,482	**0.72**	7,587	**0.84**	7,457	**0.83**
		Model fail	0		995		0		995		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power.
Model fail indicate either failure to estimate parameters or non-convergence. Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

Figure 7.4: Power of each test by sample size and time of increase for ADR - summarised over varying AE background rates and increases in background rate due to ADRs



7.3.3 Signal detection strategy utilising prior knowledge on background event rates (figure 7.5 & tables 7.6 and 7.7)

Simulations were also used to explore the impact of varying event background rates (i.e. the prevalence of the event in the population not associated to the intervention). If prior information on event background rates is available then this information can be exploited to choose the most appropriate test to screen emerging events. Performance measures of each of the tests according to background rates of 10%, 5% and 1% are summarised below and guiding principles regarding a signal detection strategy utilising background event rates follow.

AE background rate – 10%

With background event rates of 10% the combined test and the Fisher's exact test achieved 86% and 79% power respectively in trials of at least 2000 participants and exceeded 90% power in samples of 5000 (96% and 93% respectively). The beta-binomial and gamma-Poisson models both achieved 78% power with sample sizes of at least 1000, 89% and 87% in samples of 2000 and over 90% in samples of 5000 (98% and 97% respectively). With samples of 5000 participants, the double-Weibull and double-Cox models achieved 81% and 91% power, respectively.

Across investigated sample sizes the proportion of false positives ranged from 3% to 5% for the Fisher's exact test, 3% to 6% for the combined test, and 4% to 6% for the Cox proportional hazards model with Grambsch-Therneau test and the double-Cox model. The beta-binomial and gamma-Poisson tests had false-positives rates ranging from 9% to 11% and reached 15% for the gamma-Poisson in small sample sizes (n=200), with similar results observed for the double-Weibull model (range 11% to 15%), but the simple Weibull time-to-event model with ancillary parameter produced three times as many false-positives (range 14% to 18%).

AE background rate – 5%

With background event rates of 5% the combined test, the Fisher's exact test and the double-Cox method reached 88%, 81% and 80% power respectively and the beta-binomial and gamma-Poisson models achieved 91% and 90% power respectively in samples of 5000 participants. With 2000 participants, the beta-binomial and gamma-Poisson models both achieved 77% power. With background rates of 5%, there were very few failures, occurring only in the beta-binomial and gamma-Poisson models (highest failure rate of 3 out of 15000 = 0.02%).

Across investigated sample sizes the pattern of false positives was similar to that seen for background rates of 10%, with the proportion of false positives for the Fisher's exact test ranging from 3% to 4%, false positives for the combined test ranged from 4% to 6% and both Cox models had false positive rates between 3% and 5% across scenarios. The beta-binomial and gamma-Poisson models rates ranged from 5% to 12% and both Weibull models continued to produce high rates of false positives (range 12% to 18%).

AE background rate – 1%

When background event rates were low (1%) all of the examined methods were inappropriate as a means to detect signals for ADRs, with power remaining below 70% across all tests in investigated sample sizes, and failure rates as high as 39% in the smallest sample sizes.

Similar patterns as for the larger background rates are seen for the number of false positives across sample sizes for the Fisher's exact test, the combined test and both Cox models, with the number remaining below 6%, excluding an inflated rate for the combined test when n=400 which produced 12% false positives. Rates increased slightly in smaller sample sizes for the beta-binomial and

gamma-Poisson models with 24% false positives when n=200, but similar patterns were seen for larger sample sizes as in larger background rates. Both Weibull models had larger false positive rates across the range of sample sizes ranging from 13% to 21%.

If information on background event rates in the study population is known from, for example, control group event rates observed in previous trials, the following guiding principles can be used to inform an analysis strategy:

Seventh recommendation*: if the background event rate is known to be approximately 10% or greater, the combined test is recommended to screen events to detect ADRs on sample sizes of 2000 or more.*

Eighth recommendation*: With background event rates of 5% or greater the beta-binomial and gamma-Poisson models are recommended to screen events to detect ADRs in samples of at least 2000 participants but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate exceeds 5%).*

Table 7.6: Power of each test by sample size & AE background rates over: increases in background rates due to ADRs of 25%, 50% & 100%, at day 1, month 1, 3, 6 & 11

			Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
AE %	Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
			N=15,000															
10%	200	Power	3,487	**0.23**	3,506	**0.23**	2,653	**0.18**	2,753	**0.18**	4,570	**0.30**	3,328	**0.22**	6,946	**0.46**	6,401	**0.43**
		Model fail	0		0		0		0		0		0		0		0	
	400	Power	4,996	**0.33**	4,999	**0.33**	4,453	**0.30**	4,925	**0.33**	7,481	**0.50**	6,154	**0.41**	9,263	**0.62**	8,897	**0.59**
		Model fail	0		0		0		0		0		0		1		1	
	800	Power	6,673	**0.44**	6,818	**0.45**	6,741	**0.45**	7,680	**0.51**	9,917	**0.66**	8,509	**0.57**	11,353	**0.76**	10,715	**0.71**
		Model fail	0		0		0		0		0		0		0		3	
	1000	Power	7,104	**0.47**	7,446	**0.50**	7,457	**0.50**	8,588	**0.57**	10,607	**0.71**	9,311	**0.62**	11,743	**0.78**	11,706	**0.78**
		Model fail	0		0		0		0		0		0		0		3	
	2000	Power	8,596	**0.57**	9,372	**0.62**	9,621	**0.64**	11,260	**0.75**	12,842	**0.86**	11,796	**0.79**	13,373	**0.89**	13,073	**0.87**
		Model fail	0		0		0		0		0		0		0		0	
	5000	Power	10,717	**0.71**	12,079	**0.81**	11,473	**0.76**	13,721	**0.91**	14,328	**0.96**	13,976	**0.93**	14,697	**0.98**	14,570	**0.97**
		Model fail	0		0		0		0		0		0		1		1	
5%	200	Power	2,403	**0.16**	2,783	**0.19**	1,728	**0.12**	1,457	**0.10**	1,822	**0.12**	1,303	**0.09**	4,521	**0.30**	4,808	**0.32**
		Model fail	0		0		0		0		0		0		0		0	
	400	Power	3,533	**0.24**	3,534	**0.24**	2,747	**0.18**	2,907	**0.19**	4,268	**0.28**	2,929	**0.20**	6,470	**0.43**	6,379	**0.43**
		Model fail	0		0		0		0		0		0		0		0	
	800	Power	4,950	**0.33**	4,942	**0.33**	4,534	**0.30**	4,959	**0.33**	7,220	**0.48**	5,732	**0.38**	8,927	**0.60**	9,088	**0.61**
		Model fail	0		0		0		0		0		0		0		0	
	1000	Power	5,512	**0.37**	5,507	**0.37**	5,170	**0.34**	5,740	**0.38**	7,967	**0.53**	6,582	**0.44**	9,485	**0.63**	9,318	**0.62**
		Model fail	0		0		0		0		0		0		0		0	
	2000	Power	7,116	**0.47**	7,362	**0.49**	7,574	**0.50**	8,591	**0.57**	10,585	**0.71**	9,034	**0.60**	11,575	**0.77**	11,606	**0.77**
		Model fail	0		0		0		0		0		0		1		0	
	5000	Power	9,033	**0.60**	9,925	**0.66**	10,195	**0.68**	12,001	**0.80**	13,227	**0.88**	12,222	**0.81**	13,681	**0.91**	13,534	**0.90**
		Model fail	0		0		0		0		0		0		0		0	

1%	200	Power	1,117	**0.09**	1,194	**0.10**	0	**0.00**	0	**0.00**	191	**0.02**	0	**0.00**	3,669	**0.24**	3,787	**0.25**
		Model fail	2,437		3,167		201		1,146		5,847		0		0		0	
	400	Power	1,884	**0.13**	2,207	**0.16**	925	**0.06**	166	**0.01**	1,258	**0.08**	501	**0.03**	3,329	**0.22**	3,440	**0.23**
		Model fail	314		1,249		46		1,011		0		0		0		0	
	800	Power	2,334	**0.16**	2,562	**0.18**	1,518	**0.10**	1,184	**0.08**	1,218	**0.08**	1,079	**0.07**	4,368	**0.29**	3,311	**0.22**
		Model fail	0		995		1		995		0		0		0		0	
	1000	Power	2,348	**0.16**	2,518	**0.18**	1,630	**0.11**	1,418	**0.10**	1,640	**0.11**	1,286	**0.09**	4,726	**0.32**	4,799	**0.32**
		Model fail	0		995		0		995		0		0		0		0	
	2000	Power	3,584	**0.24**	3,548	**0.25**	2,822	**0.19**	2,945	**0.21**	4,136	**0.28**	2,831	**0.19**	6,952	**0.46**	6,464	**0.43**
		Model fail	0		995		0		995		0		0		0		0	
	5000	Power	5,436	**0.36**	5,330	**0.38**	5,229	**0.35**	5,537	**0.40**	7,802	**0.52**	6,289	**0.42**	10,182	**0.68**	9,285	**0.62**
		Model fail	0		995		0		995		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

Figure 7.5: Power of each test by sample size and AE background rates - summarised over varying times of increase and increased background rates due to ADRs

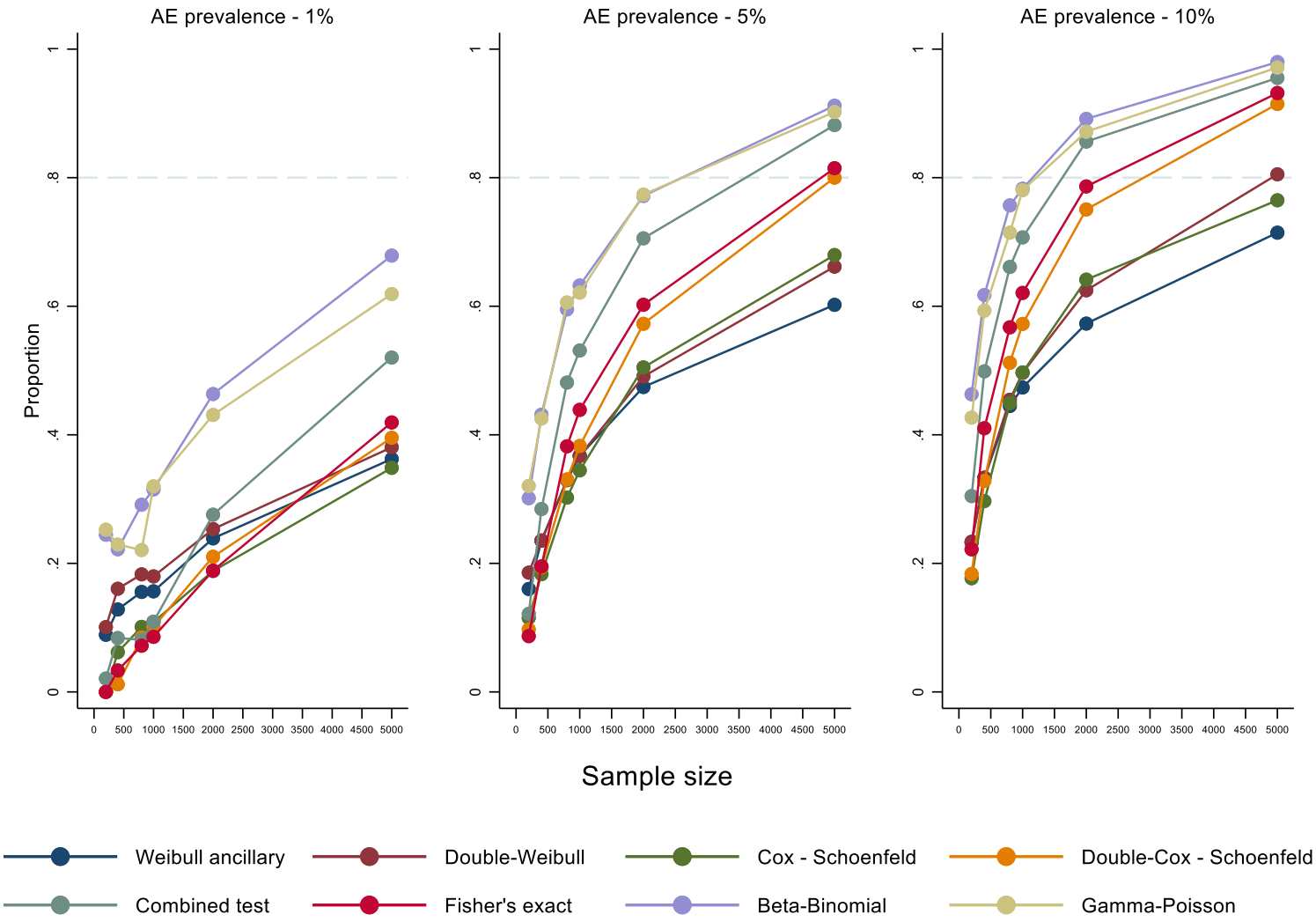


Table 7.7: False positives of each test by sample size & AE background rates over: increases in background rates due to ADRs of 25%, 50% & 100%, at day 1, month 1, 3, 6 & 11

			Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Double-Cox PH model with GT test for disproportionality using Schoenfeld residuals**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
			N=15,000															
AE background rate	Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
10%	200	False positive	2,655	**0.18**	2,205	**0.15**	840	**0.06**	750	**0.05**	720	**0.05**	435	**0.03**	1,710	**0.11**	2,220	**0.15**
		Model fail	0		0		0		0		0		0		0		0	
	400	False positive	2,055	**0.14**	1,620	**0.11**	645	**0.04**	645	**0.04**	750	**0.05**	795	**0.05**	1,530	**0.10**	1,260	**0.08**
		Model fail	0		0		0		0		0		0		0		0	
	800	False positive	2,505	**0.17**	1,950	**0.13**	795	**0.05**	840	**0.06**	840	**0.06**	660	**0.04**	1,635	**0.11**	1,680	**0.11**
		Model fail	0		0		0		0		0		0		0		0	
	1000	False positive	2,310	**0.15**	1,695	**0.11**	690	**0.05**	525	**0.04**	495	**0.03**	435	**0.03**	1,665	**0.11**	1,515	**0.10**
		Model fail	0		0		0		0		0		0		0		0	
	2000	False positive	2,340	**0.16**	1,680	**0.11**	780	**0.05**	585	**0.04**	720	**0.05**	690	**0.05**	1,650	**0.11**	1,380	**0.09**
		Model fail	0		0		0		0		0		0		15		0	
	5000	False positive	2,205	**0.15**	1,680	**0.11**	765	**0.05**	750	**0.05**	705	**0.05**	600	**0.04**	1,410	**0.09**	1,500	**0.10**
		Model fail	0		0		0		0		0		0		0		0	
5%	200	False positive	2,640	**0.18**	2,265	**0.15**	660	**0.04**	450	**0.03**	600	**0.04**	540	**0.04**	1,365	**0.09**	780	**0.05**
		Model fail	60		60		0		0		0		0		0		0	
	400	False positive	2,535	**0.17**	2,160	**0.14**	720	**0.05**	735	**0.05**	690	**0.05**	540	**0.04**	1,395	**0.09**	1,695	**0.11**
		Model fail	0		0		0		0		0		0		0		0	
	800	False positive	2,625	**0.18**	2,325	**0.15**	750	**0.05**	765	**0.05**	825	**0.06**	495	**0.03**	1,305	**0.09**	1,095	**0.07**
		Model fail	0		0		0		0		0		0		0		0	
	1000	False positive	2,445	**0.16**	1,830	**0.12**	750	**0.05**	720	**0.05**	810	**0.05**	630	**0.04**	1,740	**0.12**	1,530	**0.10**
		Model fail	0		0		0		0		0		0		0		0	
	2000	False positive	2,430	**0.16**	1,980	**0.13**	720	**0.05**	690	**0.05**	900	**0.06**	630	**0.04**	1,755	**0.12**	1,365	**0.09**
		Model fail	0		0		0		0		0		0		0		0	

	5000	False positive	2,460	**0.16**	2,010	**0.13**	825	**0.06**	885	**0.06**	645	**0.04**	660	**0.04**	1,485	**0.10**	1,275	**0.09**
		Model fail	0		0		0		0		0		0		0		0	
1%	200	False positive	1,935	**0.18**	1,740	**0.15**	30	**0.002**	0	**0.00**	45	**0.003**	0	**0.00**	3,600	**0.24**	3,585	**0.24**
		Model fail	3945		3735		225		245		120		0		0		0	
	400	False positive	2,910	**0.21**	2,580	**0.18**	375	**0.03**	15	**0.001**	1,725	**0.12**	270	**0.02**	975	**0.07**	990	**0.07**
		Model fail	960		915		30		15		0		0		0		0	
	800	False positive	2,820	**0.19**	2,520	**0.17**	885	**0.06**	315	**0.02**	750	**0.05**	420	**0.03**	2,040	**0.14**	1,980	**0.13**
		Model fail	45		45		0		0		0		0		0		0	
	1000	False positive	2,595	**0.17**	2,310	**0.15**	765	**0.05**	360	**0.02**	420	**0.03**	255	**0.02**	1,410	**0.09**	915	**0.06**
		Model fail	0		0		0		0		0		0		0		0	
	2000	False positive	2,520	**0.17**	2,100	**0.14**	900	**0.06**	750	**0.05**	960	**0.06**	840	**0.06**	1,665	**0.11**	2,100	**0.14**
		Model fail	0		0		0		0		0		0		0		0	
	5000	False positive	2,280	**0.15**	1,890	**0.13**	765	**0.05**	780	**0.05**	750	**0.05**	780	**0.05**	1,560	**0.10**	1,455	**0.10**
		Model fail	0		0		0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: ADR – adverse drug reaction; PH – proportional hazards; GT - Grambsch-Therneau

7.4 Summary and recommendations

The aim of the work described in this chapter and chapter 6 was to identify if there were reliable and objective methods to identify time dependent ADRs from the body of emerging harm outcomes. A novel method building on work from the observation setting was developed using the Weibull time-to-event model with ancillary parameter and this was compared to range of alternative approaches. The Weibull models incorporated time-to-event information into the analysis with the aim of detecting a disproportionality in hazard rates between treatment groups, indicating a time-dependent reaction. This was compared to a widely used simple approach that does not utilise information on the time events occur (the Fisher's exact test), a standard time-to-event method to detect a violation of the proportional hazards assumption (the Cox proportional hazards model with the Grambsch-Therneau test for disproportionality) and a simple modification of this method (the double-Cox model), a recently developed time-to-event method designed to account for the presence of a disproportionality in hazard rates (combined test), and two easily implementable Bayesian approaches (beta-binomial and gamma-Poisson models).

Simulations were used to create a range of typical simple RCT scenarios. Performance of each of the tests was investigated by looking at how often each test detected a signal where a signal truly existed (power) and how often each test detected a signal where there was no signal (false positives) across the simulated scenarios. Performance measures were compared across tests and based on a balance of both power and the rate of false positives in the simulated scenarios, guiding principles as to where each of the evaluated tests may be appropriately used were devised and are summarised in table 7.8. Figure 7.6 was also developed to help identify which of the evaluated methods provides the most appropriate signal detection test according to differing trial characteristics and could be used to develop a trial specific signal detection strategy for analysis of emerging harm outcomes. On completion of the simulations, results for accuracy were deemed not to be a good measure of

performance in this setting since low false positive values were resulting in misleadingly high values of accuracy. Therefore, the focus of reporting is on the power and the rate of false positive for examined tests.

Of the evaluated methods none were appropriate at screening emerging harm outcomes to detect time dependent ADRs in trials smaller than 5000 participants (n=2500 per treatment group). For screening purposes when no prior assumptions can be made about the emerging events the combined test is recommended in large samples (n≥5000). The beta-binomial and gamma-Poisson models can also be utilised as screening tools in smaller samples (n≥800) when the specific aim is to detect a doubling of risk (100% increase in background event rates) but requires large samples (n≥5000) to detect smaller increases (50% increase in background event rates). The beta-binomial and gamma-Poisson models should also be used with the understanding that there is likely to be an inflated rate of false positives (exceeding standard accepted rate of 5%). However, an inflated rate of false positives might be deemed acceptable if detected signals simply triggers closer monitoring of the event in ongoing or future studies where a confirmatory analysis approach could be adopted. That said with increasing sample sizes even the smallest of effects would be detected, and such an excessive detection rate would likely become overly burdensome making the approach impractical. The balance between under and over-detecting signals requires careful thought and is likely to be context specific, this is discussed further in the limitations section. Finally, none of the examined methods were suitable to detect small risk increases (25% increase in background event rates).

If there is prior information available on background event rates or if specific times post treatment exposure are of concern more targeted approaches can be adopted. For example, if screening events to detect reactions in the period immediately (i.e. within first few days) after drug exposure e.g. anaphylactic shock, then the Weibull model with ancillary parameter can be used in samples of 1000

or more but caution in interpretation is required as this test may over-identify signals. To avoid excessive numbers of false positives (i.e. avoid false positive rates exceeding 5%), the combined test with samples of 2000 or more can be used instead. If background event rates are known to exceed 10% then the combined test can be used in slightly smaller samples (n≥2000). However, when background event rates are as low as 1%, all of the examined methods were inappropriate as a means to detect signals for ADRs, with power remaining below 70% across methods in the investigated sample sizes, with failure rates as high as 39% in the smallest sample sizes. As such of the examined methods, none are considered appropriate to detect uncommon or rare ADRs.

Time-to-event methods offer advantages in terms of screening for early effects with the Weibull models and combined test exhibiting the largest power for detecting immediate (within the first few days of exposure) and early (within a month of exposure) effects, with sufficient power across a range of sample sizes and trial characteristics. However, the performance of the beta-Binomial model and to some extent the Fisher's exact test, neither of which utilise information on time of events, revealed that in larger sample sizes with later signals, incorporation of time is of less importance to detect signals for ADRs. Time-to-event methods can therefore be advantageous in smaller samples and where signals are raised early relative to initial exposure.

Table 7.8: Recommendations when undertaking signal detection analysis on emerging harms to detect time dependent ADRs

	Information available	Recommendation
1	*General guiding principle when no information available but assumed minimum background event rate of 1% and sample size is ≥5000*	*For screening purposes when there is no prior knowledge about background event rates in the study population, the increase in background rates due to ADRs or time increases due to ADRs are likely to occur, the **combined test** can be used as a screening tool to analyse emerging events in samples of **5000** or more where the minimum background event rate is assumed to be ≥1%.*
2(a)	*General guiding principle when no information available and sample size is ≥800*	*For screening purposes when there is no prior knowledge the **beta-binomial** and **gamma-Poisson** models are recommended to detect at least a **doubling** in event rate where an excessive number of false signals is not of concern (i.e. false positive rate can exceed 5%).*
2(b)	*General guiding principle when no information available and sample size is ≥2000*	*For screening purposes when there is no prior knowledge and excessive numbers of false positives is of concern the **combined test** or **Fisher's exact test** are recommended to detect at least a **doubling** in event rate (i.e. wish to maintain false positive rate of 5%).*
3	*General guiding principle when no information available and sample size is ≥5000*	*For screening purposes when there is no prior knowledge the **beta-binomial** or **gamma-Poisson** models are recommended to detect **50% increases** in event rates but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate can exceed 5%).*
4(a)	*Strategy based on the time of increase due to ADRs and sample size is ≥1000*	*If screening events that occur immediately (within a few days) after treatment exposure the **Weibull model with ancillary parameter** is recommended but caution is required as this test may result in an excessive number of false signals (i.e. false positive rate exceeds 5%).*
4(b)	*Strategy based on the time of increase due to ADRs and sample size is ≥2000*	*If screening events that occur immediately (within a few days) after treatment exposure e.g. anaphylactic shock and excessive numbers of false positives is of concern (i.e. wish to maintain false positive rate of 5%), the **combined test** is recommended.*
5	*Strategy based on the time of increase due to ADRs and sample size is ≥5000*	*If screening for early increases in events (one-month ± 2 weeks) e.g. suicidal ideation following exposure to a SSRI, the **combined test** is recommended.*
6	*Strategy based on the time of increase due to ADRs and sample size is ≥5000*	*If screening for longer term increases in events (month 1- 11) e.g. acute liver injury, neutropenia or later effects such as peripheral neuropathy, the **beta-binomial** and **gamma-Poisson** models are recommended but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate exceeds 5%).*
7	*Background event rates of 10% and sample size is ≥2000*	*If the background event rate is known to be approximately **10%** or more, the **combined test** is recommended to screen events.*
8	*Background event rates of 5% and sample size is ≥2000*	*With known background event rates of **5% or more,** the **beta-binomial** and **gamma-Poisson models** are recommended but the likely inflation in the number of false positives should be taken into consideration (i.e. false positive rate exceeds 5%).*
9	*Background event rates of 1%*	*When background event rates are as low as 1% all of the examined methods were inappropriate as a means to detect signals for ADRs.*

Figure 7.6: Design considerations and recommendations when developing a signal detection strategy for the analysis of emerging harm outcomes

Developing a signal detection strategy for ADRs											
General screening approach			**OR**	**Time specific screening approach**			**OR**	**Screening for specific effect size**			
Design constraint	→	**Test choice**		**Design constraint**	→	**Test choice**		**Design constraint**	→	**Test choice**	
Sample ≥ 5000	→	Combined test		Detect immediate effects in samples ≥ 1000	→	Weibull model with ancillary parameter		Detect a doubling in effect size if sample ≥ 800	→	Beta-binomial or gamma-Poisson	
Sample ≥ 2000 & background event rate ≥ 10%	→	Combined test		Detect immediate effects in samples ≥ 2000 if inflated false positive rate of concern	→	Combined test		Detect a doubling in effect size if sample ≥ 2000 & inflated false positive rate of concern	→	Combined test or Fisher's exact test	
Sample ≥ 2000 & background event rate ≥ 5%	→	Beta-binomial or gamma-Poisson		Detect early effects in samples ≥ 5000	→	Combined test		Detect a 50% increase in effect size if sample ≥ 5000	→	Beta-binomial or gamma-Poisson	
				Detect non-immediate through to late onset effects in samples ≥ 5000	→	Beta-binomial or gamma-Poisson					

7.4.1 Comparisons to existing work

Others have also explored the incorporation of time-to-event information into the analysis of emerging harms.[155, 158, 197, 275] These have typically focused on quantifying estimates of treatment effect for prespecified events of interest rather than a tool to detect signals for time-dependent ADRs. In addition, these alternative approaches have sought to tackle issues not considered in this chapter such as incorporation of information on recurrent events, variation in patterns of censoring and competing risks. Methods considered in this chapter are only suitable for handling time-to-first event data, except the gamma-Poisson model, which can be implemented on repeated events but further simulation work would be needed to examine its performance in this setting. Whilst the time-to-event approaches investigated in this work (e.g. the Cox proportional hazards model, Weibull time-to-event models and the combined test) can account for censoring, further simulation work would be required to explore the impact of different patterns of censoring due to, for example, loss to follow-up or withdrawals, the impact of such events being associated with outcomes i.e. informative censoring, and how competing risks due to events such as death should be handled.

With similar motivation to this work, Xia et al. looked at the utility of Bayesian hierarchical models to improve the analysis of emerging harms. Unlike the work described in this chapter their focus was on the analysis of binary and count outcomes (including incident densities which allow for differential periods of risk among patients) utilising logistic and Poisson regression models (summarised in chapter 3 section 3.4.4).[177] Whilst the latter can account for differential exposure, it assumes, like the gamma-Poisson model, a constant event rate over time and cannot identify time-dependent effects. In addition, the methods proposed model the entire body of events utilising the structure of events within body-systems to control for multiplicity of tests with the ultimate aim of identifying a "*flagging device or a screening tool that can strike a proper balance between "no adjustment" versus "too much adjustment""* i.e. controlling the overall type-I error rate accounting

for multiplicity of tests. The aim being to prevent over-identification of signals across the body of events as opposed to the work presented in this chapter that aimed to achieve balance between power and false positives, thus reducing the risk of missing signals of potential harm. The false positive rates presented in this work are per test rather than as a global rate to control. In addition, the Bayesian hierarchal models are based on analysis of summary data and do not utilise the subject level information available. Whilst the work reported in this chapter did not include subject level information, this could be easily incorporated into each of the time-to-event models via covariate adjustments (e.g. the Weibull and Cox time-to-event models, and the combined test).

The work by Gould is perhaps the most similar to the work undertaken in this chapter. Gould proposed screening tools for the analysis of emerging harms utilising a Bayesian framework (summarised in chapter 3 section 3.4.4).[174, 175, 203] The methods proposed utilise the beta-binomial and gamma-Poisson distributions to generate posterior probabilities that the intervention group event rate is generated by the same process as the control group event rate, raising signals if the posterior probabilities indicate the intervention group event rate is generated by a larger process. Whilst there is overlap with the Bayesian approaches proposed by Gould and those investigated in this chapter, further work would be required to investigate how the methods proposed by Gould compare to approaches that incorporate information on time events occur and that allow for non-constant risk over time.

There has also been much work outside of the harms arena that focuses on detecting treatment effects in the presence of non-proportional hazards. Royston and Palmer initially developed a joint test to retain power in the presence of a disproportionality and this was surpassed in terms of performance by development of the combined test in 2016, which is considered in this thesis.[285, 293] In the initial presentation of the combined test the authors reported that the test performed well in

terms of power in the presence of early effects but had reduced power for later effects and had a consistent false positive rate of 5% across scenarios, which is reflected in the results presented in this chapter.[285] In 2020 the same authors undertook a more comprehensive comparison of the combined test comparing it to eight alternatives through examination of power and false positive rates in the presence of early, late and near proportional hazard treatment effects.[294] In the 2020 paper, the authors also introduced two new approaches for the analysis of treatment effects in the presence of non-proportional hazards. The first was the weighted-combined (WC) test, which aimed to improve the performance of the combined test in the presence of a late treatment effect and the second was the modified versatile weighted log-rank test (mVWLR), which aimed to improve the test performance in the presence of early effects but low event rates. Results presented by the authors indicated that the combined test continued to perform well in the presence of an early effect, and that either of the new tests could be used when expecting a later effect. In addition, where no information is available on treatment effects the authors recommended the mVWLR. Whilst these new tests sound potentially very promising for screening emerging harms to detect potential time-dependent ADRs, the methods behind both have not been comprehensively published and neither are currently implementable in standard statistical software and as such have not been explored further in this chapter, but this is an area to be explored in future work.

Whilst the results for the Weibull time-to-event models presented in this chapter, in part, agree with the results from the observational setting that motivated their development i.e. the Weibull models ability to detect early effects, censoring datasets at the mid-way point through follow-up did not prove to be an effective solution to identify non-immediate effects in the RCT setting as it did in the observational setting.[124, 265] This is explained by the smaller sample sizes investigated as compared to the observational setting, and the limited number of censoring periods in the double-Weibull approach compared to the Weibull tool, and the adjustment of the p-value to detect signals to 0.025

in the double-Weibull approach, which was maintained at 0.05 for the Weibull tool in the observational setting. Future work could examine the effect of retaining a p-value of 0.05, which may be a more reasonable approach in a signal detection framework where over-identification of signals could be seen as less problematic where the aim is to detect events of interest for further monitoring. Further work is needed to explore the potential impact of each of these factors.

7.4.2 Limitations and future work

When drawing inference from the results of this chapter there are several points of note. Firstly, the assumption of a common standard deviation in all but the immediate (e.g. within a few days) reaction scenario was adopted for simplicity but is less likely if there is a suspicion that variation increases with time from initial exposure. To investigate such patterns further simulation work would be required.[124]

A threshold of p-value ≤ 0.05 (or p-value ≤ 0.025 in case of the double-Weibull and double-Cox models to adjust for multiple tests) has been used to detect signals for ADRs throughout this chapter. Differing thresholds to raise signals might be worth consideration with careful thought given to the trade-off between the rate of missed signals and rate of false positives. This trade-off will be context specific, for example, dependent on the size of trial, participant profile, approach to recording events and resource. The practical implications of varying the signal threshold should be explored in future work.

When analysing emerging events the Bayesian approaches (beta-binomial and gamma-Poisson models) require an assumption about the distribution of the event rate distributional parameters i.e. the prior distributions. The commonly used 'non-informative' or minimally informative, Jeffreys

priors are assumed in this work for both the binomial and Poisson parameters. Alternatives have been proposed such as the neutral prior, which assumes $Beta(1/3, 1/3)$ or Bayes Laplace prior, which assumes $Beta(1, 1)$ but these have not been investigated in this work. The potential pitfalls of using the non-informative Jeffreys prior is the potential to "*suppress the importance of observed data*", especially in the case of rare events which result in non-informative priors becoming far more informative than intended.[295] More specifically it has been shown that in the case of zero observed events in small samples too much probability mass is concentrated close to zero and the posterior distribution is no longer dominated by the observed data. In such scenarios an alternative from the $Beta(1, b)$ family, where $(b > 1)$ has been suggested as being more appropriate and as such further work is needed to understand the sensitivity of these results to the chosen priors. In addition, the utility of these approaches has only been investigated where it is assumed no prior information is available. It is likely that some information will be available from historical trials, at the very least on event background rates, thus negating the need for non-informative priors and potentially providing a more powerful analysis strategy. However, the practicalities of such an approach for the analysis of emerging harms will not be straightforward, with priors needing to be specified in trial set-up with input from the clinical team for all possible events. This is likely to be prohibitively time-consuming and may be hindered by clinicians' unfamiliarity with such approaches.

The Bayesian approaches also require prespecified thresholds of risk to be defined. The results presented raised signals if the probability that the risk ratio (beta-binomial approach) or incident rate ratio (gamma-Poisson approach) exceeded one was greater than 0.9. However, in different trial scenarios it might be more appropriate to choose different values based on what would be considered a clinically important effect. Again, this is reliant on clinical input from the outset, which requires a clear understanding of the thresholds to be defined and comprehension of the inferences that can be made, which without clear guidance could impede clinicians' ability to use in practice

and deter adoption. It is also possible to look at different treatment effect estimates such as the odds ratio or risk differences. The impact of looking at varying treatment effects on the overall power and false positive rates of the beta-binomial and gamma-Poisson models are presented for information in tables A7.16 to A7.19 in appendix A7.5.

Bayesian approaches have much potential under a signal detection framework. They avoid the need for interpretation based on p-values and thus the temptation to interpret under a hypothesis testing framework and the inappropriate conclusions that follow such as non-significant p-values interpreted as "*evidence of a good safety profile*".[56] However, use of Bayesian methods is still relatively rare in the clinical trials arena and in the immediate term this unfamiliarity and perceived complexity may hinder adoption of such approaches. Thus any recommendations that utilise Bayesian approaches would require clear explanation and education to support adoption by the clinical trials community.

There are reactions where calendar time may not reflect exposure, for example, Crowe et al. pointed out that "*for drugs taken on an as-needed basis, analyses using the number of doses taken may be helpful because time on study may not be highly correlated with exposure to the drug*". [7] Therefore a future avenue for exploration is a suitable signal detection tool to improve the analysis of harms for drugs taken as-needed.

It should be kept in mind that the results presented, and inferences drawn in this chapter, as with all simulation studies, are limited to the generated simulated scenarios. Whilst these aimed to cover a broad range of 'typical' simple, trial scenarios, practicalities such as the time taken to perform such simulations, limit what can be feasibly looked at. For example, the simulations were restricted to

scenarios where the increased risk above the background rate was at most double but in practice larger effects are likely to be seen, for example, the threefold increase in risk of a deep vein thrombosis in women taking the oral contraceptive pill.[296] Simulations also only looked at times of ADRs generated using the normal distribution. Future work could look at detection rates for larger effects and generating ADR times using different distributions such as the lognormal distribution. I have also assumed complete follow-up i.e. simulated scenarios did not incorporate a censoring mechanism. In practice there are always likely to be withdrawals, losses to follow-up and competing events such that some participants fail to complete follow-up. These simulations offer insight under ideal conditions. When censoring is balanced between arms these simulations also hold as the minimum number needed for full follow-up data. However, further work is required to assess the impact of censoring. It is important that these limitations, as well as those outlined above are given full consideration before the guiding principles outlined in table 7.8 and figure 7.6 are implemented in a real-world clinical trial setting.

In addition, an underlying assumption of the simulations is that events unrelated to the intervention will occur at a constant background rate (i.e. the AE background rate). However, our own recent work has drawn the validity of this assumption in the RCT setting into question. In routinely collected datasets events unrelated to the intervention can happen and are collected at any point in time, hence demonstrate a constant background rate, but the collection process for emerging events in the RCT setting, where there are specific visits for collection, may undermine this assumption. For example, we have observed that in the initial period post-randomisation there is a peak in the number of events recorded in both intervention and control groups, which are likely unrelated to the intervention but instead related to the trial processes. However, this will likely become less of an issue as more trials become embedded in the NHS setting using routinely collected data sources.

Differing patterns of the underlying background event rate and the impact this has on the examined tests to detect signals for ADRs are yet to be explored.

This work advocates a signal detection approach for the analysis of emerging harms, but there is a risk that the identified methods will be used inappropriately in a hypothesis-testing framework. The aim of hypothesis tests is to detect a statistically significant difference of a prespecified minimum clinically important difference (MCID) at a prespecified level of significance and power. Whilst a hypothesis test approach is suitable for prespecified primary and secondary outcomes, it is inappropriate for the analysis of emerging harm outcomes where we do not prespecify events in advance, nor a MCID, therefore caution is required to prevent over-interpretation of the results produced by the application of these methods. For example, non-significant p-values do not indicate "*the treatment had a good safety profile*" nor do significant results provide a threshold for selection of the events to present. As Gould remarked "*hypothesis testing in this circumstance is questionable because the hypotheses to be tested have not been defined before obtaining the data that will be used to test them*."[174] Instead the approach should be to analyse the data in order to detect signals for ADRs that lead to closer inspection of signalled events and to provide information that can be used to inform the design and monitoring of future clinical trials as well as post-marketing surveillance activities. As summarised by Xia et al. "*signal detection is a process: Once a signal is identified, it needs to be further investigated, hypothesized, characterized, verified, and quantified"* going on to add that *"the field of clinical trial signal detection is still in its infancy. More research and more experience with existing models are needed."*[177]

In addition, whilst the methods chosen for comparison purposes were informed by a comprehensive methodological review, I have subsequently identified two promising new approaches proposed by Royston and Palmer (the authors of the considered combined test).[294] At the time of writing, these tests have not been fully presented and there is no code for implementation, but to ensure

recommendations remain pertinent, these and other newly emerging methods will need to be evaluated and recommendations updated in a timely manner.

7.4.3 Conclusions

Whilst several of the investigated tests have been found to be of use for screening purposes in specific scenarios, all require large sample sizes, thus there remains a need for a tool that can detect signals for ADRs in smaller studies. This work has shown that whilst the widely used Fisher's exact test consistently maintains low levels of false positives across investigated scenarios, it is an inappropriate tool for signal detection purposes in all but the largest of trial sample sizes and should no longer be used as the default method for the analysis of emerging harm outcomes.[177]

Cornelius et al. highlighted that in the observational setting it is "*desirable to have a tool that has the ability to detect a signal regardless of when that signal occurs*" and the same is true for the clinical trial setting.[124] In this work nine potential screening tools have been evaluated and guiding principles on how they can be best utilised to help identify time-dependent ADRs from the body of emerging harm outcomes have been provided. The Bayesian approaches and combined test were identified as suitable screening tools for the analysis of emerging harms in samples of more than 5000 and more targeted strategies are required in smaller samples. However, further work is still needed to identify methods to screen emerging events in samples smaller than 5000.

8. Discussion and recommendations

8.1 Summary

RCTs designed to address questions of efficacy also collect valuable information to allow an evaluation of harm of interventions under investigation contributing to the wider picture of the developing harm profile. The underlying premise of this thesis was that this valuable information on harm was not being fully utilised and hence an opportunity to compare rates of events to help identify signals for potential ADRs was being missed. The overarching aim of this thesis was to corroborate this belief, and if confirmed, explore existing and propose new methods, and develop strategies to improve the analysis of harm outcomes in phase II and III pharmacology trials. Thus enabling presentation of more informative harm profiles and facilitating the detection of ADRs.

In chapter two I examined analysis practice for harms as presented in journal publications to gain a better perspective of what current practice looks like, specifically looking at articles in top ranked medical journals to provide insights into current 'best' practice. This helped to highlight areas for improvements but also identified examples of good practice that could be exemplified and built upon. Chapter three identified existing methods available to trialists specifically for the analysis of harm outcomes in a clinical trial setting. This helped to expose methodological avenues that were explored in later chapters and highlighted methodological gaps. Chapter four sought to understand practices beyond those presented in journal publications and to identify any barriers clinical trial statisticians experienced when analysing harms. It also helped to identify what clinical trial statisticians believe would help them overcome these barriers and which areas trialists believe should be prioritised for further research. The aim being to ensure any development work, undertaken in this thesis or future work was designed with solutions to these obstacles in mind. Chapter five describes the work undertaken to directly address one of the main barriers and priorities for future work identified in chapter four and which was flagged as a priority by journal

editors as a result of early dissemination activities – that being a lack of guidance on appropriate graphical summaries for harm outcomes. This consequently led to the development, in conjunction with the UKCRC statisticians' operations group, to a set of recommendations for visualising harm outcomes, the process and outcome of which are detailed in chapter five. Based on a methodological gap identified in chapter three and potential solutions called for in chapter four, the work described in chapters six and seven explored methods that could be utilised under a signal detection framework to detect ADRs. Shifting the focus from one of, inappropriately, conducting hypothesis tests in line with the analysis of primary and secondary outcomes and reframing the analysis of harm outcomes to one of detecting signals of harm that can be investigated further in ongoing or future studies – this idea is discussed in chapter seven. In this final chapter, I discuss the main findings of this thesis, the potential implications they have on analysis practice and how they relate to recent work in this area. I also provide recommendations that can be adopted in both the immediate and short term and highlight future work that I believe needs to be undertaken.

8.2 Main findings relating to current practice

8.2.1 Summary

The review of current practice described in chapter two confirmed the supposition that data on harm outcomes is not being fully utilised, providing evidence that inappropriate and inconsistent practices are often being undertaken, thus preventing comprehensive summaries of harm profiles from being established.[45] Specifically evidence gathered indicates that there is a reliance on simple approaches with frequency tables dominating practice and an inappropriate use of hypothesis tests for the analysis of harm outcomes. There was evidence of data misuse with a pervasive practice of dichotomising continuous outcomes such as laboratory and vital signs data, resulting in a waste of valuable information, a practice also often seen when analysing efficacy outcomes.[112] There was also a suggestion that industry funded studies were more likely to present a more comprehensive summary of events including information on severity, seriousness, relatedness, duration and timing

of events in comparison to academic led trials. In addition, there was great variation in the means by which events were chosen to include in journal articles, many of which relied on arbitrary rules, the impact of which and alternatives to, will be explored further in future work.

8.2.2 What could clinical trial statisticians be doing?

Somewhat unexpectedly, the scoping review for analysis methods undertaken in chapter three revealed a broad range of methods for the analysis of both prespecified and emerging harms. There is limited evidence of the application of any of these methods.[130] Whilst many of the identified methods could be adopted into practice immediately, with some being explored in later chapters of this thesis, without a more formal quantitative comparison it is unclear which, if any of the identified statistical methods should be promoted for use. A better understanding of how the methods compare to each other both in terms of accuracy of results produced and ease of implementation, coupled with a better understanding of the reasons for a lack of uptake would enable this, thus allowing researchers to make an informed choice about the best analytical approach to adopt. This too is an area to be explored further in future work.

8.2.3 Feedback from clinical trial statisticians

The survey of clinical trial statisticians from across UK academic institutions and industry confirmed routine use of sub-optimal analysis practices. While there was a moderate level of awareness of alternative approaches identified in chapter three, reported use in applied practice was limited.[216] The overriding message from this piece of work was that trialists needed guidance and training on appropriate methods for the analysis of harm outcomes. It also revealed that results presented in journal articles are likely to reflect a subset of results produced by trial statisticians and that perhaps we should be looking beyond the practices of statisticians to further enable change. This was also indicated in responses that suggested that the potential influence of journal editors and regulators

resulted in many of the practices observed and that we should be arming trialists with the confidence to deny requests for inappropriate analysis. With one survey participant responding that we should "*encourage statisticians to push back on 'bad' practices that are asked of them (it can work....) but comes with risks that trial team may not be willing to take (i.e. articles rejected) and perhaps is something that those with more experience would be more confident in tackling*." In addition, examination of guidelines from the ICH offer some insight into the rationale behind observed practices. For example, the ICH E3 guidelines on structure and content of clinical study reports suggest that "*tables should list each adverse event, the number of patients in each treatment group in whom the event occurred, and the rate of occurrence*" and that "*adverse events should be grouped by body system. Each event may then be divided into defined severity categories (e.g., mild, moderate, severe) if these were used. The tables may also divide the adverse events into those considered at least possibly related to drug use and those considered not related, or use some other causality scheme".*[39] The ICH E9 states "*methods to reduce the effect of the background noise may also be appropriate such as ignoring adverse events of mild severity or requiring that an event should have been observed at repeated visits to qualify for inclusion in the numerator. Such methods should be explained and justified in the protocol.*"[6] This not only highlights the need for more objective means to summarise harm profiles but for a cross-industry update on our thinking around the analysis and reporting of harms and the opportunity for cross-industry learnings to support any proposed changes beyond those responsible for conducting analysis.

8.2.4 Differences in analysis practice between academia and industry

The initial work presented in chapters two to four also revealed potentially important differences in analysis and reporting practices between academic led trials compared to industry led trials. The results of the systematic review suggested that industry funded trials are more likely to provide a comprehensive overview of harms. The scoping review to identify existing methods also found that the majority of the methodological work stemmed from within industry, whether solely or

collaboratively undertaken. Whilst overall the survey found similar results according to participants' background, a notable difference related to the use of hypothesis tests, which were used more often by academics, and a greater concern about the acceptability of methods to regulators from industry participants. This highlighted that there may be lessons for academic trials to learn from industry with regard to the analysis of harm outcomes and that opinions of regulators and journal editors may be worth seeking to ensure they are reflected in any proposed changes and that any proposals for change are supported across the clinical trial arena.

8.3 Recommendations

8.3.1 Changes for immediate adoption

Through the work undertaken in this thesis, wider collaborative projects undertaken alongside it and staying abreast of the work of other researchers in this field, I have identified a number of changes that I believe could be adopted into current practice to improve both the results reported and the analysis undertaken for both prespecified and emerging harm outcomes.

Reporting

An appraisal of the state of play as presented in journal articles has led me to conclude that several simple changes are likely to lead to improvements in reporting of harm outcomes. Thus enabling a more comprehensive presentation of the harm profile. Specifically, I believe the following strategies could lead to reports that are more transparent:

1. Encourage trialists to give more consideration to harm outcomes at the design stage and specify analysis plans for both prespecified and emerging events as one would for primary and secondary outcomes. This entails giving careful consideration to the estimate of interest for each analysis and planning how each analysis will be used to draw inferences. For example thinking about the impact of different analysis populations and assessing how

sensitive the results are to such decisions. The 2017 guidance for statistical analysis plans encourage this at the planning stage stating "s*ufficient detail on summarizing safety data, e.g., information on severity, expectedness, and causality; details of how adverse events are coded or categorized; how adverse event data will be analyzed, i.e., grade 3/4 only, incidence case analysis, intervention emergent analysis*".[87] This proposal is also supported by others in the field with Xia et al. stating that *"It is critical to proactively plan for the evaluation of safety data and to ensure that safety signals are detected in a timely manner."* and James et al. who states that *"reporting standards should make clear which datasets were used for analyses."* [297, 298]

2. Promoting implementation of the CONSORT harms checklist when reporting the results of RCTs to ensure comprehensive and clear harm profiles are presented.[28] This would also help standardise reporting practices and could enable a more accurate synthesis of harm data. The work in this thesis and collaborative projects undertaken alongside it have shown adoption of this checklist to be sub-optimal.[80] Whilst the checklist does not guard against all pitfalls identifed in current practice and does not provide specific guidance on analysis methods, its widespread adoption would go some way towards improving the current situation, ensuring more transparent reporting. It is also apparent that journal editors themselves are aware of these inadequacies as evidenced in a recent request from one journal editor for a commentary piece written as a "*call to arms*" highlighting "*what needs doing to solve this ongoing problem*" (personal communication February 2021). Promotion and endorsement from journal editors, as there is for the original CONSORT checklist would be one simple step to help instigate change. In addition, the CONSORT group responsible for the harms checklist recognise its limited impact and an international collaboration to update it with the aim of improving uptake is currently underway. One potential solution being explored is to integrate the items from CONSORT harms into the main statement, thus

highlighting the importance of providing a balanced summary of efficacy and harm outcomes alongside each other in the main report.

3. Another important step to ensure that trialists provide a balanced summary between efficacy and harm outcomes in journal articles is asking them to refrain from the inadequate practice of depositing all information on harm outcomes into supplementary material, whilst also avoiding the use of arbitrary rules for the selection of events to report in the main journal article. This would allow readers to make a risk-benefit assessment solely from the information presented in the main article and would also ensure that we were working towards a more consistent approach of reporting across trials. Work to give clearer guidance on the choice of events to report is planned in future work based on the earlier work of Cornelius et al. and Mayo-Wilson et al. who called for "*standards … to determine which AEs to include in reports of clinical trials and how to report AEs completely in other public sources such as trial registers.*"[56, 57]

Adoption of these strategies, I believe, would ultimately ensure clearer and more consistent reporting of harm outcomes in journal articles.

Analysis

In terms of analysis practices, immediate changes I believe that trialists could make to ensure a more efficient use of data informed by the work in this thesis include:

1. Adopting approaches that utilise all information available on harm outcomes, reducing information loss when analysing at participant level. For example, using information on recurrent events rather than presenting as those who experienced at least one event, and retaining continuous outcomes in their natural form rather than dichotomising. This could include simple changes such as reporting incident rate ratios with a measure of uncertainty

to account for differences in exposure times or using linear mixed effects models to incorporate information on repeated continuous laboratory results or plotting the mean cumulative function to account for recurrent events to explore the burden of harm.

2. Avoid underpowered hypothesis tests and using p-values and measures of precision as null hypothesis tests.[25]
3. Utilise visualisations to improve the communication of complex data on harms in clinical trial manuscripts and reports, enabling clearer summaries of harm profiles to be presented and helping to identify signals for potential ADRs. For example the dot plot and stacked bar chart could help to identify the potential burden of harm participants experience, the Kaplan-Meier plot can help identify potential signals for ADRs for further monitoring in future clinical trials, as well post-marketing surveillance studies, and the mean cumulative function plot can help identify periods when interventions might be consider well tolerated. These ideas are explored fully in chapter 5 and I comment further on this work below.
4. Adopt statistical models that utilise information on time events occur to detect signals for potential ADRs. For example, for screening purposes when no prior assumptions can be made about the emerging events, the combined test from Royston and Palmer is recommended in samples larger than 5000 participants. The beta-binomial and gamma-Poisson models can also be utilised as screening tools in smaller samples (n≥800) when the specific aim is to detect a doubling of risk (100% increase in background event rates) but requires large samples (n≥5000) to detect smaller increases (50% increase in background event rates). These ideas are explored fully in chapters 6 and 7 and I comment further on this work below.

Changes to adopt informed by collaborative work I have undertaken alongside this PhD include:

1. When dichotomisation of continuous harm outcomes is justified, for example to aid clinical interpretation, a distributional approach should be used to analyse the difference in proportions between treatment groups to retain statistical power and ensure an efficient use of available data.[299]
2. Refrain from interpreting results under the same null hypothesis-testing framework we use for primary and secondary outcomes. Instead, reframe the analysis of emerging harm outcomes to one of signal detection using the output of analysis to detect events for further investigation in ongoing or future studies. This would not only help inform the focus of future studies (in terms of data collection, analysis and reporting) but also provide useful insights for any future systematic reviews and post-marketing studies. As Drago et al. comments "*a need for inferential statistics should not be confused with hypothesis testing using p–values, and confidence intervals should not be used as null hypothesis tests*".[25] I explore this premise using visualisations and statistical models utilising information on time events occur in chapters five to seven and comment further on the findings of these chapters below. Bayesian approaches also have much potential in this field avoiding the need for interpretation based on statistical significance and provide an efficient means to incorporate prior knowledge and allow for a cumulative assessment of information. Methods such as the beta-binomial model and gamma-Poisson model are easy to implement, with easy to interpret output and use of such methods in analysis strategies could improve efficiencies across the drug development pathway. Bayesian approaches should be explored further in this setting.

Whilst journal editors acknowledge issues in the analysis of harm outcomes, experience indicates that this is not a topic of huge interest or priority. A continued effort to challenge the status quo is needed. Resources to help trialists move away from simple descriptive approaches are needed and researchers should be encouraged to move to a more suitable framework for the analysis of harm

outcomes. These themes are explored further in an opinion piece led my primary supervisor to be submitted for publication in the Journal of Clinical Epidemiology.

8.3.2 Incorporation of visualisations

Consensus guidelines developed with contribution from the UKCRC CTU statisticians' operations group described in chapter five were devised to help trialists in their choice of visualisations for analysing and reporting harm outcomes in journal articles. Recommendations are given for the variety of different outcome types (e.g. binary, count, time-to-event and continuous outcomes) and for both prespecified and emerging events. The results demonstrate, along with external collaborative work published in a 2020 Trials paper, that visualisations have much to offer for both the evaluation and communication of harm information, addressing one of the key challenges identified in my introduction about how best to communicate and present vast amounts of complex information on harm.[259] Visualisations can also help identify signals of harm in line with the proposed move toward a signal detection framework for the analysis of harm outcomes. The value of visualisations in this setting is a view supported by many working in clinical trials, as evidenced by the investment industry have made into resources to develop better visualisation but also from academics as seen from the ongoing collaboration and support of the UKCRC CTU statisticians' operations group into this piece of work. Unfortunately, evidence obtained in this thesis (chapters two and four) indicates that, to date, this support has not translated into applied practice reported in journal articles. Therefore, as well as practical guidance on which visualisations to use, software and code across the three most common statistical packages (Stata, R and SAS) is promoted alongside each of the recommended plots to ease implementation, with some code specifically being developed as part of this thesis.[204, 205] As also called for in the survey of statisticians, a case study demonstrating the implementation and impact on interpretation of some of these plots has also been prepared and published in collaboration with my supervisor.[259] Future work should focus

on getting these recommendations adopted into practice with potential strategies including development of good practice examples that incorporate visualisations into the analysis section for harm outcomes in statistical analysis plans and development of standard operating procedures detailing good practice for the analysis of harm outcomes that could be implemented nationally across CTUs. The full list of recommended visualisations can be found in tables 5.3 and 5.4 of chapter five.

8.3.3 Signal detection methods

Signal detection approaches explored in chapters six and seven propose a move away from the widely used but inappropriate practice of conducting hypothesis tests for emerging harm outcomes where outputs are interpreted as significant or not to confirm a difference between treatment groups. Instead using statistical analysis to detect signals for potential harm for further investigation in line with practice in the pharmacovigilance setting. The construct of type I error and its inflation due to multiple tests being performed is less problematic in signal detection setting where instead of making definitive conclusions, the output simply triggers closer monitoring of the event in ongoing or future studies. However, there are added complexities that need careful consideration. For example, thresholds at which to raise a signal need to be specified and careful consideration needs to be given to the trade-off between the rate of missed signals and rate of false positives, which is likely to be context specific, varying by intervention and/or population under investigation. Thus, this is a more complex analytical framework than the traditional hypothesis-testing framework.

The aim of the work I undertook was to identify if there was an objective, signal detection method that could identify time-dependent ADRs from the body of emerging harm outcomes. The supposition underlying the investigated approaches is that events not associated with an intervention will tend to occur at a constant rate over time, in contrast to ADRs where the causal

mechanism can mean that their occurrence is temporal. A novel approach based on the Weibull proportional hazards model was proposed, building on work from the observational setting and this method was compared to a range of alternative approaches that also utilised information on time events occurred, as well as the often used Fisher's exact test. Unfortunately, of the evaluated methods none were appropriate for use as a screening tool in trials smaller than 5000 participants. Approaches such as the recently developed combined test from Royston and Palmer and the Bayesian beta-binomial model showed utility in specific scenarios such as when there is prior information available on background event rates or if specific times post treatment exposure are of concern and guiding principles have been developed outlining recommendations for use.[285, 287] This work also confirmed that the widely used Fisher's exact test is an inappropriate means to analyse emerging harms in all but the largest of trials and trialists should refrain from using it as a default statistical test for analysis of emerging harm outcomes. The full list of recommendations for use can be found in table 7.8 and figure 7.6 of chapter seven along with limitation and cautions for use in section 7.4.2. This also highlights there is still much to be explored in this setting. Whilst further development work is underway to identify suitable statistical approaches for signal detection in trials smaller than 5000 participants with one simple potential solution being to reduce the threshold for a signal (e.g. from 0.05 to 0.1) this still requires further work to explore the trade-off between the rate of missed signals and rate of false positives. In the immediate term, the visualisations recommended in chapter 5 have much to offer.

8.4 How does this work compare with recent research in the field?

A recent 2020 review by Patson et al. showed continued suboptimal analysis for harm outcomes as demonstrated in chapter two. They also provided recommendations on broad analytical approaches that could be adopted according to outcome type, which were broadly in line with my own recommendations to adopt analytical approaches that utilise all information available. For example,

the authors propose implementing survival methods for time-to-event outcomes and fitting Poisson or negative binomial models for count outcomes.

Also recognising the potential importance of time in the analysis of harms, Stegherr and colleagues recently proposed and demonstrated the application of an analysis strategy under a time-to-event framework to account for censoring and competing events to quantify treatment effects when analysing prespecified harm outcomes of interest.[275] With a similar focus, Unkel et al. suggested a framework for harm outcomes according to the estimand framework, proposing analysis strategies for "*time to the occurrence of the first AE of a specific type".*[62] However, unlike the time-to-event approaches I explore, neither paper consider analysis of emerging harm outcomes, which this thesis confirmed, in practice remains reliant on simple approaches. I provide a more detailed summary of the signal detection methods I explored compared to alternative time-to-event approaches proposed for harm outcomes in the literature in chapter 7 section 7.4.1.

In 2009, the SPERT suggested trialists operate under a program safety analysis plan which is a combined analysis plan for all stages along the entire development pathway.[7] This industry viewpoint suggests discussing such plans with the FDA with a focus on getting new products to market – "*approach calls for standardization of data collection methods, early and repeated safety assessments during development including periodic meta-analysis, and review of safety data from all available sources at regular intervals during the marketed use of a product."* Ample resources and guaranteed funding make programmes of development feasible in the industry setting. Unfortunately, the nature of public funding bodies hinders such an approach in academic led trials. However, the authors recognition of the value of data on harm outcomes obtained in RCTs is very much in line with the viewpoint from which this thesis is written i.e. "*data from randomized trials provide the most interpretable evaluation of safety".* They also propose several key principles that

could be adopted in clinical trials outside of industry. For example, in line with the work presented in this thesis they propose making the distinction between prespecified and emerging events and advocate that analysis plans should describe "*standard data collection plan and analysis section*". However, very much in keeping with the subsequent work of Stegherr et al. and Unkel et al. the more sophisticated analytical approaches proposed are for prespecified events and the authors endorse simple descriptive approaches for emerging events. As such, the work of Chuang-Stein in 2013 highlighting the importance of statisticians in the assessment of drug harm and the contributions that the sector can make to the "*development of methods and tools for risk assessment and signal detection"* is still pertinent in 2021, especially in regard to the analysis of emerging harm outcomes.[139] However, as highlighted in the recent work of Lopes et al. a "*substantial shift*" in how trialists report and interpret information on harms would be needed to allow statisticians to change their analysis practices. However, unlike Lopes and colleagues I am optimistic that continued efforts, with the support of key stakeholders such as the UKCRC CTU statisticians' operations group, can lead to change.[300]

8.5 Limitations of the work undertaken in this thesis

This thesis focused on the final analysis of trial outcome data and whilst I identified methods specifically designed to monitor harm outcomes in ongoing studies in chapter three, I did not explore these further as there were too many to be undertaken within the timelines for this PhD. This is an area being explored by others as exemplified by the work on producing DMC reports discussed in chapter five.[52-54, 261] I also framed this thesis specifically to trials of pharmacological interventions, non-CTIMPS or trials of complex interventions are likely to present their own challenges as discussed by Moody et al. and Papaioannou et al.[49, 301] However, feedback obtained at the UKCRC CTU statisticians' operations group, specifically from those working in trials of complex interventions indicated that the research presented in this thesis could act as a foundation for further research in those areas.

Whilst I sought the perspective of those directly responsible for undertaking analysis on many aspects of this thesis there was no formal involvement of regulators or funders or journal editors, thus the findings of this thesis lacks their perspectives. It is clear from the findings of chapter three that these parties influence practice and that endorsement of any proposed changes will be needed when undertaking dissemination, as well as incorporating their perspectives to progress the area and is an important avenue to explore in future work. Perspectives and priorities sought in my PhD work were primarily UK focused and there is a potential that priorities in other countries might differ. My involvement in the CONSORT harm checklist, which is being updated by a team comprised of a wide-ranging international cohort of researchers working in clinical trials, demonstrated to me that the issues are similar outside of the UK and it is apparent that the need for change is echoed internationally.

Solutions to improve practice investigated in chapters five to seven focused on implementation in parallel arm trials. Visualisations for multi-arm settings were considered but such studies were not the focus and need more careful thought as to the most appropriate approaches. In addition, it would have been helpful to explore the utility of the signal detection tests identified in chapter seven in clinical trial datasets with known ADRs. However, finding interventions where ADRs are considered 'known' is difficult as there is often dispute about causality. In contrast to efficacy outcomes where selecting suitable trials as case studies to demonstrate application of a method is eased by the fact that trials are designed around efficacy outcomes. There also needs to be careful consideration of the applicability of the simulation results on individual trials before the recommendations are adopted, as the findings are reliant on the underlying assumptions of the simulations and how accurately they reflect real world datasets. For example, the assumption of a constant background event rate is only likely to hold in placebo controlled trials.[302]

More specific limitations of each piece of work undertaken in this thesis are reflected upon and discussed in the discussion section of each relevant chapter.

8.6 Limitation of RCTs for the analysis of harm outcomes

RCTs have smaller sample sizes and follow-up periods compared to observational studies, which restrict their ability to detect rare events and those with long latency. Strict inclusion criteria in RCTs can make results less generalizable and less likely to detect drug interactions and events in more complex populations. However, no phase of the clinical research pathway is sufficient alone for the detection of harms and there are limitations at each stage. For example in the post-marketing setting there is *"evidence of significant and widespread under-reporting of ADRs to spontaneous reporting systems including serious or severe ADRs".*[27] My own view, echoed by many in the field, is rather than advocating that confirmatory conclusions on the harm profile be made from phase III clinical trial results, that instead results add to the body of data from earlier phases and inform later phases to help develop the harm profile.[7, 25]

8.7 Strengths of this thesis

I believe one of the key strengths of this thesis was working with and gaining feedback directly from clinical trialists to find out what they needed to help improve the analysis of harms. In addition, I sought their input on priorities to inform the direction I took in this thesis, as well as their contributions directly in the development of the recommendations for visualisations described in chapter five.

I also believe early dissemination of the results to the clinical trial community via publication and oral presentations at a number of conferences helped inform the direction of this thesis. For example, one such dissemination activity resulted in informal input from journal editors that led to the work to develop the recommendations for visualisations. I was also able to disseminate results via less traditional platforms such as blog posts, Twitter and contributions to podcasts such as the Effective Statistician. Such activities helped to establish my status as an active researcher in this field, which led to ongoing international collaborations some of which have already resulted in outputs that aim to change practices in the field of analysis and reporting of harm outcomes.[80]

One of the main outputs of this thesis is a set of practical recommendations for trialists to use to inform their choice of visualisation to present a more informative and comprehensive summary of the harm profile. This set of recommendations with signposting to software for implementation directly addresses one of the key needs identified by trialists i.e. the need for guidance and examples of use to inform improved practice.

I also developed and explored the utility of a novel, signal detection approach to address the problems of analysing harm outcomes under a null hypothesis-testing framework. While the utility of investigated tests was limited, there was some important findings with practical implications for practice. For example, it confirmed that the widely used Fisher's exact test was not an appropriate method to identify signals and revealed promising avenues for exploration in future work, such as the newly developed weighted-combined test and the modified versatile weighted log-rank test recently proposed by Royston and Palmer.[294]

In addition, I have been actively involved in a number of related projects throughout this thesis contributing to the wider body of research in this area.

8.8 Ongoing collaborative projects

Related work I am actively involved in that is already underway includes involvement in the update of CONSORT harms checklist for reporting harms and several collaborative projects with my supervisor. The latter includes: a project looking at how trialists select events to report in the main publication of trial results; a project comparing how clinically informed selection of events to present in the main publication of trial results compares to using the outputs of simple regression analyses and events highlighted of importance by patients; and exploration of a distributional approach to screen clinical and biological data to detect signals for potential ADRs. We are also exploring whether there is a need and interest for a special interest group focusing on harm outcomes as part of the MRC-NIHR Trials Methodology Research Partnership (TMRP) Outcomes Working Group.

8.9 Future research

The research undertaken in this thesis has already highlighted and motivated further avenues for research. This includes:

1. Further methodological work to identify the most appropriate approaches for analysis. For example, a suitable signal detection tool to screen emerging events was not identified for trials smaller than 5000 participants, which are more common in the academic setting and alternative approaches are still needed. A quantitative comparison of existing methodology as identified in chapter three is also needed so that recommendations on methods to use can be developed.
2. Exploration of Bayesian approaches such as ordinal models to incorporate severity into analysis thus avoiding the need for interpretation based on statistical significance, providing

an efficient means to incorporate prior knowledge and allowing for a cumulative assessment of information.

3. Exploration of more transparent and objective ways to 'select' which events to incorporate in journal publication. With one possible solution to build on the idea first proposed by Cornelius et al. to develop core outcomes by drug class.[56]
4. Work to understand the implications of improved analysis of harm outcomes in phase II/III trials on the wider drug development pathway, as well practical consequences on resource utilisation.
5. Development of case studies to demonstrate application of any proposed methods to encourage and ease adoption.
6. Seeking input and engaging journal editors and regulators to gather their perspectives on any proposed changes and to ensure support for any proposed changes to analysis and reporting practices.
7. Adopting strategies to implement change such as establishing special interest groups with a collective goal of improving analysis of harm outcomes.

All with the ultimate aim of informing the development of universally adopted guidelines and resources to promote best practice for the analysis of harm outcomes in RCTs.

8.10 Overall conclusions

At the start of this thesis, I set out to gain a better understanding of what current practice looked like and explore what, if any, improvements could be made. The initial work described in chapter two confirmed my belief that information on harm was being underutilised and highlighted several practices that could be targeted for change. In chapter three I explored existing methodology for the analysis of harms, which provided a broad overview of the current methodology available to

researchers and helped establish methodological gaps. It also highlighted the need to compare and contrast these methods to provide clarity on which, if any, of the identified methods should be promoted for use. In chapter four, I was able to confirm the findings of chapter two, further highlighting that improvements are needed and gained an understanding of why methods identified in chapter three were not being implemented. Namely, that clinical trial statisticians require guidance on appropriate methods for analysis of harm outcomes with training to support change. It also revealed the importance of seeking journal editors and regulators perspectives to inform and support any proposed changes. This directly informed the direction of future chapters such that in chapter five I set out to develop a set of recommendations for researchers to inform their choice of visualisations when reporting harm outcomes in journal articles. This work helped demonstrate the value of visualisations as a tool to communicate clearer harm profiles and highlighted the support of the UKCRC CTU statisticians for such a resource. In chapters six and seven exploration of statistical tests as signal detection tools to detect ADRs revealed whilst time-to-event based methods offer some utility in large trials, further work is needed to identify a tool to detect signals in a wider range of typical clinical trial scenarios. In addition, it revealed that the popular Fisher's exact test was not suitable for use in any but the largest of samples.

In conclusion, this work revealed that there is an understanding and agreement from within the clinical trials community that how we analyse and report harm outcomes in RCTs needs to change. Efforts are already being made, but these are overwhelming being driven by industry with a focus on the analysis of prespecified harm outcomes, with little thought given publically to the analysis of emerging harms. A cross-sector effort is needed to address the analysis of emerging harm outcomes. Within this thesis I have proposed several solutions for immediate adoption informed by my own work and involvement in collaborative projects, as well as the independent work of others in the field. Visualisations provide a powerful tool to communicate harm offering alternative perspectives

to the traditional frequency tables. Implementation of the recommendations provided in this thesis have the potential to help improve communication of harm outcomes in clinical trial manuscripts and reports, enabling clearer summaries of harm profiles to be presented. However, there remains the need for an easy to implement, objective, signal detection approach that is suitable across a wider range of typical clinical trial scenarios. In addition, clear guidelines for best analysis practice that can be adopted across the clinical trial arena and endorsed by key stakeholders would enable a more coherent and consistent path for change.

Glossary

Active collection: events that are collected by prompting such as by asking participants non-leading questions or using a prespecified checklist.

Adverse events (AEs): any untoward medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment (Source: Edwards IR and Biriell C. Harmonisation in pharmacovigilance. *Drug safety* 1994; 10: 93-102. DOI: 10.2165/00002018-199410020-00001)

Adverse drug reactions (ADRs): a response to a drug which is noxious and unintended …' where a causal relationship is 'at least a reasonable possibility' (Source: Edwards IR and Biriell C. Harmonisation in pharmacovigilance. *Drug safety* 1994; 10: 93-102. DOI: 10.2165/00002018-199410020-00001 and The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Harmonised Tripartite Guideline. E2A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. 1994.)

Bayesian statistics: branch of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event. (Source: Gelman, A; Carlin, JB.; Stern, HS.; Dunson, DB.; Vehtari, A; Rubin, DB. (2013). Bayesian Data Analysis, Third Edition. Chapman and Hall/CRC. ISBN 978-1-4398-4095-5.)

Body-system: alternative term used to **system organ class** to describe grouping of adverse events. See system organ class.

Clinical Research Organisations (CROs): an outsourcing company that a sponsoring company hires as an independent contractor to lead clinical trials and other research support services on its behalf.

Clinical trials of investigational medicinal products (CTIMPs): a clinical trial that evaluates the effects of drug (Source: https://www.kingshealthpartners.org/research/getstarted/ctimps)

Clinical Trials Units (CTUs): specialised biomedical research units, which design, centrally coordinate and analyse clinical trials and other studies

Confidence interval: a confidence interval derived from a valid analysis will, over unlimited repetitions of the study, contain the true parameter with a frequency no less than its confidence level (often 95% is the stated level, but other levels are also used) (Source: Last JM. A dictionary of epidemiology. Oxford: International Journal of Epidemiology, 1988. https://www.oxfordreference.com/view/10.1093/acref/9780195314496.001.0001/acref-9780195314496-e-369?rskey=NkeMGE&result=361)

Consolidated Standards of Reporting Trials (CONSORT): encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials. (Source: http://www.consort-statement.org/)

Data monitoring committees: group of experts external to a study that reviews accumulating data from an ongoing clinical trial to ensure patients participating in trials are at no unavoidable increased risk for harm whilst ensuring that a trial continues for an adequate period and is not stopped too early to answer its scientific questions. (Source: Guideline on Data Monitoring Committees, European Medicines Agency)

Discontinuation of intervention: see withdrawal of treatment

Effectiveness: trials which determine the intervention effect under "real-world" settings. (Source: https://www.nihr.ac.uk/glossary?letter=E&postcategory=-1)

Efficacy: trials which determine whether an intervention works as intended under ideal circumstances. (Source: https://www.nihr.ac.uk/glossary?letter=E&postcategory=-1)

Emerging event: these are events that have not been prespecified to be of interest at the start of the trial. They are events that are reported and collected during the trial and may be unexpected. Includes AEs, and laboratory and vital sign data indicative of harm.

European Medicines Agency: a decentralised agency of the European Union (EU) responsible for the scientific evaluation, supervision and safety monitoring of medicines in the EU. (Source: https://www.ema.europa.eu/en/about-us/who-we-are)

EudraVigilance system: the system for managing and analysing information on suspected adverse reactions to medicines, which have been authorised or being studied in clinical trials in the European Economic Area (EEA). (Source: https://www.ema.europa.eu/en/human-regulatory/research-development/pharmacovigilance/eudravigilance)

Exposure time: period of time in which a person, group or population receive an intervention.

False positives: see type I error

Follow-up: observation over a period of time of a person, group or population to observe changes in predefined outcomes. (Source: https://www.nice.org.uk/Glossary?letter=F)

Food and Drug Administration: is a federal agency of the United States Department of Health and Human Services. It is responsible for protecting public health by ensuring the safety, efficacy, and security of human and veterinary drugs, biological products, and medical devices. (Source: https://www.fda.gov/about-fda/what-we-do)

Frequentist statistics: a branch of statistics based on making assumptions about the process that generated the data and infinitely many replications of them. (Source: https://www.fharrell.com/post/journey/)

Group sequential design: statistical approach in clinical trials where data is sequentially evaluated as it is collected. (Source: https://toolbox.eupati.eu/glossary/group-sequential-design/)

Harm outcomes: Individual events encompassing emerging events and prespecified events of interest.

Harm profile: The summary or burden of the cumulative effect of all harm outcomes.

Health Research Authority (HRA): the central body in the UK that is responsible for the regulation and approval of different aspects of health and social care research. (Source: https://www.hra.nhs.uk/)

Interim analysis: analysis of data before data collection is completed.

MedDRA: standardised medical terminology to classify medical events developed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (Source: https://www.meddra.org/how-to-use/basics/hierarchy)

Multiple testing: simultaneously testing more than one hypothesis. (Source: Altman DG. Practical Statistics for Medical Research. London: Chapman)

Outcome: the impact that a treatment has on a person, group or population. (Source: https://www.nice.org.uk/Glossary?letter=O)

Pharmacovigilance: the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem. (Source: European Medicines Agency. Pharmacovigilance: Overview, https://www.ema.europa.eu/en/human-regulatory/overview/pharmacovigilance-overview (accessed 28/10/2020).)

Phase I: trials of the initial administration of a new investigational product into humans, often conducted in healthy volunteers. (Source: ICH Topic E8 General Considerations for Clinical Trials)

Phase II: trials beginning to look at therapeutic efficacy in patients. (Source: ICH Topic E8 General Considerations for Clinical Trials)

Phase III: trials beginning to look at demonstrating or confirming therapeutic benefit. (Source: ICH Topic E8 General Considerations for Clinical Trials)

Phase IV/Post marketing/surveillance: trials that begin after drug approval that might look at drug-drug interaction or safety studies. (Source: ICH Topic E8 General Considerations for Clinical Trials)

Placebo: an inactive drug or treatment used in a clinical trial

Power: the probability of rejecting the null hypothesis when it is false. (Source: Gardner MJ Altman DG, editors. Statistics with Confidence. London: BMJ Publishing Group. Differences between means: type I and type II errors and power)

Preferred term: distinct descriptor (single medical concept) for a symptom, sign, disease diagnosis, therapeutic indication, investigation, surgical or medical procedure, and medical social or family history characteristic within the MedDRA classification system. (Source: https://www.meddra.org/how-to-use/basics/hierarchy)

Prespecified event: individual events that are listed in advance as harm outcomes of interest to follow. They already be known or suspected to be associated to the intervention, or followed for reasons of interest.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA): an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses. (Source: http://www.prisma-statement.org/)

Protocol: a document that describes the background and rationale for a clinical trial with a detailed plan of how it will be conducted. (Source: http://www.ct-toolkit.ac.uk/routemap/protocol-development/)

Randomised controlled trial: a study in which people are randomly assigned to two or more groups to test a specific drug, treatment or intervention. One group (the experimental group) receives the intervention being tested and the other group (the control group) receives an alternative intervention. Groups are then follow up to see how they compare in response to predefined outcome(s). (Source: https://www.nice.org.uk/glossary?letter=r)

Safety Planning, Evaluation and Reporting Team (SPERT): group formed by the Pharmaceutical Research and Manufacturers of America to propose a pharmaceutical industry standard for safety planning, data collection, evaluation, and reporting, beginning with planning first in-human studies and continuing through the planning of the post-product approval period. (Source: Crowe, B. J., et al. (2009). "Recommendations for safety planning, data collection, evaluation and reporting during

drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team." Clinical Trials **6**(5): 430-440.)

Safety studies: Trials that aim to establish the absence of harm.

Serious Adverse Events (SAEs): used to describe a patient/event outcome or action criteria usually associated with events that pose a threat to a patient's life or functioning. Seriousness (not severity) serves as a guide for defining regulatory reporting obligations. (Source: The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Harmonised Tripartite Guideline. E2A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. 1994.)

Signal: will be used to refer to the information that detects or 'flags' the possibility of a causal relationship between the intervention and the harm outcome. Signals indicate that closer examination of an outcome is needed; this might involve closer examination of the event in ongoing studies or inform outcomes to prespecify in future studies including subsequent RCTs, systematic reviews or post-marketing research.

Spontaneous collection: refers to events that are reported without prompt, these are likely to be participant reports but can also be recorded by clinicians and trial staff

Statistical analysis plan: a document detailing the planned analysis for a clinical trial.

System organ class: groupings by etiology (e.g. *Infections and infestations*), manifestation site (e.g. *Gastrointestinal disorders*) or purpose (e.g. *Surgical and medical procedures*) within the MedDRA classification system. (Source: https://www.meddra.org/how-to-use/basics/hierarchy)

Type I error: the probability of rejecting a true null hypothesis. (Source: Gardner MJ Altman DG, editors. Statistics with Confidence. London: BMJ Publishing Group. Differences between means: type I and type II errors and power)

Type II error: the probability of failing to reject the null hypothesis when it is false. (Source: Gardner MJ Altman DG, editors. Statistics with Confidence. London: BMJ Publishing Group. Differences between means: type I and type II errors and power)

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH): a joint regulator and industry initiative that aims to harmonise drug development

UKCRC CTU Network: A UK network of academic clinical trials units (CTUs) who have been assessed by an international panel of experts in clinical trials research, who promote academic trials units and, through its activities, provides its members with information, guidance, and representation in support of the conduct of high-quality, effective, efficient, and sustainable clinical trials research. (Source: https://ukcrc-ctu.org.uk/)

Withdrawal from study: participants choose to discontinue study intervention and all study procedures

Withdrawal of treatment: participants or clinical team discontinue study intervention but participants remain in the study for follow-up and collection of data on outcomes

Yellow Card Scheme: a scheme run by the MHRA and is the UK system for collecting and monitoring information on safety concerns such as suspected side effects or adverse incidents involving medicines and medical devices. (Source: https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/)

References

1. Snodin DJ and Suitters A. Toxicology and Adverse Drug Reactions. In: Talbot J and Aronson JK (eds) *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice*. Sixth ed.: John Wiley and Sons, 2012, pp.157-214.
2. European Medicines Agency (EMA). Guideline on strategies to identify and mitigate risks for first-in-human and early clinical trials with investigational medicinal products, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-strategies-identify-mitigate-risks-first-human-early-clinical-trials-investigational_en.pdf (2017, accessed 07/05/2020).
3. Hauben M and Aronson JK. Gold Standards in Pharmacovigilance. *Drug Safety* 2007; 30: 645-655. DOI: 10.2165/00002018-200730080-00001.
4. Wheeler GM, Mander AP, Bedding A, et al. How to design a dose-finding study using the continual reassessment method. *BMC Medical Research Methodology* 2019; 19: 18. DOI: 10.1186/s12874-018-0638-z.
5. Thall PF and Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004; 60: 684-693. 2004/09/02. DOI: 10.1111/j.0006-341X.2004.00218.x.
6. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Topic E9 Statistical Principles for Clinical Trials. 1998.
7. Crowe BJ, Xia HA, Berlin JA, et al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. *Clinical Trials* 2009; 6: 430-440. DOI: 10.1177/1740774509344101.
8. Davis SK, B.; Raine, J.M. Spontaneous reporting – UK. In: Mann R AE (ed) *Pharmacovigilance*. 2nd ed. New York: Wiley, 2007, pp.199-215.
9. European Medicines Agency (EMA). EudraVigilance, http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000679.jsp&mid=WC0b01ac05800250b5 (accessed 24/07/2018 2018).
10. Glasser SP, Salas M and Delzell E. Importance and Challenges of Studying Marketed Drugs: What Is a Phase IV Study? Common Clinical Research Designs, Registries, and Self-Reporting Systems. *The Journal of Clinical Pharmacology* 2007; 47: 1074-1086. DOI: https://doi.org/10.1177/0091270007304776.
11. Suvarna V. Phase IV of Drug Development. *Perspectives in clinical research* 2010; 1: 57-60.
12. Talbot J, Keisu M and Ståhle L. Clinical Trials - Collecting Safety Data and Establishing the Adverse Drug Reactions Profile. In: Talbot J and Aronson JK (eds) *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice, Sixth Edition*. Sixth Edition ed.: John Wiley and Sons, 2012, pp.215-289.
13. Institute of Medicine (US) Committee on Conflict of Interest in Medical Research Education and Practice. E, The Pathway from Idea to Regulatory Approval: Examples for Drug Development. In: Lo B. and Field MJ (eds) *Conflict of Interest in Medical Research, Education, and Practice*. Washington (DC): National Academies Press, 2009.
14. O'Quigley J, Pepe M and Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 1990; 46: 33-48. 1990/03/01.
15. Braun TM. The current design of oncology phase I clinical trials: progressing from algorithms to statistical models. *Chin Clin Oncol* 2014; 3: 2. 2015/04/07. DOI: 10.3978/j.issn.2304-3865.2014.02.01.
16. Bate A and Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety* 2009; 18: 427-436. DOI: doi:10.1002/pds.1742.
17. European Medicines Agency. Pharmacovigilance: Overview, https://www.ema.europa.eu/en/human-regulatory/overview/pharmacovigilance-overview (accessed 28/10/2020).
18. Drug Safety Research Unit (DSRU). https://www.dsru.org/ (accessed 28/10/2020).
19. Uppsala Monitoring Centre. https://www.who-umc.org/ (accessed 28/10/2020).

20. Onakpoya IJ, Heneghan CJ and Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine* 2016; 14: 10. DOI: 10.1186/s12916-016-0553-2.
21. Goldman SA. Limitations and strengths of spontaneous reports data. *Clinical Therapeutics* 1998; 20: C40-C44. DOI: https://doi.org/10.1016/S0149-2918(98)80007-6.
22. Senn S. Safety Data, Harms, Drug Monitoring and Pharmaco-Epidemiology. *Statistical Issues in Drug Development*. Third ed.: Wiley Blackwell, 2021.
23. Ioannidis JA. Adverse events in randomized trials: Neglected, restricted, distorted, and silenced. *Archives of Internal Medicine* 2009; 169: 1737-1739. DOI: 10.1001/archinternmed.2009.313.
24. Singh S and Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials* 2012; 13. DOI: 10.1186/1745-6215-13-138.
25. Drago JZ, Gönen M, Thanarajasingam G, et al. Inferences About Drug Safety in Phase III Trials in Oncology: Examples From Advanced Prostate Cancer. *JNCI: Journal of the National Cancer Institute* 2020. DOI: 10.1093/jnci/djaa134.
26. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Topic E8 General Considerations for Clinical Trials. 2006.
27. Hazell L and Shakir SA. Under-reporting of adverse drug reactions : a systematic review. *Drug Saf* 2006; 29: 385-396. 2006/05/13. DOI: 10.2165/00002018-200629050-00003.
28. Ioannidis JA, Evans SW, Gøtzsche PC, et al. Better reporting of harms in randomized trials: An extension of the consort statement. *Annals of Internal Medicine* 2004; 141: 781-788. DOI: 10.7326/0003-4819-141-10-200411160-00009.
29. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Harmonised Tripartite Guideline. E2A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. 1994.
30. Edwards IR and Biriell C. Harmonisation in pharmacovigilance. *Drug safety* 1994; 10: 93-102. DOI: 10.2165/00002018-199410020-00001.
31. Foulkes MA. *Safety assessment versus efficacy assessment*. 2007, p.323-334.
32. Carpenter J and Kenward M. *Missing data in randomised controlled trials: a practical guide*. Birmingham: Health Technology Assessment Methodology Programme, 2007.
33. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials. 2019.
34. Lineberry N, Berlin JA, Mansi B, et al. Recommendations to improve adverse event reporting in clinical trial publications: A joint pharmaceutical industry/journal editor perspective. *BMJ (Online)* 2016; 355: i5078.
35. Zorzela L, Loke YK, Ioannidis JP, et al. PRISMA harms checklist: improving harms reporting in systematic reviews. *BMJ* 2016; 352. DOI: 10.1136/bmj.i157.
36. European Commission. Communication from the Commission — Detailed guidance on the collection, verification and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use ('CT-3'). *Official Journal of the European Union* 2011; C 172/1.
37. Food and Drug Administration. Guidance for Industry and Investigators. Safety Reporting Requirements for INDs and BA/BE studies. In: U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), (eds.). 2012.
38. Food and Drug Administration. Safety Assessment for IND Safety Reporting Guidance for Industry. In: U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), (eds.). 2015.
39. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Topic E3 Structure and Content of Clinical Study Reports. 1996.

40. Council for International Organizations of Medical Sciences (CIOMS) Working Group VI. *Management of safety information from clinical trials*. Geneva 2005.
41. Seltzer JH, Li J and Wang W. Interdisciplinary Safety Evaluation and Quantitative Safety Monitoring: Introduction to a Series of Papers. *Therapeutic Innovation & Regulatory Science* 2019; 0: 2168479018793130. DOI: 10.1177/2168479018793130.
42. Zink RC, Marchenko O, Sanchez-Kam M, et al. Sources of Safety Data and Statistical Strategies for Design and Analysis:Clinical Trials. *Therapeutic Innovation & Regulatory Science* 2018; 52: 141-158. DOI: 10.1177/2168479017738980.
43. Evans SJW and Nitsch D. Statistics: Analysis and Presentation of Safety Data. In: Talbot J and Aronson JK (eds) *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice*. Sixth Edition ed.: John Wiley and Sons, 2012, pp.349-388.
44. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. *Journal of Biopharmaceutical Statistics* 2009; 19: 889-899.
45. Phillips R, Hazell L, Sauzet O, et al. Analysis and reporting of adverse events in randomised controlled trials: a review. *BMJ Open* 2019; 9: e024537. 2019/03/04. DOI: 10.1136/bmjopen-2018-024537.
46. The Medicines for Human Use (Clinical Trials) Regulations. *SI 1031*. United Kingdom2004.
47. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008; 337: a1655. DOI: 10.1136/bmj.a1655.
48. Horigian VE, Robbins MS, Dominguez R, et al. Principles for defining adverse events in behavioral intervention research: lessons from a family-focused adolescent drug abuse trial. *Clinical Trials* 2010; 7: 58-68. DOI: 10.1177/1740774509356575.
49. Papaioannou D, Cooper C, Mooney C, et al. Adverse event recording failed to reflect potential harms: a review of trial protocols of behavioral, lifestyle and psychological therapy interventions. *Journal of Clinical Epidemiology* 2021; 136: 64-76. DOI: https://doi.org/10.1016/j.jclinepi.2021.03.002.
50. Health Research Authortiy (HRA). Safety Reporting, https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/safety-reporting/ (accessed 28/10/2020).
51. European Medicines Agency. *Guidance on data monitoring committees*. 2005. EMA.
52. Harrell F. Example Closed Meeting Data Monitoring Committee Report, https://biostatdata.app.vumc.org/fh/talks/RCTGraphics/greportEx2.pdf (2017, accessed 01/02/2021).
53. Harrell F. DSMB Report for EXAMPLE Trial, https://biostatdata.app.vumc.org/fh/talks/RCTGraphics/greportEx1.pdf (2017, accessed 01/02/2021).
54. Thomas SM, Jung K, Sun H, et al. Enhancing clarity of clinical trial safety reports for data monitoring committees. *Journal of Biopharmaceutical Statistics* 2020: 1-15. DOI: 10.1080/10543406.2020.1815034.
55. Aronson JK. Adverse Drug Reactions: History, Terminology, Classification, Causality, Frequency, Preventability. In: Talbot J and Aronson JK (eds) *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice*. Sixth ed.: John Wiley and Sons, 2012, pp.1-119.
56. Cornelius VR, Sauzet O, Williams JE, et al. Adverse event reporting in randomised controlled trials of neuropathic pain: Considerations for future practice. *PAIN* 2013; 154: 213-220. DOI: 10.1016/j.pain.2012.08.012.
57. Mayo-Wilson E, Fusco N, Hong H, et al. Opportunities for selective reporting of harms in randomized clinical trials: Selection criteria for non-systematic adverse events. *Trials* 2019; 20: 553. DOI: 10.1186/s13063-019-3581-3.
58. Favier R and Crépin S. The reporting of harms in publications on randomized controlled trials funded by the "Programme Hospitalier de Recherche Clinique," a French academic funding scheme. *Clinical Trials* 2018; 0: 1740774518760565. DOI: 10.1177/1740774518760565.

59. Ioannidis JA and Lau J. Completeness of safety reporting in randomized trials: An evaluation of 7 medical areas. *JAMA* 2001; 285: 437-443. DOI: 10.1001/jama.285.4.437.
60. Ma H, Ke C, Jiang Q, et al. Statistical Considerations on the Evaluation of Imbalances of Adverse Events in Randomized Clinical Trials. *Therapeutic Innovation and Regulatory Science* 2015; 49: 957-965. DOI: http://dx.doi.org/10.1177/2168479015587363.
61. Food and Drug Administration. *Attachment B: Clinical Safety Review of an NDA or BLA of the Good Review Practice. Clinical Reveiw Template (MAPP 6010.3 Rev. 1)*. 2010.
62. Unkel S, Amiri M, Benda N, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharmaceutical Statistics* 2019; 18: 166-183. DOI: 10.1002/pst.1915.
63. Brown EG and Harrison JE. Dictionaries and Coding in Pharmacovigilance. In: Talbot J and Aronson JK (eds) *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice*. Sixth Edition ed.: John Wiley and Sons, 2012, pp.545-572.
64. Brown EG, Wood L and Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf* 1999; 20: 109-117. 1999/03/19. DOI: 10.2165/00002018-199920020-00002.
65. Medical Dictionary for Regulatory Activities (MedDRA). MedDRA Maintenance and Support Services Organization Website, https://www.meddra.org/.
66. NIH National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE), https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm (accessed 27/10/2020).
67. Uppsala Monitoring Centre. What is WHO-ART?, https://www.who-umc.org/vigibase/services/learn-more-about-who-art/ (accessed 27/10/2020).
68. Zhang S, Liang F and Tannock I. Use and misuse of common terminology criteria for adverse events in cancer clinical trials. *BMC Cancer* 2016; 16: 392. DOI: 10.1186/s12885-016-2408-9.
69. Medical Dictionary for Regulatory Activities (MedDRA). MedDRA Hierarchy, https://www.meddra.org/how-to-use/basics/hierarchy (accessed 28/10/2020).
70. PROTECT Benefit-Risk Group. Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT), http://protectbenefitrisk.eu/index.html (accessed 27/10/2020).
71. Mt-Isa S, Hallgreen CE, Wang N, et al. Balancing benefit and risk of medicines: a systematic review and classification of available methodologies. *Pharmacoepidemiology and Drug Safety* 2014; 23: 667-678. DOI: 10.1002/pds.3636.
72. Hughes D, Waddingham E, Mt-Isa S, et al. Recommendations for benefit–risk assessment methodologies and visual representations. *Pharmacoepidemiology and Drug Safety* 2016; 25: 251-262. DOI: 10.1002/pds.3958.
73. Ioannidis JPA and Contopoulos-Ioannidis DG. Reporting of safety data from randomised trials. *The Lancet* 1998; 352: 1752-1753. DOI: 10.1016/S0140-6736(05)79825-1.
74. Edwards JE, McQuay HJ, Moore RA, et al. Reporting of Adverse Effects in Clinical Trials Should Be Improved. *Journal of Pain and Symptom Management* 1999; 18: 427-437. DOI: 10.1016/S0885-3924(99)00093-7.
75. Pitrou I, Boutron I, Ahmad N, et al. Reporting of safety results in published reports of randomized controlled trials. *Archives of Internal Medicine* 2009; 169: 1756-1761. DOI: 10.1001/archinternmed.2009.306.
76. Maggi CB, Griebeler IH and Dal Pizzol Tda S. Information on adverse events in randomised clinical trials assessing drug interventions published in four medical journals with high impact factors. *Int J Risk Saf Med* 2014; 26: 9-22. 2014/05/07. DOI: 10.3233/JRS-140609.
77. Smith SM, Wang AT, Katz NP, et al. Adverse event assessment, analysis, and reporting in recent published analgesic clinical trials: ACTTION systematic review and recommendations. *PAIN* 2013; 154: 997-1008. DOI: 10.1016/j.pain.2013.03.003.

78. Peron J, Maillet D, Gan HK, et al. Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. *J Clin Oncol* 2013; 31: 3957-3963. 2013/09/26. DOI: 10.1200/JCO.2013.49.3981.
79. Hum SW, Golder S and Shaikh N. Inadequate harms reporting in randomized control trials of antibiotics for pediatric acute otitis media: a systematic review. *Drug Safety* 2018 May 08. DOI: 10.1007/s40264-018-0680-0.
80. Junqueira D, Phillips R, Zorzela L, et al. Commentary: Time to improve the reporting of harms in randomized controlled trials. *Journal of Clinical Epidemiology* 2021.
81. Meeting abstracts from the 4th International Clinical Trials Methodology Conference (ICTMC) and the 38th Annual Meeting of the Society for Clinical Trials. *Trials* 2017; 18: 200. DOI: 10.1186/s13063-017-1902-y.
82. *Journal Citation Reports (Clarivate Analytics, 2018)*.
83. Schulz KF, Altman DG, Moher D, et al. Consort 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine* 2010; 152: 726-732. DOI: 10.7326/0003-4819-152-11-201006010-00232.
84. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC, 2017.
85. Stephens MD, Talbot JC and Routledge PA. *The Detection of New Adverse Reactions*. 4 ed. London: Macmillan Reference 1998.
86. Bent S, Padula A and Avins AL. Brief communication: Better ways to question patients about adverse medical events: a randomized, controlled trial. *Annals of Internal Medicine* 2006; 144: 257-261. DOI: 10.7326/0003-4819-144-4-200602210-00007.
87. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA* 2017; 318: 2337-2343. DOI: 10.1001/jama.2017.18556.
88. Tsang R, Colley L and Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of Clinical Epidemiology* 2009; 62: 609-616.
89. Litonjua AA, Carey VJ, Laranjo N, et al. Effect of Prenatal Supplementation With Vitamin D on Asthma or Recurrent Wheezing in Offspring by Age 3 Years: The VDAART Randomized Clinical Trial. *JAMA* 2016; 315: 362-370. 2016/01/28. DOI: 10.1001/jama.2015.18589.
90. Miller PD, Hattersley G, Riis BJ, et al. Effect of Abaloparatide vs Placebo on New Vertebral Fractures in Postmenopausal Women With Osteoporosis: A Randomized Clinical Trial. *JAMA* 2016; 316: 722-733. 2016/08/18. DOI: 10.1001/jama.2016.11136.
91. Libman IM, Miller KM, DiMeglio LA, et al. Effect of Metformin Added to Insulin on Glycemic Control Among Overweight/Obese Adolescents With Type 1 Diabetes: A Randomized Clinical Trial. *JAMA* 2015; 314: 2241-2250. 2015/12/02. DOI: 10.1001/jama.2015.16174.
92. Writing Committee for the Diabetic Retinopathy Clinical Research Network, Gross JG, Glassman AR, et al. Panretinal Photocoagulation vs Intravitreous Ranibizumab for Proliferative Diabetic Retinopathy: A Randomized Clinical Trial. *JAMA* 2015; 314: 2137-2146. 2015/11/14. DOI: 10.1001/jama.2015.15217.
93. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *N Engl J Med* 2016; 375: 311-322. 2016/06/14. DOI: 10.1056/NEJMoa1603827.
94. Beardsley J, Wolbers M, Kibengo FM, et al. Adjunctive Dexamethasone in HIV-Associated Cryptococcal Meningitis. *N Engl J Med* 2016; 374: 542-554. 2016/02/11. DOI: 10.1056/NEJMoa1509024.
95. Myles PS, Smith JA, Forbes A, et al. Stopping vs. Continuing Aspirin before Coronary Artery Surgery. *N Engl J Med* 2016; 374: 728-737. 2016/03/05. DOI: 10.1056/NEJMoa1507688.
96. Nichol A, French C, Little L, et al. Erythropoietin in traumatic brain injury (EPO-TBI): a double-blind randomised controlled trial. *The Lancet* 2015; 386: 2499-2506. DOI: 10.1016/s0140-6736(15)00386-4.

97. Billings FTt, Hendricks PA, Schildcrout JS, et al. High-Dose Perioperative Atorvastatin and Acute Kidney Injury Following Cardiac Surgery: A Randomized Clinical Trial. *JAMA* 2016; 315: 877-888. 2016/02/26. DOI: 10.1001/jama.2016.0548.
98. Kor DJ, Carter RE, Park PK, et al. Effect of Aspirin on Development of ARDS in At-Risk Patients Presenting to the Emergency Department: The LIPS-A Randomized Clinical Trial. *JAMA* 2016; 315: 2406-2414. 2016/05/18. DOI: 10.1001/jama.2016.6330.
99. Aitken E, Jackson A, Kearns R, et al. Effect of regional versus local anaesthesia on outcome after arteriovenous fistula creation: a randomised controlled trial. *The Lancet* 2016; 388: 1067-1074. DOI: 10.1016/s0140-6736(16)30948-5.
100. Kaul U, Bangalore S, Seth A, et al. Paclitaxel-Eluting versus Everolimus-Eluting Coronary Stents in Diabetes. *N Engl J Med* 2015; 373: 1709-1719. 2015/10/16. DOI: 10.1056/NEJMoa1510188.
101. Cohen MS, Chen YQ, McCauley M, et al. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. *N Engl J Med* 2016; 375: 830-839. 2016/07/19. DOI: 10.1056/NEJMoa1600693.
102. Isanaka S, Langendorf C, Berthe F, et al. Routine Amoxicillin for Uncomplicated Severe Acute Malnutrition in Children. *N Engl J Med* 2016; 374: 444-453. 2016/02/04. DOI: 10.1056/NEJMoa1507024.
103. Natalucci G, Latal B, Koller B, et al. Effect of Early Prophylactic High-Dose Recombinant Human Erythropoietin in Very Preterm Infants on Neurodevelopmental Outcome at 2 Years: A Randomized Clinical Trial. *JAMA* 2016; 315: 2079-2085. 2016/05/18. DOI: 10.1001/jama.2016.5504.
104. Sheehan WJ, Mauger DT, Paul IM, et al. Acetaminophen versus Ibuprofen in Young Children with Mild Persistent Asthma. *N Engl J Med* 2016; 375: 619-630. 2016/08/18. DOI: 10.1056/NEJMoa1515990.
105. Burger JA, Tedeschi A, Barr PM, et al. Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia. *N Engl J Med* 2015; 373: 2425-2437. 2015/12/08. DOI: 10.1056/NEJMoa1509388.
106. McInnes IB, Mease PJ, Kirkham B, et al. Secukinumab, a human anti-interleukin-17A monoclonal antibody, in patients with psoriatic arthritis (FUTURE 2): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet* 2015; 386: 1137-1146. DOI: 10.1016/s0140-6736(15)61134-5.
107. Herbst RS, Baas P, Kim D-W, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *The Lancet* 2016; 387: 1540-1550. DOI: 10.1016/s0140-6736(15)01281-7.
108. Lincoff AM, Mehran R, Povsic TJ, et al. Effect of the REG1 anticoagulation system versus bivalirudin on outcomes after percutaneous coronary intervention (REGULATE-PCI): a randomised clinical trial. *The Lancet* 2016; 387: 349-356. DOI: 10.1016/s0140-6736(15)00515-2.
109. Ribrag V, Koscielny S, Bosq J, et al. Rituximab and dose-dense chemotherapy for adults with Burkitt's lymphoma: a randomised, controlled, open-label, phase 3 trial. *The Lancet* 2016; 387: 2402-2411. DOI: 10.1016/s0140-6736(15)01317-3.
110. Nguyen QD, Merrill PT, Jaffe GJ, et al. Adalimumab for prevention of uveitic flare in patients with inactive non-infectious uveitis controlled by corticosteroids (VISUAL II): a multicentre, double-masked, randomised, placebo-controlled phase 3 trial. *The Lancet* 2016; 388: 1183-1192. DOI: 10.1016/s0140-6736(16)31339-3.
111. Rodes-Cabau J, Horlick E, Ibrahim R, et al. Effect of Clopidogrel and Aspirin vs Aspirin Alone on Migraine Headaches After Transcatheter Atrial Septal Defect Closure: The CANOA Randomized Clinical Trial. *JAMA* 2015; 314: 2147-2154. 2015/11/10. DOI: 10.1001/jama.2015.13919.
112. Altman DG and Royston P. The cost of dichotomising continuous variables. *BMJ* 2006; 332: 1080. DOI: 10.1136/bmj.332.7549.1080.
113. Ruperto N, Pistorio A, Oliveira S, et al. Prednisone versus prednisone plus ciclosporin versus prednisone plus methotrexate in new-onset juvenile dermatomyositis: a randomised trial. *The Lancet* 2016; 387: 671-678. DOI: 10.1016/s0140-6736(15)01021-1.

114. Whitehead KJ, Sautter NB, McWilliams JP, et al. Effect of Topical Intranasal Therapy on Epistaxis Frequency in Patients With Hereditary Hemorrhagic Telangiectasia: A Randomized Clinical Trial. *JAMA* 2016; 316: 943-951. 2016/09/07. DOI: 10.1001/jama.2016.11724.
115. Shamseer L, Hopewell S, Altman DG, et al. Update on the endorsement of CONSORT by high impact factor journals: a survey of journal "Instructions to Authors" in 2014. *Trials* 2016; 17: 301. DOI: 10.1186/s13063-016-1408-z.
116. CONSORT. Impact of CONSORT, http://www.consort-statement.org/about-consort/impact-of-consort (accessed 28/10/2020).
117. Haidich AB, Birtsou C, Dardavessis T, et al. The quality of safety reporting in trials is still suboptimal: survey of major general medical journals. *J Clin Epidemiol* 2011; 64: 124-135. 2010/12/22. DOI: 10.1016/j.jclinepi.2010.03.005.
118. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Harmonised Tripartite Guideline Structure and Content of Clinical Study Reports E3. 1996.
119. Sacks CA, Miller PW and Longo DL. Talking about Toxicity — "What We've Got Here Is a Failure to Communicate". *New England Journal of Medicine* 2019; 381: 1406-1408. DOI: 10.1056/NEJMp1908310.
120. Altman DG and Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 1995; 311: 485. DOI: 10.1136/bmj.311.7003.485.
121. Jonville-Béra A, Giraudeau B and Autret-Leca E. Reporting of drug tolerance in randomized clinical trials: When data conflict with authors' conclusions. *Annals of Internal Medicine* 2006; 144: 306-307. DOI: 10.7326/0003-4819-144-4-200602210-00024.
122. Detry MA and Lewis RJ. The intention-to-treat principle: How to assess the true effect of choosing a medical treatment. *JAMA* 2014; 312: 85-86. DOI: 10.1001/jama.2014.7523.
123. Patson N, Mukaka M, Otwombe KN, et al. Systematic review of statistical methods for safety data in malaria chemoprevention in pregnancy trials. *Malaria Journal* 2020; 19: 119. DOI: 10.1186/s12936-020-03190-z.
124. Cornelius VR, Sauzet O and Evans SJW. A Signal Detection Method to Detect Adverse Drug Reactions Using a Parametric Time-to-Event Model in Simulated Cohort Data. *Drug Safety* 2012; 35: 599-610. journal article. DOI: 10.2165/11599740-000000000-00000.
125. Cohen J. The Cost of Dichotomization. *Applied Psychological Measurement* 1983; 7: 249-253. DOI: 10.1177/014662168300700301.
126. Austin PC and Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 2004; 23: 1159-1178. DOI: 10.1002/sim.1687.
127. Verbaanderd C, Rooman I and Huys I. Exploring new uses for existing drugs: innovative mechanisms to fund independent clinical research. *Trials* 2021; 22: 322. DOI: 10.1186/s13063-021-05273-x.
128. Food and Drug Administration. The Drug Development Process, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process (2018).
129. Senn S. *Statistical Issues in Drug Development*. Third ed.: Wiley Blackwell, 2021.
130. Phillips R, Sauzet O and Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. *BMC Medical Research Methodology* 2020; 20: 288. DOI: 10.1186/s12874-020-01167-9.
131. Arksey H and O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 2005; 8: 19-32. DOI: 10.1080/1364557032000119616.
132. O'Brien PC and Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics* 1979; 35: 549-556. DOI: 10.2307/2530245.

133. DeMets DL and Lan G. The alpha spending function approach to interim data analyses. In: Thall PF (ed) *Recent Advances in Clinical Trial Design and Analysis*. Boston, MA: Springer US, 1995, pp.1-27.
134. Tricco AC, Lillie E, Zarin W, et al. Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Annals of Internal Medicine* 2018. DOI: 10.7326/M18-0850.
135. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine* 2009; 6: e1000097. DOI: 10.1371/journal.pmed.1000097.
136. Phillips R, Cornelius V and Sauzet O. An overview of statistical methods developed to analyse adverse events in clinical trials: protocol for a methodological review, https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=97442 (2018).
137. Amit O, Heiberger RM and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 2008; 7: 20-35.
138. Chuang-Stein C, Le V and Chen W. Recent Advancements in the Analysis and Presentation of Safety Data. *Drug Information Journal* 2001; 35: 377-397. DOI: 10.1177/009286150103500207.
139. Chuang-Stein C and Xia HA. The practice of pre-marketing safety assessment in drug development. *Journal of Biopharmaceutical Statistics* 2013; 23: 3-25. Review. DOI: 10.1080/10543406.2013.736805.
140. Karpefors M and Weatherall J. The Tendril Plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials. *Journal of the American Medical Informatics Association* 2018; 25: 1069-1073. DOI: 10.1093/jamia/ocy016.
141. Southworth H. Detecting outliers in multivariate laboratory data. *Journal of Biopharmaceutical Statistics* 2008; 18: 1178-1183.
142. Trost DC and Freston JW. Vector Analysis to Detect Hepatotoxicity Signals in Drug Development. *Therapeutic Innovation & Regulatory Science* 2008; 42: 27-34. Article. DOI: 10.1177/009286150804200106.
143. Zink RC, Wolfinger RD and Mann G. Summarizing the incidence of adverse events using volcano plots and time intervals. *Clinical Trials* 2013; 10: 398-406.
144. Bolland K and Whitehead J. Formal approaches to safety monitoring of clinical trials in life-threatening conditions. *Statistics in Medicine* 2000; 19: 2899-2917. Article. DOI: 10.1002/1097-0258(20001115)19:21<2899::AID-SIM597>3.0.CO;2-O.
145. Fleishman AN and Parker RA. Stopping guidelines for harm in a study designed to establish the safety of a marketed drug. *Journal of Biopharmaceutical Statistics* 2012; 22: 338-350. Article. DOI: 10.1080/10543406.2010.536872.
146. Lieu TA, Kulldorff M, Davis RL, et al. Real-Time Vaccine Safety Surveillance for the Early Detection of Adverse Events. *Medical Care* 2007; 45: S89-S95. DOI: 10.1097/MLR.0b013e3180616c0a.
147. Liu JP. Rethinking statistical approaches to evaluating drug safety. *Yonsei Medical Journal* 2007; 48: 895-900. Review. DOI: 10.3349/ymj.2007.48.6.895.
148. Shih MC, Lai TL, Heyse JF, et al. Sequential generalized likelihood ratio tests for vaccine safety evaluation. *Statistics in Medicine* 2010; 29: 2698-2708.
149. Agresti AaK, B. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Appl Statist* 2005; 54: 691-706.
150. Bristol DR and Patel HI. A Markovian model for comparing incidences of side effects. *Statistics in Medicine* 1990; 9: 803-809.
151. Chuang-Stein C, Mohberg NR and Musselman DM. Organization and analysis of safety data using a multivariate approach. *Statistics in Medicine* 1992; 11: 1075-1089. DOI: doi:10.1002/sim.4780110809.
152. Huang L, Zalkikar J and Tiwari R. Likelihood ratio based tests for longitudinal drug safety data. *Statistics in Medicine* 2014; 33: 2408-2424. Article. DOI: 10.1002/sim.6103.

153. Mehrotra DV and Adewale AJ. Flagging clinical adverse experiences: Reducing false discoveries without materially compromising power for detecting true signals. *Statistics in Medicine* 2012; 31: 1918-1930.
154. Mehrotra DV and Heyse JF. Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research* 2004; 13: 227-238. Article.
155. Allignol A, Beyersmann J and Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics* 2016; 15: 297-305.
156. Borkowf CB. Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. *Statistics in Medicine* 2006; 25: 3679-3695.
157. Gong Q, Tong B, Strasak A, et al. Analysis of safety data in clinical trials using a recurrent event approach. *Pharmaceutical Statistics* 2014; 13: 136-144. DOI: doi:10.1002/pst.1611.
158. Hengelbrock J, Gillhaus J, Kloss S, et al. Safety data from randomized controlled trials: applying models for recurrent events. *Pharmaceutical Statistics* 2016; 15: 315-323. Conference Paper. DOI: http://dx.doi.org/10.1002/pst.1757.
159. Lancar R, Kramar A and Haie-Meder C. Non-parametric methods for analysing recurrent complications of varying severity. *Statistics in Medicine* 1995; 14: 2701-2712. Clinical Trial

Randomized Controlled Trial.
160. Leon-Novelo LG, Zhou X, Bekele BN, et al. Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. *Biometrics* 2010; 66: 966-974.
161. Frank LG, Junyuan W, Kenneth L, et al. Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in Medicine* 2006; 25: 1275-1286. DOI: doi:10.1002/sim.2335.
162. Nishikawa M, Tango T and Ogawa M. Non-parametric inference of adverse events under informative censoring. *Statistics in Medicine* 2006; 25: 3981-4003.
163. O'Gorman TW, Woolson RF and Jones MP. A comparison of two methods of estimating a common risk difference in a stratified analysis of a multicenter clinical trial. *Controlled Clinical Trials* 1994; 15: 135-153.
164. Rosenkranz G. Analysis of adverse events in the presence of discontinuations. *Drug Information Journal* 2006; 40: 79-87. DOI: 10.1177/009286150604000110.
165. Sogliero-Gilbert G, Ting, N. and Zubkoff, L. . A statistical comparison of drug safety in controlled clinical trials: The Genie score as an objective measure of lab abnormalities. *Therapeutic Innovation & Regulatory Science* 1991; 25. DOI: https://doi.org/10.1177/009286159102500109.
166. Wang J and Quartey G. Nonparametric estimation for cumulative duration of adverse events. *Biometrical Journal* 2012; 54: 61-74.
167. Wang J and Quartey G. A semi-parametric approach to analysis of event duration and prevalence. *Computational Statistics & Data Analysis* 2013; 67: 248-257. DOI: https://doi.org/10.1016/j.csda.2013.05.023.
168. Berry DA. Monitoring accumulating data in a clinical trial. *Biometrics* 1989; 45: 1197-1211. Article.
169. French JL, Thomas N and Wang C. Using historical data with Bayesian methods in early clinical trial monitoring. *Statistics in Biopharmaceutical Research* 2012; 4: 384-394. DOI: http://dx.doi.org/10.1080/19466315.2012.707088.
170. Yao B, Zhu L, Jiang Q, et al. Safety monitoring in clinical trials. *Pharmaceutics* 2013; 5: 94-106. 2013/12/05. DOI: 10.3390/pharmaceutics5010094.
171. Zhu L, Yao B, Xia HA, et al. Statistical Monitoring of Safety in Clinical Trials. *Statistics in Biopharmaceutical Research* 2016; 8: 88-105. DOI: 10.1080/19466315.2015.1117017.
172. Berry SM and Berry DA. Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics* 2004; 60: 418-426. Article. DOI: 10.1111/j.0006-341X.2004.00186.x.

173. Chen W, Zhao N, Qin G, et al. A bayesian group sequential approach to safety signal detection. *Journal of Biopharmaceutical Statistics* 2013; 23: 213-230. Article. DOI: 10.1080/10543406.2013.736813.
174. Gould AL. Detecting potential safety issues in clinical trials by Bayesian screening. *Biometrical Journal* 2008; 50: 837-851. Article. DOI: 10.1002/bimj.200710469.
175. Gould AL. Detecting potential safety issues in large clinical or observational trials by bayesian screening when event counts arise from poisson distributions. *Journal of Biopharmaceutical Statistics* 2013; 23: 829-847. Article. DOI: 10.1080/10543406.2013.789887.
176. McEvoy BW, Nandy RR and Tiwari RC. Bayesian Approach for Clinical Trial Safety Data Using an Ising Prior. *Biometrics* 2013; 69: 661-672.
177. Xia HA, Ma H and Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; 21: 1006-1029.
178. Whone A, Luz M, Boca M, et al. Randomized trial of intermittent intraputamenal glial cell line-derived neurotrophic factor in Parkinson's disease. *Brain* 2019; 142: 512-525. DOI: 10.1093/brain/awz023.
179. Trost DC, Overman EA, Ostroff JH, et al. A model for liver homeostasis using modified mean-reverting Ornstein-Uhlenbeck process. *Computational and Mathematical Methods in Medicine* 2010; 11: 27-47.
180. Du Toit G, Roberts G, Sayre PH, et al. Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy. *New England Journal of Medicine* 2015; 372: 803-813. DOI: 10.1056/NEJMoa1414850.
181. Lineberry N, Berlin JA, Mansi B, et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective.[Erratum appears in BMJ. 2017 Mar 8;356:j1228; PMID: 28274948]. *BMJ* 355: i5078.
182. Xia HA, Crowe BJ, Schriver RC, et al. Planning and core analyses for periodic aggregate safety data reviews. *Clinical Trials* 2011; 8: 175-182. DOI: 10.1177/1740774510395635.
183. Cooper AJP, Lettis S, Chapman CL, et al. Developing tools for the safety specification in risk management plans: lessons learned from a pilot project. *Pharmacoepidemiology and Drug Safety* 2008; 17: 445-454. DOI: doi:10.1002/pds.1576.
184. Lewis S and Clarke M. Forest plots: trying to see the wood and the trees. *BMJ* 2001; 322: 1479-1480. DOI: 10.1136/bmj.322.7300.1479.
185. Lieu TA, Kulldorff M, Davis RL, et al. Real-time vaccine safety surveillance for the early detection of adverse events. *Medical Care* 2007; 45: S89-S95. Article.
186. Whitehead J. Sequential Methods for Clinical Trials. *Wiley StatsRef: Statistics Reference Online*. 2014.
187. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995; 57: 289-300.
188. Diaz-Mitoma F, Halperin SA, Tapiero B, et al. Safety and immunogenicity of three different formulations of a liquid hexavalent diphtheria–tetanus–acellular pertussis–inactivated poliovirus–Haemophilus influenzae b conjugate–hepatitis B vaccine at 2, 4, 6 and 12–14 months of age. *Vaccine* 2011; 29: 1324-1331. DOI: https://doi.org/10.1016/j.vaccine.2010.11.053.
189. Group MV-S, Priddy FH, Novak RM, et al. Safety and Immunogenicity of a Replication-Incompetent Adenovirus Type 5 HIV-1 Clade B gag/pol/nef Vaccine in Healthy Adults. *Clinical Infectious Diseases* 2008; 46: 1769-1781. DOI: 10.1086/587993.
190. Phillips R and Cro S. AEFDR: Stata module to perform false discovery rate p-value adjustment for adverse event data. *Statistical Software Components S458733*. Boston College Department of Economics, 2020.
191. Liu GF, Wang J, Liu K, et al. Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in Medicine* 2006; 25: 1275-1286. DOI: doi:10.1002/sim.2335.

192. Andersen PK and Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann Statist* 1982; 10: 1100-1120. DOI: 10.1214/aos/1176345976.
193. Prentice RL, Williams BJ and Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981; 68: 373-379. DOI: 10.1093/biomet/68.2.373.
194. Fine JP and Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999; 94: 496-509. DOI: 10.1080/01621459.1999.10474144.
195. Cabarrou B, Longué M, Filleron T, et al. The importance of jointly analyzing treatment administration and toxicity associated with targeted therapies: a case study of regorafenib in soft tissue sarcoma patients. *Annals of Oncology* 2018; 29: 1588-1593. DOI: 10.1093/annonc/mdy168.
196. Christie D, Denham J, Steigler A, et al. Delayed rectal and urinary symptomatology in patients treated for prostate cancer by radiotherapy with or without short term neo-adjuvant androgen deprivation. *Radiotherapy and Oncology* 2005; 77: 117-125. DOI: https://doi.org/10.1016/j.radonc.2005.10.005.
197. Proctor T and Schumacher M. Analysing adverse events by time-to-event models: the CLEOPATRA study. *Pharmaceutical Statistics* 2016; 15: 306-314. DOI: 10.1002/pst.1758.
198. Tsuboi A, Myoui A, Sugiyama H, et al. A Phase I/II Trial of a WT1 (Wilms' Tumor Gene) Peptide Vaccine in Patients with Solid Malignancy: Safety Assessment Based on the Phase I Data. *Japanese Journal of Clinical Oncology* 2006; 36: 231-236. DOI: 10.1093/jjco/hyl005.
199. O'Brien S, Rizzieri DA, Vey N, et al. Elacytarabine has single-agent activity in patients with advanced acute myeloid leukaemia. *British Journal of Haematology* 2012; 158: 581-588. DOI: doi:10.1111/j.1365-2141.2012.09186.x.
200. Spiegelhalter DJ, Abrams KR and Myles JP. An Overview of the Bayesian Approach. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, 2004.
201. Gelman A, Hill J and Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons *Journal of Research on Educational Effectiveness* 2012; 5: 189-211. DOI: 10.1080/19345747.2011.618213.
202. Southworth H and O'Connell M. Data Mining and Statistically Guided Clinical Review of Adverse Event Data in Clinical Trials. *Journal of Biopharmaceutical Statistics* 2009; 19: 803-817. DOI: 10.1080/10543400903105232.
203. Gould AL. Unified screening for potential elevated adverse event risk and other associations. *Statistics in Medicine* 2018; 37: 2667-2689. DOI: doi:10.1002/sim.7686.
204. Phillips R and Cro S. AEDOT: Stata module to produce dot plot for adverse event data. *Statistical Software Components S458735*. Boston College Department of Economics, 2020.
205. Phillips R and Cro S. AEVOLCANO: Stata module to produce volcano plot for adverse event data. *Statistical Software Components S458736*. Boston College Department of Economics, 2020.
206. Cornelius V, Cro S and Phillips R. Visualisations to evaluate and communicate adverse event information in Randomised Controlled Trials. *Under review* 2020.
207. Wang W, Whalen E, Munsaka M, et al. On Quantitative Methods for Clinical Safety Monitoring in Drug Development. *Statistics in Biopharmaceutical Research* 2018; 10: 85-97. DOI: 10.1080/19466315.2017.1409134.
208. Phillips R, Sauzet O and Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a review of available methods. *(Unpublished)*.
209. Phillips R, Cornelius V and Sauzet O. An evaluation and application of statistical methods designed to analyse adverse event data in RCTs. *Trials Conference: 5th International Clinical Trials Methodology Conference, ICTMC 2019 United Kingdom* 2019; 20.
210. Food and Drug Administration. *Guidance for industry e9 statistical principles for clinical trials*. 1998. Food and Drug Administration.
211. Harrell F. Continuous Learning from Data: No Multiplicities from Computing and Using Bayesian Posterior Probabilities as Often as Desired. Statistical Thinking2018.

212. J A. Is power analysis necessary in Bayesian Statistics?, https://stats.stackexchange.com/questions/65754/is-power-analysis-necessary-in-bayesian-statistics.
213. Spiegelhalter DJ. Incorporating Bayesian Ideas into Health-Care Evaluation. *Statistical Science* 2004; 19: 156-174.
214. Ball G. Continuous safety monitoring for randomized controlled clinical trials with blinded treatment information Part 4: One method. *Contemporary Clinical Trials* 2011; 32: S11-S17. Article. DOI: 10.1016/j.cct.2011.05.008.
215. Gould AL and Wang WB. Monitoring potential adverse event rate differences using data from blinded trials: the canary in the coal mine. *Statistics in Medicine* 2017; 36: 92-104. DOI: http://dx.doi.org/10.1002/sim.7129.
216. Phillips R and Cornelius V. Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry. *BMJ Open* 2020; 10: e036875. DOI: 10.1136/bmjopen-2020-036875.
217. Love SB, Brown S, Weir CJ, et al. Embracing model-based designs for dose-finding trials. *British journal of cancer* 2017; 117: 332-339. 06/29. DOI: 10.1038/bjc.2017.186.
218. Dimairo M, Julious SA, Todd S, et al. Cross-sector surveys assessing perceptions of key stakeholders towards barriers, concerns and facilitators to the appropriate use of adaptive designs in confirmatory trials. *Trials* 2015; 16: 585-585. DOI: 10.1186/s13063-015-1119-x.
219. Kelley K, Clark B, Brown V, et al. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care* 2003; 15: 261-266. DOI: 10.1093/intqhc/mzg031.
220. Jaki T. Uptake of novel statistical methods for early-phase clinical studies in the UK public sector. *Clinical Trials* 2013; 10: 344-346.
221. Rhodes C, Hutton G and Ward M. Research and Development spending. In: SN04223 HoCLBPn, (ed.). London: House of Commons Library, 2020.
222. The Association of the British Pharmaceutical Industry. *Clinical trials - How the UK is researching medicines of the future*. 2019.
223. Special Issue:Analysis of Adverse Event Data. *Pharmaceutical Statistics* 2016; 15: 287-379.
224. Colopy MW, Gordon R, Ahmad F, et al. Statistical Practices of Safety Monitoring: An Industry Survey. *Therapeutic Innovation & Regulatory Science* 2019; 53: 293-300. DOI: 10.1177/2168479018779973.
225. Davis S, Sun H and Jung K. Best practices for reporting safety data to data monitoring committees. *Trials Conference: 4th International Clinical Trials Methodology Conference , ICTMC and the 38th Annual Meeting of the Society for Clinical Trials United Kingdom* 2017; 18.
226. Furey A and Bechhofer R. Effective graphical analyses of adverse events in DMC reports. *Trials Conference: 4th International Clinical Trials Methodology Conference , ICTMC and the 38th Annual Meeting of the Society for Clinical Trials United Kingdom* 2017; 18.
227. Vandemeulebroecke M, Baillie M, Carr D, et al. How can we make better graphs? An initiative to increase the graphical expertise and productivity of quantitative scientists. *Pharmaceutical Statistics* 2019; 18: 106-114. DOI: 10.1002/pst.1912.
228. Wang W, Revis R, Nilsson M, et al. Clinical Trial Drug Safety Assessment With Interactive Visual Analytics. *Statistics in Biopharmaceutical Research* 2020: 1-12. DOI: 10.1080/19466315.2020.1736142.
229. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38: 2074-2102. DOI: https://doi.org/10.1002/sim.8086.
230. Cro S, Morris TP, Kenward MG, et al. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine* 2020; 39: 2815-2842. DOI: https://doi.org/10.1002/sim.8569.

231. Duke SP, Bancken F, Crowe B, et al. Seeing is believing: good graphic design principles for medical research. *Statistics in Medicine* 2015; 34: 3040-3059. DOI: doi:10.1002/sim.6549.
232. Tufte ER. *The Visual Display of Quantitative Information*. 2nd ed. Chesapeake, CT: Graphics Press, 2001.
233. Harrell FE. *Principles of Graph Construction*.
234. Lineberry N, Berlin JA, Mansi B, et al. Recommendations to improve adverse event reporting in clinical trial publications: A joint pharmaceutical industry/journal editor perspective. *BMJ (Online)* 2016; 355 (no pagination).
235. Gordon I and Finch S. Statistician Heal Thyself: Have We Lost the Plot? *Journal of Computational and Graphical Statistics* 2015; 24: 1210-1229. DOI: 10.1080/10618600.2014.989324.
236. Gelman A, Pasarica C and Dodhia R. Let's Practice What We Preach. *The American Statistician* 2002; 56: 121-130. DOI: 10.1198/000313002317572790.
237. Unwin A. Why is Data Visualization Important? What is Important in Data Visualization? *Harvard Data Science Review* 2020; 2. DOI: 10.1162/99608f92.8ae4d525.
238. Black N, Murphy M, Lamping D, et al. Consensus Development Methods: A Review of Best Practice in Creating Clinical Guidelines. *Journal of Health Services Research & Policy* 1999; 4: 236-248. DOI: 10.1177/135581969900400410.
239. Moher D, Schulz KF, Simera I, et al. Guidance for Developers of Health Research Reporting Guidelines. *PLOS Medicine* 2010; 7: e1000217. DOI: 10.1371/journal.pmed.1000217.
240. Ballarini NM, Chiu Y-D, König F, et al. A critical review of graphics for subgroup analyses in clinical trials. *Pharmaceutical Statistics* 2020 25 March 2020. DOI: 10.1002/pst.2012.
241. Thanarajasingam G, Atherton PJ, Novotny PJ, et al. Longitudinal adverse event assessment in oncology clinical trials: The Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. *The Lancet Oncology* 2016; 17: 663-670. DOI: http://dx.doi.org/10.1016/S1470-2045%2816%2900038-3.
242. Thanarajasingam G, Atherton PJ, Pederson L, et al. Beyond maximum grade: A novel method to assess toxicity over time in clinical trials of targeted therapy in lymphoma. *Journal of Clinical Oncology Conference* 2016; 34.
243. Newell J, Kay JW and Aitchison TC. Survival ratio plots with permutation envelopes in survival data problems. *Computers in Biology and Medicine* 2006; 36: 526-541. DOI: https://doi.org/10.1016/j.compbiomed.2005.03.005.
244. Bel EH, Wenzel SE, Thompson PJ, et al. Oral Glucocorticoid-Sparing Effect of Mepolizumab in Eosinophilic Asthma. *New England Journal of Medicine* 2014; 371: 1189-1197. DOI: 10.1056/NEJMoa1403291.
245. Berard R, Fong R, Carpenter DJ, et al. An international, multicenter, placebo-controlled trial of paroxetine in adolescents with major depressive disorder. *J Child Adolesc Psychopharmacol* 2006; 16: 59-75. 2006/03/24. DOI: 10.1089/cap.2006.16.59.
246. Ortega HG, Liu MC, Pavord ID, et al. Mepolizumab Treatment in Patients with Severe Eosinophilic Asthma. *New England Journal of Medicine* 2014; 371: 1198-1207. DOI: 10.1056/NEJMoa1403290.
247. Emslie GJ, Wagner KD, Kutcher S, et al. Paroxetine Treatment in Children and Adolescents With Major Depressive Disorder: A Randomized, Multicenter, Double-Blind, Placebo-Controlled Trial. *Journal of the American Academy of Child & Adolescent Psychiatry* 2006; 45: 709-719. DOI: https://doi.org/10.1097/01.chi.0000214189.73240.63.
248. Singh S and Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials* 2012; 13: 138. 2012/08/22. DOI: 10.1186/1745-6215-13-138.
249. Phillips R and Cro S. Stata module to produce dot plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458735.html2020.
250. CTSPEDIA: Adverse Events Clinical Questions Addressed, https://www.ctspedia.org/do/view/CTSpedia/ListingsAdverseEventsVetted (2014, accessed 08/03/2021).

251. Proctor T and Schumacher M. Analysing adverse events by time-to-event models: The CLEOPATRA study. *Pharmaceutical Statistics* 2016.
252. Rogers JK, Pocock SJ, McMurray JJV, et al. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *European journal of heart failure* 2014; 16: 33-40. 2013/12/18. DOI: 10.1002/ejhf.29.
253. Morris TP, Jarvis CI, Cragg W, et al. Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open* 2019; 9: e030215. DOI: 10.1136/bmjopen-2019-030215.
254. Rost LC. Thoughts & How To's How to pick more beautiful colors for your data visualizations. Chartable2020.
255. Vandemeulebroecke M, Baillie M, Margolskee A, et al. Effective Visual Communication for the Quantitative Scientist. *CPT: Pharmacometrics & Systems Pharmacology* 2019; 8: 705-719. DOI: https://doi.org/10.1002/psp4.12455.
256. Vickers AJ, Assel MJ, Sjoberg DD, et al. Guidelines for Reporting of Figures and Tables for Clinical Research in Urology. *European Urology* 2020; 78: 97-109. DOI: 10.1016/j.eururo.2020.04.048.
257. Adobe Color, https://color.adobe.com/create/color-wheel (accessed 11/03/2021).
258. Jenny B. https://colororacle.org/ (accessed 11/03/2021).
259. Cornelius V, Cro S and Phillips R. Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials. *Trials* 2020; 21: 1028. DOI: 10.1186/s13063-020-04903-0.
260. Cleveland WS. Graphs in Scientific Publications. *The American Statistician* 1984; 38: 261-269. DOI: 10.2307/2683400.
261. Clinical Trials Statistical Data Analysis Center DoBaMI, University of Wisconsin Madison. Sample Closed Session DMC Report, https://www.biostat.wisc.edu/sites/default/files/Sample_Report_Closed_20160920.pdf (2016, accessed 26/03/2018 2018).
262. Meyboom RH, Egberts AC, Edwards IR, et al. Principles of signal detection in pharmacovigilance. *Drug Saf* 1997; 16: 355-365. 1997/06/01.
263. Agency EM. ICH Topic E9 Statistical Principles for Clinical Trials. 1998.
264. Cleves C, Gould WW and Marchenko YV. *An introduction to survival analysis using Stata*. Revised third edition ed.: Stata Press, 2016.
265. Sauzet O, Carvajal A, Escudero A, et al. Illustration of the Weibull Shape Parameter Signal Detection Tool Using Electronic Healthcare Record Data. *Drug Safety* 2013; 36: 995-1006. DOI: 10.1007/s40264-013-0061-7.
266. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)* 1972; 34: 187-220.
267. Kaplan EL and Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 1958; 53: 457-481. DOI: 10.2307/2281868.
268. Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* 1972; 14: 945-966. DOI: 10.2307/1267144.
269. Aalen O. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 1978; 6: 701-726.
270. Royston P, Choodari-Oskooei B, Parmar MKB, et al. Combined test versus logrank/Cox test in 50 randomised trials. *Trials* 2019; 20: 172. journal article. DOI: 10.1186/s13063-019-3251-5.
271. Casella G and Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury, 2002.
272. Rodríguez G. Censoring and The Likelihood Function, https://data.princeton.edu/wws509/notes/c7s2 (accessed 11/02/2021).
273. Trinquart L, Jacot J, Conner SC, et al. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. *Journal of Clinical Oncology* 2016; 34: 1813-1819. DOI: 10.1200/jco.2015.64.2488.

274. Trifirò G, Pariente A, Coloma PM, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiology and Drug Safety* 2009; 18: 1176-1184. DOI: 10.1002/pds.1836.
275. Stegherr R, Beyersmann J, Jehl V, et al. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. *Biometrical Journal* 2020; n/a: 1-21. DOI: https://doi.org/10.1002/bimj.201900347.
276. Banerjee PJ, Cornelius VR, Phillips R, et al. Adjunctive intraocular and peri-ocular steroid (triamcinolone acetonide) versus standard treatment in eyes undergoing vitreoretinal surgery for open globe trauma (ASCOT): study protocol for a phase III, multi-centre, double-masked randomised controlled trial. *Trials* 2016; 17: 339. DOI: 10.1186/s13063-016-1445-7.
277. O'Neill RT. Assessment of safety. In: Peace KE (ed) *Biopharmaceutical statistics for drug development*. New York: Marcel Dekker, 1988, pp.543-604.
278. Conover WJ. *Practical Nonparametric Statistics*. 3rd Edition ed. New York: Wiley, 1999.
279. StataCorp. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press, 2017.
280. Fienberg SE. *The Analysis of Cross-Classified Categorical Data*. 2nd edition ed. Cambridge, MA: MIT Press, 1980.
281. Altman DG. *Practical statistics for medical research*. London: Chapman & Hall/CRC, 1991.
282. Zelterman D and Louis T. *Contingency tables in medical studies.* Boston: Dekker, 1992.
283. Grambsch P and Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; 81: 515-526. DOI: 10.1093/biomet/81.3.515.
284. Schoenfeld D. Partial Residuals for The Proportional Hazards Regression Model. *Biometrika* 1982; 69: 239-241. DOI: 10.2307/2335876.
285. Royston P and Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology* 2016; 16: 16. DOI: 10.1186/s12874-016-0110-x.
286. Royston P. A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome. *Stata Journal* 2017; 17: 405-421.
287. Yao B, Zhu L, Jiang Q, et al. Safety monitoring in clinical trials. *Pharmaceutics* 2013; 5: 94-106. Article. DOI: 10.3390/pharmaceutics5010094.
288. Kraemer HC. Events per person-time (incidence rate): a misleading statistic? *Stat Med* 2009; 28: 1028-1039. 2009/01/20. DOI: 10.1002/sim.3525.
289. Thompson J, Palmer T and Moreno S. Bayesian analysis in Stata with WinBUGS. *The Stata journal* 2006; 6: 530-549.
290. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; 25: 4279-4292. DOI: doi:10.1002/sim.2673.
291. Agency. EM. A guideline on summary of product characteristics. In: Agency. EM, (ed.). 2009.
292. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *arXivorg* 2018; Preprint.
293. Royston P and Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* 2014; 15: 314. DOI: 10.1186/1745-6215-15-314.
294. Royston P and B. Parmar MK. A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. *Trials* 2020; 21: 315. DOI: 10.1186/s13063-020-4153-2.
295. Tuyl F, Gerlach R and Mengersen K. A Comparison of Bayes–Laplace, Jeffreys, and Other Priors. *The American Statistician* 2008; 62: 40-44. DOI: 10.1198/000313008X267839.
296. Lowe GDO, Rumley A, Woodward M, et al. Oral contraceptives and venous thromboembolism. *The Lancet* 1997; 349: 1623. DOI: 10.1016/S0140-6736(05)61660-1.
297. Xia HA and Jiang Q. Statistical Evaluation of Drug Safety Data. *Therapeutic Innovation and Regulatory Science* 2014; 48: 109-120. Review. DOI: http://dx.doi.org/10.1177/2168479013510917.

298. James EC, Dunn D, Cook AD, et al. Overlap between adverse events (AEs) and serious adverse events (SAEs): a case study of a phase III cancer clinical trial. *Trials* 2020; 21: 802. 2020/09/19. DOI: 10.1186/s13063-020-04718-z.
299. Chis Ster A, Phillips R, Sauzet O, et al. Improving analysis practice of continuous adverse event outcomes in randomised controlled trials - a distributional approach. *Trials* 2021; 22: 419. DOI: 10.1186/s13063-021-05343-0.
300. Lopes GS, Tournigand C, Olswold CL, et al. Adverse event load, onset, and maximum grade: A novel method of reporting adverse events in cancer clinical trials. *Clinical Trials* 2021; 18: 51-60. DOI: 10.1177/1740774520959313.
301. Moody G, Addison K, Cannings-John R, et al. Monitoring adverse social and medical events in public health trials: assessing predictors and interpretation against a proposed model of adverse event reporting. *Trials* 2019; 20: 804. DOI: 10.1186/s13063-019-3961-8.
302. Crowe B, Brueckner A, Beasley C, et al. Current Practices, Challenges, and Statistical Issues With Product Safety Labeling. *Statistics in Biopharmaceutical Research* 2013; 5: 180-193.
303. Chen WF, Zhao NQ, Qin GY, et al. A Bayesian Group Sequential Approach to Safety Signal Detection. *Journal of Biopharmaceutical Statistics* 2013; 23: 213-230. DOI: 10.1080/10543406.2013.736813.

Appendices

Appendix A2.1: Data items extracted from publications
Appendix A2.2: Selection criteria used to select events presented in the main journal report
Appendix A2.3: Selection criteria used to select events presented in the appendix
Appendix A2.4: Characteristics of included studies by funding source (categorical variables)
Appendix A2.5: Characteristics of included studies by funding source (continuous variables)
Appendix A3.1: Search terms by database
Appendix A3.2 - Data extraction sheet
Appendix A3.3: Completed PRISMA checklist
Appendix A3.4: Summary of hypothesis tests to analyse prespecified harm outcomes in phase II/III RCTs
Appendix A3.5: Summary of hypothesis tests to analyse emerging harm outcomes in phase II/III RCTs
Appendix A3.6: Summary of estimation approaches to analyse emerging harm outcomes in phase II/III RCTs
Appendix A3.7: Summary of decision making probability methods to analyse prespecified harm outcomes in phase II/III RCTs
Appendix A3.8: Summary of decision making probability methods to analyse emerging harm outcomes in phase II/III RCTs
Appendix A4.1: Survey questions
Appendix A4.2: Email invitation sent to senior statistician representatives at UKCRC registered CTUs or personal contacts within industry
Appendix A4.3: Text from participant information sheet for CTU participants
Appendix A4.4: Clinical areas participants indicating they predominantly worked on
Appendix A4.5: Free text comments regarding other information presented on emerging harm outcomes
Appendix A4.6: Free text comments regarding methods participants are aware of specifically for the analysis of harm outcomes
Appendix A4.7: Free text comments regarding participants' use of specialist methods for analysis of emerging harm outcomes
Appendix A5.1: Consensus group attendees
Appendix A5.2: Plots for consideration - multiple binary outcomes
Appendix A5.3: Plots for consideration - multiple time-to-event outcomes
Appendix A5.4: Plots for consideration - Single binary outcomes
Appendix A5.5: Plots for consideration - Single continuous outcomes
Appendix A5.6: Plots for consideration - single time-to-event outcomes
Appendix A5.7: Plots for consideration - multiple continuous outcomes
Appendix A5.8: Questions clinicians were asked to consider for each plot during interviews
Appendix A5.9: Table and figures summarising initial appraisals of all plots by outcome type
Appendix A5.10: Free text comments summarised for each plot
Appendix A5.11: Tables summarising Mentimeter votes to decide which plots to take forward and amendments
Appendix A7.1: Power and false positive rates of the Fisher's exact test and Chi-squared test
Appendix A7.2: Power and false positive rates of the of the alternative Weibull survival models

Appendix A7.3: Sample sizes required by each of the other investigated tests to achieve 80% power and the specific power of each test
Appendix A7.4: False positives across scenarios
Appendix A7.5: Power of Bayesian methods with varying thresholds of risk to detect
Appendix A.8: Table of permissions for all reused copyrighted works in this thesis
Appendix A.9: Copies of permission documents to republish copyrighted works (my own work)
Appendix A.10: Copies of permission documents to republish third party copyrighted works

Appendix A2.1: Data items extracted from publications

			Items collected	Instructions
Study details		1	Study number	
		2	Journal	
		3	Funding source: public, private, both or unspecified.	Studies will be assumed to be funded by industry only if this is explicitly stated.
Study characteristics		4	Control: placebo, active or both	Select placebo if no active treatment is given, else active. Both should be selected for trials with multiple arms where there is at least one group receiving no active treatment and one group receiving an active treatment.
		5	Number of centres	
		6	Number randomised	
		7	Study duration (length of trial follow-up)	
Methods	Details of how AE outcomes were defined (coding, attribution) and were collected (mode of collection, timing)	8	Describe the collection method: passive surveillance, patient prompted, clinical examinations (e.g. vital signs or urine samples), and laboratory tests. (Select all that apply)	*Passive*: If authors state that AEs were collected throughout the study with no further information we will assume that collection was passive. *Prompted*: Prompted methods include, but are not limited to: questions about both specific events and AEs in general, questionnaires, or diaries.
		9	Stated the timing of collection.	
		10	Mention dictionary for coding of events: Researcher defined, MedDRA, CTCAE, WHO-ART, COSTART, ICD-10, other or not applicable	
		11	Describe who undertook the assessment of attribution to study drug: blinded assessor, unblinded assessor or not specified.	
Planned analysis	Details of any plans for analysing AE outcomes	12	Describe analysis for AE outcomes in the statistical methods.	Reference must be made to harmful events e.g. AEs or a specific harm event, this cannot be simply how binary events will be analysed.
		13	Define a 'safety' population for analysis.	
		14	Specify a planned interim analysis with stopping criteria: based on efficacy, based on safety, based on both efficacy and safety, yes but no other details given, no planned interim analysis or unclear	Criteria for stopping must be set out, it is not enough to say that the DMC reviewed the data.

Results	Details of what was reported and where	15	What was reported in the main paper: summaries of type of AEs (e.g. AE, SAE, AR, ADR), actual AE terms, both, neither or not applicable?	Not applicable is relevant when for example authors explicitly state that there are no events or there is only one event so summaries are inappropriate.
		16	What was reported in the appendix: summaries of type of AEs (e.g. AE, SAE, AR, ADR), actual AE terms, both, neither or not applicable?	Not applicable is relevant when for example authors explicitly state that there are no events or there is only one event so summaries are inappropriate. We will only search the appendix/supplementary material for AE data if the main article makes reference to it.
		17	Who was the AE analysis performed on: all randomised, participants who took at least a single dose, other or not specified?	
		18	How were number of drop-outs/withdrawals reported: By treatment arm, overall, not reported or not applicable?	Not applicable is relevant when there are no drop-outs/withdrawals. This does not include discontinuation of treatment.
		19	Were drop-outs/withdrawals due to AEs reported: Yes, no or not applicable?	Not applicable if drop-outs/withdrawals are not reported or if it is reported that there are no drop-outs/withdrawals.
		20	Were specific AEs causing withdrawals reported: Yes, no or not applicable?	Not applicable if drop-outs/withdrawals due to AEs are not reported or if it is reported that there are no drop-outs/withdrawals due to AEs.
		21	What was the selection criteria for the AEs reported?	Free text response where possibilities can include for example**:** most frequent, above a severity threshold, SAEs. Include details of what is in the main journal article and what is in the appendix separately.
	Details of how AEs were summarised and presented - binary outcomes	22	What summary information was given: Number of people, number of events, both, unclear, not summarised or not applicable?	Only select 'number of events' if presented for each individual event not just overall number of events. Not applicable is only relevant when report that there are no AEs.
		23	What analysis was performed: frequencies, percentages, differences and 95% confidence intervals, significance tests, other? (Select all that apply)	

	Details of how AEs were summarised and presented - continuous outcomes	24	Were continuous outcomes dichotomised: Yes for all, yes for some, no or not applicable?	This includes measures that will have been captured as continuous and then dichotomised for example blood levels, blood pressure etc.
		25	If continuous outcomes were analysed as continuous what analysis was performed: differences in measures of central tendency, significance tests, other? (Select all that apply)	
	Details of how AEs were summarised and presented	26	Were signal detection methods used?	
		27	Were any graphical summaries of AEs presented?	
		28	Were severity ratings given: Yes for all, yes for some, no or not applicable?	
		29	Were numbers of serious events presented: Yes by treatment arm, yes overall, no or not applicable?	If death is reported as part of the efficacy outcome it is not enough to constitute reporting serious events.
		30	Were serious events coded as treatment related: Yes for all, yes for some, no or not applicable?	
		31	Provided information on the duration of events?	This refers to the length of the actual AE i.e. how long did it last.
		32	Provided information on the timing of events?	This refers to the time of onset of the AE.
		33	Accounted for multiplicity of statistical tests?	
		34	Referenced CONSORT extension for harms?	

Appendix A2.2: Selection criteria used to select events presented in the main journal report

Selection criteria	n	%
All AEs presented	20	10.87
AEs in greater than x% in any group	10	5.43
AEs in greater than x% in treatment group	4	2.17
AEs in greater than x% in all patients	1	0.54
Most common (no criteria specified)	9	4.89
Predefined AEs	26	14.13
SAEs	15	8.15
AEs leading to study drug discontinuation/interruption	3	1.63
Treatment related AEs	5	2.72
Grade 3>= events	9	4.89
AEs in greater than x% in any group & predefined/special interest AEs	4	2.17
AEs in greater than x% in any group & frequency between groups differed by more than y% & predefined/special interest AEs	1	0.54
AEs in greater than x% in all patients & predefined/special interest AEs	3	1.63
AEs in greater than x% in treatment group & AEs of special interest	2	1.09
AEs in greater than x% in any group & all SAEs	2	1.09
AEs in greater than x% in all patients & all SAEs	1	0.54
AEs in greater than x% in any group & SAEs related to treatment	1	0.54
Most common (no criteria specified) & predefined/special interest AEs	3	1.63
Most common (no criteria specified) & all SAEs	4	2.17
Most common (no criteria specified) & all SAEs & AEs leading to study drug discontinuation/interruption	1	0.54
Most common (no criteria specified) & treatment related SAEs	1	0.54
AEs where frequency between groups differed by more than y% & all SAEs	1	0.54
AEs of special interest	6	3.26
Grade >=3 AEs in greater than x% of patients	1	0.54
Grade >=3 AEs in greater than x% in intervention & y% in control	1	0.54
Most common (no criteria specified) grade 3>= AEs	1	0.54
Most common SAEs (no criteria specified)	1	0.54
SAEs & AE of special interest	1	0.54
Treatment related AEs in greater than x% of patients	1	0.54
Treatment related AEs in greater than x% in any group	1	0.54
AEs in greater than x% in treatment group & SAEs	1	0.54
AEs in greater than x% in treatment group & SAEs & predefined AEs	2	1.09
AEs in greater than x% in any group & significantly different & SAEs	1	0.54
AEs in greater than x% in any group & treatment related AEs/SAEs	2	1.09
AEs in greater than x% in treatment group & treatment related AEs & SAEs	1	0.54
AEs in greater than x% in treatment group & treatment related AEs in greater than y% in all patients	1	0.54
AEs in greater than x% in any group & Grade 3>= events	1	0.54
AEs in greater than x% in all patients & Grade 3>= events	1	0.54
AEs in greater than x% in all patients & Grade 2>= treatment related AEs	1	0.54
AEs in greater than x% in any group & Grade 3>= events in greater than y% in any group	1	0.54
AEs in greater than x% in any group & SAEs in treatment group	1	0.54
AEs in greater than x% in any group & AEs of special interest & most common (no criteria specified) AEs leading to treatment discontinuation/interruption & predefined AEs	1	0.54

AEs in greater than x% in any group, AEs of special interest in greater than y% in treatment group & treatment related deaths	1	0.54
AEs in greater than x% in treatment group & SAEs in greater than y% in any group	1	0.54
AEs and SAEs occurring more often in treatment group than control	1	0.54
AEs in greater than x% in treatment group & occurred more often in treatment group than control & predefined/special interest AEs	1	0.54
AEs in greater than x% in any group & frequency between groups differed by more than y%, SAEs in greater than z% in any group & all grade >=3 AEs	1	0.54
AEs in greater than x% patients & more than y% difference between treatment groups & AEs leading to treatment discontinuation/interruption & most common SAEs (no criteria specified) & death	1	0.54
Predefined AEs, AEs leading to hospitalisation/death/study drug discontinuation/interruption & SUSARS	2	1.09
Some form of overall summary	6	3.26
Not specified how selected	6	3.26
Not summarised in main paper	11	5.98

Appendix A2.3: Selection criteria used to select events presented in the appendix

Selection criteria	n	%
All AEs	18	9.78
SAEs	18	9.78
All AEs & SAEs	4	2.17
AEs in greater than x% in any group	7	3.8
AEs in greater than x% in treatment group	2	1.09
AEs in greater than x% in all patients	1	0.54
AEs in greater than x% in any group & all SAEs	2	1.09
AEs in greater than x% in treatment group & all SAEs	1	0.54
AEs in greater than x% in all patients & all SAEs	3	1.63
AEs in greater than x% in treatment group & all SAEs	1	0.54
AEs in greater than x% in treatment group & greater than in control group & all SAEs	1	0.54
SAEs in greater than x% in any group	1	0.54
AEs in greater than x% in any group & SAEs in greater than y% in any group	1	0.54
AEs in greater than x% in any group & AEs of special interest	2	1.09
Treatment related AEs	5	2.72
Treatment related AEs in greater than x% in any group	2	1.09
Grade 3>= events	2	1.09
Predefined AEs	8	4.35
AEs of special interest	1	0.54
AEs leading to study drug discontinuation/interruption	2	1.09
AEs leading to study drug discontinuation & SAEs	1	0.54
Grade 3>= events leading to study drug discontinuation & grade 3>= laboratory results	1	0.54
Treatment related AEs & AEs leading to study drug discontinuation	1	0.54
AEs in greater than x% in all patients leading to treatment discontinuations, SAEs in greater than x% in any group, serious predefined/special interest AEs and clinically significant laboratory results	1	0.54
AEs in greater than x% in any group, treatment related AEs in greater than x% in any group, treatment related SAEs and select AEs	1	0.54
Clinical laboratory data	1	0.54
Predefined AEs, AEs leading to hospitalisation/death/study drug discontinuation/interruption & SUSARS	3	1.63
Deaths	2	1.09
Some form of overall summary	5	2.72
Not specified how selected	2	1.09
Not summarised in the appendix	84	45.65

Appendix A2.4: Characteristics of included studies by funding source (categorical variables)

Characteristic		Public (N=70)		Industry (N=80)		Both (N=33)	
		n	%	n	%	n	%
Journal	BMJ	2	2.9	1	1.3	0	0.0
	JAMA	18	25.7	13	16.3	6	18.2
	Lancet	19	27.1	30	37.5	13	39.4
	NEJM	31	44.3	36	45.0	14	42.4
Centre	Single centre	7	10.0	2	2.8	3	10.0
	Multi-centre	63	90.0	70	97.2	27	90.0
Control	Placebo	38	54.3	41	51.3	16	48.5
	Active	31	44.3	34	42.5	15	45.5
	Both	1	1.4	4	5.0	2	6.1
	Neither[a]	0	0.0	1	1.3	0	0.0

[a] One trial compared interventional drug to behavioural change intervention

Appendix A2.5: Characteristics of included studies by funding source (continuous variables)

Characteristic	Public (N=70)			Industry (N=80)			Both (N=33)		
	Median	(IQR)	min, max	Median	(IQR)	min, max	Median	(IQR)	min, max
Sample size	575	(300,1273)	73, 205513	556	(259, 1599)	34, 16590	544	(260, 2127)	30, 12705
Centres[a]	13	(4, 29)	1, 251	76	(35, 148)	1, 1368	41	(13, 209)	1, 410
Participants per centre[a]	60	(20, 225)	2, 15809	7	(4, 12)	1, 4464	14	(8, 30)	3, 1922
Trial duration (weeks)[b]	39	(22, 104)	0.3, 390	52	(26, 100)	1, 261	104	(20, 228)	1, 521

Abbreviations: *IQR = Inter-quartile range; min = minimum; and max = maximum*

[a] 11 reports did not specify the number of centres

[b] 2 reports did not specify trial duration

Appendix A3.1: Search terms by database

Medline via Ovid
1. Models, Statistical/
2. Models, Theoretical/
3. Biostatistics/
4. Statistics, Nonparametric/
5. Statistics as Topic/
6. Bayes Theorem/
7. Biometry/
8. Statistical Model*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
9. Statistical Method*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
10. Bayes* Model*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
11. Bayes* Theor*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
12. Bayes* Method*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
13. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14. exp "Drug-Related Side Effects and Adverse Reactions"/
15. (Side effect* adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
16. ((adverse or undesirable or harm* or serious or toxic) adj3 (effect* or reaction* or event* or outcome* or experience*) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
17. ((toxicity or complication* or noxious or tolerability) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, original title, name

of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

18. ((AE or SAE or ADR) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

19. (Safety adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

20. 14 or 15 or 16 or 17 or 18 or 19

21. exp Clinical Trials as Topic/

22. Clinical trial*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

23. Clinical stud*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

24. 21 or 22 or 23

25. 13 and 20 and 24

EMBASE via Ovid
1. statistical model/
2. biostatistics/
3. nonparametric test/
4. statistics/
5. bayes theorem/
6. biometry/
7. statistical analysis/
8. mathematical model/
9. Statistical Model*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
10. Statistical Method*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
11. Bayes* Model*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
12. Bayes* Theor*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
13. Bayes* Method*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
14. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13
15. exp adverse drug reaction/
16. side effect/
17. (Side effect* adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
18. ((adverse or undesirable or harm* or serious or toxic) adj3 (effect* or reaction* or event* or outcome* or experience*) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
19. ((toxicity or complication* or noxious or tolerability) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]
20. ((AE or SAE or ADR) adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, heading word, drug trade name, original

title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]

21. (Safety adj5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]

22. 15 or 16 or 17 or 18 or 19 or 20 or 21

23. exp "clinical trial (topic)"/

24. Clinical trial*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]

25. Clinical stud*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]

26. 23 or 24 or 25

27. 14 and 22 and 26

Web of Science

TOPIC: (((Statistical near/0 Model*) or (Statistical near/0 Method*) or (Bayes* near/0 Model*) or (Bayes* near/0 Theor*) or (Bayes* near/0 Method*)) and (((Side effect*) near/5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)) or ((adverse or undesirable or harm* or serious or toxic) near/3 (effect* or reaction* or event* or outcome* or experience*) near/5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)) or ((toxicity or complication* or noxious or tolerability) near/5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)) or ((AE or SAE or ADR) near/5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*)) or ((Safety) near/5 (monitor* or analys* or flag* or signal* or detect* or evaluat* or screen* or assess* or identif*))) and ((Clinical near/0 trial*) or (Clinical near/0 stud*)))

Scopus
((TITLE-ABS-KEY (statistical W/0 model*)) OR (TITLE-ABS-KEY (statistical W/0 method*)) OR (TITLE-ABS-KEY (bayes* W/0 model*)) OR (TITLE-ABS-KEY (bayes* W/0 theor*)) OR (TITLE-ABS-KEY (bayes* W/0 method*))) AND ((TITLE-ABS-KEY ("Side effect*" W/5 ((monitor*) OR (analys*) OR (flag*) OR (signal*) OR (detect*) OR (evaluat*) OR (screen*) OR (assess*) OR (identif*)))) OR (TITLE-ABS-KEY ((adverse OR undesirable OR harm* OR serious OR toxic) W/3 (effect* OR reaction* OR event* OR outcome* OR experience*) W/5 (monitor* OR analys* OR flag* OR signal* OR detect* OR evaluat* OR screen* OR assess* OR identif*))) OR (TITLE-ABS-KEY ((toxicity OR complication* OR noxious OR tolerability) W/5 (monitor* OR analys* OR flag* OR signal* OR detect* OR evaluat* OR screen* OR assess* OR identif*))) OR (TITLE-ABS-KEY ((ae OR sae OR adr) W/5 (monitor* OR analys* OR flag* OR signal* OR detect* OR evaluat* OR screen* OR assess* OR identif*))) OR (TITLE-ABS-KEY ((safety) W/5 (monitor* OR analys* OR flag* OR signal* OR detect* OR evaluat* OR screen* OR assess* OR identif*)))) AND ((TITLE-ABS-KEY (clinical W/0 trial*)) OR (TITLE-ABS-KEY (clinical W/0 stud*)))

Appendix A3.2 - Data extraction sheet

Study/author characteristics			
1	Study title		
2	Authors		
3	Author affiliations: Academic, Public, Pharmaceutical, Other (select all that apply)		
4	Funding body/organisation		
5	Journal		
6	Year of publication		
Methodological characteristics			
8	Single or multiple methods proposed		
9	Number of methods proposed		
	For each method		
10	Describe the method		
11	Pre-specify the event for the method or the method is designed to screen emerging events. *(If you don't have to specify an event to set up a hypothesis etc. then answer screen emerging events)*		
12	Method applies to specific events (i.e. applied to each individual event) or events are aggregated (i.e. events are grouped together)		
i	If events are aggregated for analysis how are they grouped e.g. by system organ class (SOC), overall number of AEs etc.		
13	Type of outcome(s): Continuous, number, proportion, count, incidence rates, time-to-event, other (specify) (select all that apply)		
14	If applicable describe any test that is performed		
15	Describe any assumptions the method makes		

16	Incorporates prior/external information: Yes/No		
i	What prior/external information can be incorporated?		
ii	Bayesian methods: Yes/No		
17	Output: summary statistic, test-statistic, p-value, plot, other (specify) (select all that apply)		
18	Software/code for implementation: Yes/No		
i	Which software?		
ii	Is the code open-source?		
iii	Source/Available at		
iv	Details		
Other			
20	Useful references to follow-up (not already included) - Comment on relevance/eligibility of each, with reasons if ineligible		
21	Linked references (from eligible studies)		
22	Other relevant information		
23	Subjective opinion: strenghts and weaknesses of the method		

Appendix A3.3: Completed PRISMA checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	Not applicable for thesis write up
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	Not applicable for thesis write up
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	Introduction: section 3.1, paragraph 2 and Methods: section 3.3 paragraph 1
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	Aims: section 3.2,
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	Methods: section 3.3
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	Methods: section 3.3.2
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	Methods: section 3.3.1
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	Supplementary material: item A3.1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	Methods: section 3.3.1 and 3.3.2
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	Methods: section 3.3.3
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	Methods: section 3.3.3 Supplementary material: item A3.2
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how	NA

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
		this information was used in any data synthesis (if appropriate).	
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	Methods: section 3.3.4
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	Results: section 3.4.1 and Figure 3.1
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	Results: section 3.4.2
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	NA
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Results: section 3.4.4, Table 3.3, Table 3.4 and Supplementary material tables A.3.4-A3.8
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	Results: section 3.4.3, table 3.2 and Figure 3.2
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	Discussion: section 3.5.1
Limitations	20	Discuss the limitations of the scoping review process.	Discussion: section 3.5.3
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	Discussion – Conclusions: section 3.5.5
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	Not applicable for thesis write up

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. ;169:467–473. doi: 10.7326/M18-0850

Appendix A3.4: Summary of hypothesis tests to analyse prespecified harm outcomes in phase II/III RCTs

Outcome	Data type	Model	Reference	Brief Description
Prespecified harm outcome	Binary	Logit	Bolland & Whitehead[144]	Alpha-spending function for sequential monitoring
	Binary, count, time-to-event or continuous	Not applicable	Fleishman & Parker[145]	Redefine significance threshold for sequential monitoring
	Binary, count, time-to-event or continuous	Not applicable		Conditional power at each interim analysis for sequential monitoring
	Time-to-event	Exponential		Alpha-spending function for sequential monitoring
	Binary or incidence rate	Binomial or Poisson	Lieu et al.[185]	Maximised sequential probability ratio test for sequential monitoring
	Binary, count - incident rate	Poisson	Shih, Lai, Heyse & Chen[148]	Sequential generalised likelihood ratio test for sequential monitoring
	Binary, count, time-to-event or continuous	Not applicable	Liu[147]	Non-inferiority test for final analysis

Appendix A3.5: Summary of hypothesis tests to analyse emerging harm outcomes in phase II/III RCTs

Outcome	Data type	Model	Reference	Brief Description
Emerging adverse events (multiple)	Binary	Not applicable	Mehrotra & Heyse[154]	P-value adjustment
	Binary	Not applicable	Mehrotra & Adewale[153]	P-value adjustment
Emerging adverse events (single)	Time-to-event	Poisson	Huang, Zalkikar & Tiwari[152]	Likelihood ratio test to compare relative risk for time-to-first event
	Time-to-event	Poisson		Likelihood ratio test to compare relative risk allowing recurrent events
Overall harm profile	Binary	Multivariate - Markov chain	Bristol & Patel[150]	Multivariate likelihood ratio test with Markov chain of order one
	Binary	Multivariate - chi-squared	Chuang-Stein Mohberg & Musselman[151]	Multivariate test with chi-squared distribution
	Binary	Multinomial - logit	Agresti & Klingenberg[149]	Likelihood ratio test using a logit models to test for equality of two vectors for the marginal distributions
	Binary	Exact permutation distribution		Likelihood ratio test using the exact permutation distribution to test joint distributions
	Binary	Multinomial - logistic normal random intercept model		Likelihood ratio test using a logistic normal random intercept model to compare marginal distributions whilst modelling the joint distributions

Appendix A3.6: Summary of estimation approaches to analyse emerging harm outcomes in phase II/III RCTs

Outcome	Data type	Estimate	Reference	Brief Description
Emerging adverse events (multiple)	Ordinal	Posterior probability	Leon-Novelo, Zhou, Nebiyou Bekele & Muller[160]	Posterior probability of each grade of an AE (participant maximum grade used) allowing multiple different events per participant
Emerging adverse events (single)	Binary	Frequencies & percentage	Evans & Nitsch[43]	Standard estimates for AE analysis including frequencies, percentages, risk differences and odds ratios
	Binary	Regression models	Evans & Nitsch[43]	Regression based approaches for AE analysis e.g. Poisson regression
	Binary	Confidence interval	O'Gorman, Woolson, Jones[163]	Two methods to estimate CIs for risk-difference when combing data across multiple sites
	Count	Confidence interval	Liu, Wang, Liu & Snavely[191]	Four methods to estimate CIs for exposure adjusted incident ratios
	Binary	Confidence interval	Borkowf[156]	Alternative to the Clopper-Pearson CI for proportions
	Binary	Mean cumulative function	Siddiqui[44]	Non-parametric estimate of mean cumulative number of recurrent events
	Binary	Prevalence	Lancar, Kramar & Haie-Meder[159]	Non-parametric estimate of prevalence of event allowing for recurrence
	Binary	Mean frequency function	Gong, Tong, Strasak & Fang[157]	Non-parametric estimate of mean cumulative number of recurrent events in presence of competing risks
	Binary	Mean cumulative duration	Wang & Quartey, 2012[166]	Non-parametric estimate of mean cumulative duration for recurrent events
	Binary	Mean cumulative duration	Wang & Quartey, 2013[167]	Semi-parametric estimate of mean cumulative duration and prevalence of recurrent events
	Binary	Dependence between AEs and discontinuation	Rosenkranz[164]	Three methods to estimate the level of dependence between AE and discontinuation by treatment group that corrects for any dependence in the treatment effect estimate
	Time-to-event	Hazard ratio	Henglebrock, Gillhaus, Kloss & Leverkus[158]	Two methods to estimate hazard ratio for recurrent events

	Time-to-event	Cumulative incidence function	Allignol, Beyersmann & Schmoor[155]	Two methods to estimate the probability of an event in presence of competing risks
	Time-to-event	Conditional cumulative incidence function	Nishikawa, Tango & Ogawa[162]	Probability of a recurrent event in presence of competing risks
Laboratory & vital signs	Continuous	GENIE score	Sogliero-Gilbert, Ting, & Zubkoff[165]	Weighted linear combination of absolute normalised deviations from the reference range to indicate abnormalities

Appendix A3.7: Summary of decision making probability methods to analyse prespecified harm outcomes in phase II/III RCTs

Outcome	Data type	Model	Prior	Reference	Brief Description
Prespecified harm outcome	Binary	Beta-Binomial	Beta	Berry[168]	Posterior probability that event rate or incidence rate (incorporating exposure time) is greater in the treatment group compared to control group
	Time-to-event	Exponential	Not specified		
	Binary	Beta-Binomial	Beta	Yao, Zhu, Jiang & Xia[170]	Beta-binomial model to give posterior probability that predefined risk difference threshold is exceeded
	Count	Gamma-Poisson	Gamma	Zhu, Yao, Xia & Jiang[171]	Gamma-Poisson model to give posterior probability that predefined risk difference (incorporating exposure time) threshold is exceeded
	Binary	Logit model	Normal	French, Thomas and Wang[169]	Logit model and a piecewise exponential model to give posterior probabilities that predefined risk difference threshold is exceeded
	Time-to-event	Piecewise exponential	Normal & Gamma		

Appendix A3.8: Summary of decision making probability methods to analyse emerging harm outcomes in phase II/III RCTs

Outcome	Data type	Model	Prior	Reference	Brief Description
Emerging adverse events (multiple)	Binary	Logit	Mixed	Berry & Berry[172]	Bayesian hierarchical logit model to give posterior probability that event rate greater in treatment compared to control group
	Binary	Logit	Normal	Xia, Ma &Carlin[177]	Bayesian hierarchical logit and log-linear (incorporating exposure time) models to give posterior probability that event rate greater in treatment compared to control group
	Count	Log (Poisson)	Mixed		
	Count	Log (Poisson)	Normal		
	Binary	Logit	Mixed	Chen[303]	Sequential method. Bayesian hierarchical logit model to give posterior probability that event rate greater in treatment compared to control group for interim analysis
	Binary	Beta-Binomial	Isling	McEvoy[176]	Multivariate approach to give posterior probability of difference in event rates based on indicator functions
	Binary	Beta-Binomial	Beta	Gould[174]	Posterior probability that AEs in treatment group produced by a larger process than AE in control group
	Count	Gamma-Poisson	Poisson	Gould[175]	Posterior probability that AEs in treatment group produced by a larger process than AE in control group accounting for exposure time

Appendix A4.1: Survey questions

Study Title: Statisticians survey on statistical methods for adverse event data analysis in randomised controlled trials

This survey pertains to the final analysis of emerging harms reported or screened for in clinical trials. Not predefined harm outcomes of interest or analysis to monitor ongoing trials (interim analyses).

Number	Question	Response options				
1	How long have you worked as a clinical trial statistician? (Please specify the number of years)					
2	Do you work for:	Academic institution	NHS trust	Pharmaceutical company	Clinical Research Organisation	Other (please specify)
3	Is there a clinical area you predominantly work on? If yes, please specify	No	Yes			
4	What is the typical size of the trials you work on?	1-10	11-50	51-100	101-500	>500
5	What is the typical phase of the trials you work on?	Phase I/Dose-finding	Phase II/III	Phase IV		

Before you proceed we thought it would be helpful for you to know about our recent findings.

We undertook a systematic review of RCT journal reports and found that trials typically report AE data using frequencies (94%) and percentages (87%). They often ignore repeated events (84%) and 47% undertake hypothesis tests despite a lack of power. There is also a common practice to categorise continuous clinical and laboratory outcomes and present as frequencies and percentages (59%). A small proportion (12%) incorporated graphics into the AE analysis.

	Thinking about analysis methods for AEs:					
6	How often would you say the following influences the analysis performed?					
i	Statistician prefers simple approaches e.g. tables of frequencies and percentages	Always	Often	Not very often	Never	Don't know

	ii	Chief investigator prefers simple approaches e.g. tables of frequencies and percentages	Always	Often	Not very often	Never	Don't know
	iii	Journal prefers simple approaches e.g. tables of frequencies and percentages	Always	Often	Not very often	Never	Don't know
	iv	Regulator prefers simple approaches e.g. tables of frequencies and percentages	Always	Often	Not very often	Never	Don't know
	v	Trial sample size	Always	Often	Not very often	Never	Don't know
	vi	The number of different AEs experienced across the trial	Always	Often	Not very often	Never	Don't know
	vii	AE rates	Always	Often	Not very often	Never	Don't know
		Thinking about AE analysis you typically perform.					
7		In your experience the following is a barrier when analysing AEs:					
	i	Lack of awareness of appropriate methods	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	ii	Lack of knowledge to implement appropriate methods	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iii	Lack of training opportunities to learn what methods are appropriate	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iv	Lack of statistical software/code to implement appropriate methods	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iv	Trial sample size	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	v	The number of different AEs experienced across the trial	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	vi	AE rates	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
		Thinking about AE analysis.					
8		In your opinion:					
	i	Statisticians don't give AE data the same priority as the primary efficacy outcome	Strongly agree	Agree	Disagree	Strongly disagree	Don't know

	ii	Chief investigators don't give AE data the same priority as the primary efficacy outcome	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iii	Journals don't give AE data the same priority as the primary efficacy outcome	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iv	Regulators don't give AE data the same priority as the primary efficacy outcome	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	v	There are a lack of appropriate analysis methods	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	vi	There are a lack of examples of the use of appropriate analysis methods in the applied literature	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
9		Are you aware of any published methods specifically to analyse AEs?	Yes	No	Don't know		
		If yes, please specify					
10		If answer is 'yes' to question 9 In your opinion why are those methods not being more widely used:					
	i	Available methods are technically too complex	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	ii	Available methods are too resource intensive	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iii	Available methods are not suitable for typical trial sample sizes	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iv	Available methods are not suitable for the number of different AEs typically experienced across a trial	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	v	Available methods are not suitable for typical AE rates observed	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
11		Are there any reasons other than those mention above why those methods are not being more widely used?	Yes	No			
		If yes, please specify					

		Thinking about available methods for AE analysis					
12		How concerned are you about the following:					
	i	Difficulties in interpreting the results/output	Not at all	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
	ii	Robustness of methods	Not at all	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
	iii	Acceptability of methods to chief investigator	Not at all	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
	iv	Acceptability of methods to journal	Not at all	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
	v	Acceptability of methods to regulator	Not at all	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
13		Do you have any other thoughts about current practice for AE analysis?	Yes	No			
		If yes, please specify					
14		To what extent do you agree that the following would support a change in AE analysis practice					
	i	Software/code development is needed	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	ii	Training specifically for AE analysis is needed	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
	iii	Guidance on appropriate AE analysis is needed e.g. case studies, tutorials within open access journals	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
15		Are there any other solutions in addition to those above that would support a change in AE analysis practice?	Yes	No			
		If yes, please specify					
16		When analysing AEs do you present (please select all that apply):					

	i	Number of participants with at least one event	Yes	No
	ii	Number of events	Yes	No
	iii	Other	Yes	No
		If yes, please specify		
17		When analysing AEs which summary statistic would you typically use (please select all that apply)		
	i	Frequency	Yes	No
	ii	Percentage	Yes	No
	iii	Risk difference	Yes	No
	iv	Odds ratio	Yes	No
	v	Risk ratio	Yes	No
	vi	Incidence rate ratio	Yes	No
	vii	Other	Yes	No
		If yes, please specify		
18		In your experience how are AE rates typically compared between treatment groups (please select all that apply)		
	i	Subjective comparison	Yes	No
	ii	Exclusion of null through 95% confidence interval	Yes	No
	iii	Hypothesis test/p-value	Yes	No
	iv	Other	Yes	No
		If yes, please specify		
19		Have you undertaken any specialist AE analysis not mentioned in your previous responses? Please explain your answer. If 'yes', please include details of the method(s) used for the analysis performed	Yes	No

Appendix A4.2: Email invitation sent to senior statistician representatives at UKCRC registered CTUs or personal contacts within industry

Study Title: **Statisticians survey on statistical methods for adverse event data analysis in randomised controlled trials**

Dear colleague,

We are writing to invite you or a member of your organisation/unit to take part in a survey to measure awareness and opinions of statistical methods to analyse adverse events in pharmacological randomised controlled trials (RCTs). We require one response per unit so we ask that you either complete the survey yourself or disseminate this invitation email appropriately.

This research is part of a wider NIHR research fellowship aiming to identify ways to improve the monitoring and analysis of AE in phase II/III pharmacological trials to facilitate earlier detection of adverse drug reactions.

Before you decide whether to take part it is important for you to understand why the research is being done and what it will involve. Please take time to read the attached participant information sheet carefully and discuss it with others if you wish. Please ask us if there is anything that is not clear or if you would like more information, contact details can be found at the bottom of this email.

What action is required?

Once you have read the participant information sheet please follow the link below to access the survey. We approximate that the survey will take no longer than 15 minutes to complete. You will have an eight-week window to complete the survey. Reminder emails will be sent at week 4 and week 6.

Link: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Please complete the survey by 10 June 2019.

Please note that completing the survey and clicking 'Submit' automatically implies your consent to participate. Participation is voluntary and you are free to withdraw at any point

whilst completing the survey. However please note retraction or removal of individual survey answers is not possible once the 'Submit' button has been selected.

Contact information:

Should you have any questions concerning this study, please contact the research team using the details provided below:

Rachel Phillips
Imperial Clinical Trials Unit, Imperial College London, 1st Floor Stadium House, 68 Wood Lane, London, W12 7RH
Email: r.phillips@imperial.ac.uk
Tel: 020 759 49356

We appreciate your consideration to participate in this project.

Many thanks

Rachel Phillips

NIHR Doctoral Research Fellow

FUNDED BY
NIHR | National Institute for Health Research

Appendix A4.3: Text from participant information sheet for CTU participants

Study Title: Statisticians survey on statistical methods for adverse event data analysis in randomised controlled trials

What is the purpose of this study?

This survey will allow an exploration of awareness of statistical methods available to flag AEs as potential adverse drug reactions (ADRs) and identify any potential barriers to their use, as well as gain feedback on ideas for new statistical methods.

Why have I been chosen?

You are eligible to participate in the survey if you satisfy the following inclusion criteria:

i) Your current role is as a senior statistician or equivalent at a UKCRC CTU;

ii) You have experience of planning and preparing final analysis reports for pharmacological RCTs.

We ask you to provide your personal views.

Do I have to take part?

Participation in the study is voluntary. It is up to you to decide whether to take part. If you decide to take part, you are still free to withdraw at any time without having to give a reason. However, retraction or removal of your survey answers is not possible once the 'Submit' button has been selected.

What are the possible disadvantages and risks of taking part?

There are no disadvantages that we are aware of from taking part in this study.

What if something goes wrong?

We are not aware of any risks involved in taking part in this study.

Will my taking part in this study be kept confidential?

All personal records relating to this study will be kept confidential. We will use SurveyMonkey to capture your responses. No personal data will be collected in the survey, as such your responses to this survey will be anonymous. Responses will be kept in a secure password-protected and encrypted file and stored on Box cloud content management platform. Data in Box is stored securely and automatically backed up. The Box platform is fully General Data Protection Regulation (GDPR) compliant. Upon completion of the study the research data will be uploaded to an approved data-sharing repository. This will be maintained for at least ten years from the time the research study is complete.

What will happen to the results of the research study?

The results of this study will be analysed and published in an open access peer reviewed scientific journal. The work will also be submitted for oral presentation at a range of academic conferences targeting statisticians and the wider clinical trial community. If you would like help in locating and viewing the published results please contact us using the details below. Study data will be stored for ten years post end of study in keeping with Imperial College London research policy.

No identifying data will be published.

Will I receive payment for participating in the study?

You will not be paid for taking part in this study but upon successful completion of the survey, you will be entered into a prize draw for a chance to win £50 worth of Amazon vouchers.

Who is organising and funding the research?

This study is being organised and sponsored by Imperial College London. This study is funded by the National Institute for Health Research (NIHR) (grant reference number DRF-2017-10-131). Please note that the views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Who has reviewed this study?

This study has been reviewed by the Head of Imperial Clinical Trials Unit and granted ethical approval by the Imperial College Joint Research Compliance Office (JRCO).

What action is required?

Please follow the link in the invitation email to access the survey. We approximate that the survey will take no longer than 15 minutes to complete. You will have an eight-week window to complete the survey. Reminder emails will be sent at week 4 and week 6.

Please note that completing the survey and clicking 'Submit' automatically implies your consent to participate. Participation is voluntary and you are free to withdraw at any point whilst completing the survey. However please note retraction or removal of individual survey answers is not possible once the 'Submit' button has been selected.

Contact information:

Should you have any questions concerning this study, please contact the research team using the details provided below:

Rachel Phillips

Imperial Clinical Trials Unit, Imperial College London, 1st Floor Stadium House, 68 Wood Lane, London, W12 7RH

Email: r.phillips@imperial.ac.uk

Tel: 020 759 49356

We thank you for your consideration to participate in this project.

Appendix A4.4: Clinical areas participants indicating they predominantly worked on

	Public (N=23)		Industry (N=11)		Overall (N=34)	
Clinical area	n	%	n	%	n	%
Adaptive designs*	1	4.3	0	0	1	2.9
Breast cancer	1	4.3	0	0	1	2.9
Cancer	2	8.7	0	0	2	5.9
Cancer, HIV, TB	1	4.3	0	0	1	2.9
Oncology	3	13	4	36.4	7	20.6
Oncology, neurosciences	0	0	1	9.1	1	2.9
Intensive care/oncology	1	4.3	0	0	1	2.9
Phase 1/2 Oncology	0	0	1	9.1	1	2.9
Skin diseases, oncology	0	0	1	9.1	1	2.9
Cardiovascular	0	0	1	9.1	1	2.9
Cochlear implants	0	0	1	9.1	1	2.9
Complex intervention trials*	1	4.3	0	0	1	2.9
Diabetes trial	1	4.3	0	0	1	2.9
Health services	1	4.3	0	0	1	2.9
Immuno-inflammation	0	0	1	9.1	1	2.9
Infectious diseases	1	4.3	0	0	1	2.9
Musculoskeletal disorders	1	4.3	0	0	1	2.9
Neonatal medicine	1	4.3	0	0	1	2.9
Non-CTIMP*	1	4.3	0	0	1	2.9
Ophthalmology	1	4.3	0	0	1	2.9
Pain	0	0	1	9.1	1	2.9
Mental Health	1	4.3	0	0	1	2.9
Primary care, mental health, rehabilitation	1	4.3	0	0	1	2.9
Primary care	1	4.3	0	0	1	2.9
Psychological interventions for people with chronic physical illnesses	1	4.3	0	0	1	2.9
Surgery	1	4.3	0	0	1	2.9
Transplantation and blood transfusion	1	4.3	0	0	1	2.9

*Participant(s) indicated a non-clinical area

Appendix A4.5: Free text comments regarding other information presented on emerging harm outcomes

Other information presented
We present as a proportion as ITT and also as proportion exposed (requirement for EudraCT). We present specific toxicities and the proportions at each grade.
Number of patients with at least one G3+ events Number of patients with at least one treatment emergent Events of special interest
maximum grade over treatment by subject
Number of participants by worst grade of event (CTCAE), time to specified toxicity event
Number of events by highest CTCAE grade
Frequency of worst CTCAE grade of each AE for each patient during the treatment and follow-up periods
More frequently reported Events by severity SAEs
Relatedness
Number of events presented only for overall summary of aes, teaes, related aes and aes leading to treatment discontinuation. No summary of number of aes by soc and pt
Numbers of patients experiencing 0, 1, 2, ... events
Dependant on the trial. Most commonly the "Number of participants with at least one event" (sometimes by different treatment periods if appropriate). For trials with lengthy "maintenance" type treatments we are moving away from this and may present things like number of AEs per patient or time experiencing certain events.
median number of events in both those experiencing at least one event and out of those randomised.
And percentage by group of course.
Dependant on the trial. Most commonly the "Number of participants with at least one event" (sometimes by different treatment periods if appropriate). For trials with lengthy "maintenance" type treatments we are moving away from this and may present things like number of AEs per patient or time experiencing certain events.
Proportions and %s, making clear what the denominators are
Sometimes both, depending on the AE
In a few occasions, the client asked for confidence intervals, or the prevalence of AEs tested across arms via a Fisher exact test. On only 1 trial in 17 years of time, time to onset analyses were required, with estimation of incidence rates abd associated CI, in person-years.
Rate over the periid of exposure.
Usually both of above and incidence rate. For some events we also include rate per 100 PY
exposure time in years + incidence rates (though this varies from study to study)
incidences per group, incidence rate ratios with uncertainty (depending on the situation)
competing risk analysis

Appendix A4.6: Free text comments regarding methods participants are aware of specifically for the analysis of harm outcomes

Bayesian approaches (n=1):
"Bayesian methods to analyse low frequency event data."
Modelling approaches (n=6):
"I don't think there is anything special about AEs/SAEs that require special methods. Statistical methods for the analysis of events (yes/no) or repeated events accounted for differential follow-up or/and overdispersion already exist in statistical literature (e.g., poisson or negative binomial regression model). of course, it depends on the underlying distribution"
"Classical Poisson/Negative Binomial/ZIP Regression for incidence rates"
"Extreme Value methods"
"…,survival analysis for comparison of treatment and for time to specific event"
"Survival methods"
"GEE"
Incidence rate (n=5):
"crude incidence rates, exposure-adjusted incidence rates, mean cumulative function (MCF)"
"Rate analyses,…"
"Cumulative incidence plots, life table plots, and prevalence plots. Many methods not specific to analysis of AEs"
"Incidence rates and confidence intervals (in person-years). Time to onset."
"Rate ratio,…"
Meta-analysis (n=2):
"…examples of meta analyses to appropriately analyse AE data"
" Meta analysis of Rare events"
Graphics (n=2):
"Cumulative incidence plots, life table plots, and prevalence plots. Many methods not specific to analysis of AEs"
"Graphics for biological parameters (ellipse ci)"
Theoretical and applied examples (n=6):
" CLEOPATRA Study Repeated Measures (i.e. not just counting first event)"
"Various methods published by Harry Southworth. These are predominantly useful for pharma trials rather than Phase 4 trials unit trials."
"Volume15, Issue4 Special Issue: Analysis of Adverse Event Data July/August 2016 Pages 297-305"
"http://dx.doi.org/10.1136/bmj.i5078"
"https://onlinelibrary.wiley.com/toc/15391612/2016/15/4"
"possible use of estimands to analyse AEs (for example https://arxiv.org/abs/1805.01834)"
Other comments:
"Not meaningfully within an early phase setting, because of sample size. Monitoring based approaches are becoming used and machine learning based methods are available."
"AE tables and summary"
"The statistical literature is awash with methods"
"zz"

Appendix A4.7: Free text comments regarding participants' use of specialist methods for analysis of emerging harm outcomes

Time to event analysis (n=2):
"In characterising safety signals I have used Time to Event, Event rates, prevalence."
"Time-to-event analyses; exposure-adjusted AE rates"
Data visualisations (n=1):
"Data visualisation (which is more or less equivalent to frequencies and percentages)"
Bayesian methods
"Bayesian methods for sparse adverse events data meta-analysis"
Incorporating repeated event (n=1):
"For within-patient repeated events we have produced comparisons with a 2-d frequency table (arm vs # events)"
Other comments:
"Not sure I understood what is meant by specialist AE analysis. I used various statistical methods depending on the situation."
"Safety analysis in phase III cancer clinical trial"

Appendix A5.1: Consensus group attendees

Participant name	Affiliation
Rachel Phillips	Imperial Clinical Trials Unit, Imperial College London
Victoria Cornelius	Imperial Clinical Trials Unit, Imperial College London
Suzie Cro	Imperial Clinical Trials Unit, Imperial College London
Catherin Hewitt	York Trials Unit, University of York
Graham Wheeler	Cancer Research UK Cancer Trials Centre, University College London
Andre Lopes	Cancer Research UK Cancer Trials Centre, University College London
Simon Bond	Cambridge Clinical Trials Unit, Cambridge University Hospitals NHS Foundation Trust
Chris Harbron	Roche Pharmaceuticals
Carrol Gamble	Liverpool Clinical Trials Centre, University of Liverpool
Graeme MacLennan	Centre for Healthcare Randomised Trials, University of Aberdeen
Tim Morris	MRC Clinical Trials Unit, University College London
Sharon Love	Institute of Clinical Trials & Methodology, University College London
Sarah Pirrie	Cancer Research UK Clinical Trials Unit, University of Birmingham
Colin Everett	Clinical Trials Research Unit, University of Leeds
Rachel Evans	North Wales Organisation for Randomised Trials in Health, Bangor University
Jane Holmes	Oxford Clinical Trials Research Unit, University of Oxford
Siobhan Creanor	Exeter Clinical Trials Unit, University of Exeter
Clare Peckitt	Royal Marsden Clinical Trials Unit, The Royal Marsden NHS Foundation Trust
Laura Collett	Bristol Clinical Trials and Evaluation Unit, University of Bristol
Norin Ahmed	Comprehensive Clinical Trials Unit, University College London
Iryna Schlackow	Nuffield Department of Population Health, University of Oxford
Amanda Kirkham	Cancer Research UK Clinical Trials Unit, University of Birmingham
Will Stahl-Timmins	BMJ

Appendix A5.2: Plots for consideration - multiple binary outcomes

Figure A5.1 Volcano plot

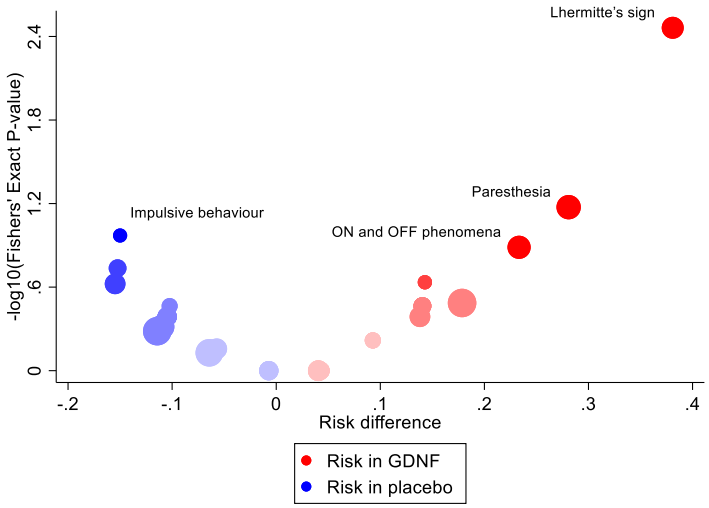


Figure description: Each bubble/circle represents a distinct AE. Bubble size/area is proportional to the total number of AEs across treatment arms. The x-axis indicates the size of the treatment effect. The colour of the bubbles is used to indicate the direction of the treatment effect. The y-axis is used to display log transformed p-values. The colour saturation of each bubble corresponds to the size of the p-value. Under the null hypothesis we would expect to see a U-shape curve with a random scatter of events around the null value, in this case a risk-difference of 0. *Original plot first proposed in: Zink RC, Wolfinger RD and Mann G. Summarizing the incidence of adverse events using volcano plots and time intervals. Clinical Trials 2013; 10: 398-406. Data taken from: Whone A, Luz M, Boca M, et al. Randomized trial of intermittent intraputamenal glial cell line-derived neurotrophic factor in Parkinson's disease. Brain 2019; 142: 512-525*

Figure A5.2: Alternative volcano 1 proposed by BMJ graphic designer

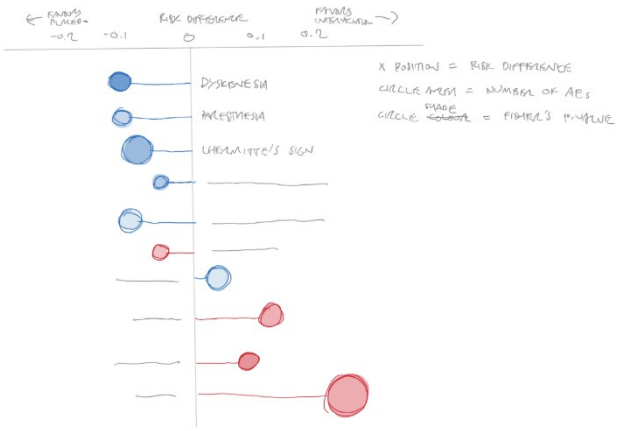


Figure description: Displays the risk difference across the x-axis. The direction of treatment effect is indicated by colour. The size of the p-value is indicated by the colour shade/saturation. Total number of events indicated by the circle area. Allows incorporation of labels for all events.

Figure A5.3 Alternative volcano 2 proposed by BMJ graphic designer

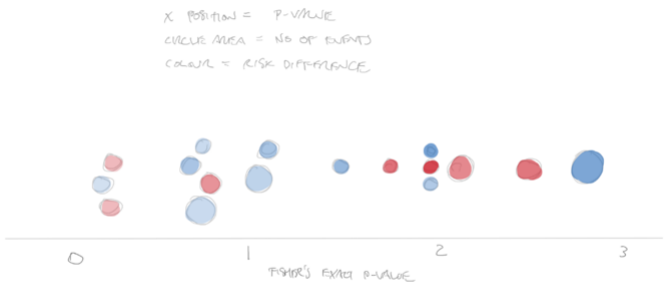


Figure description: The size of the risk difference is reflected in the colour saturation of each circle. The direction of the treatment effect is indicated by the colour of the circles. The size of the p-value is displayed across the x-axis. The total number of events is indicated by the circle area. The y-axis is not used to display a metric but instead used to stack circles/bubbles to prevent overlap.

Figure A5.4: Alternative volcano 3 proposed by BMJ graphic designer

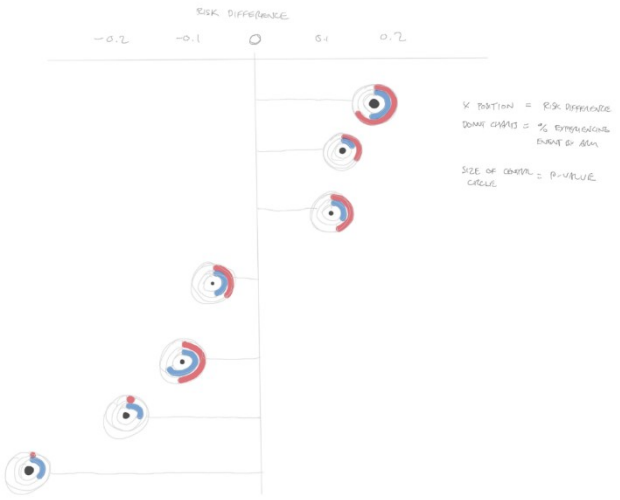


Figure description: Shows size and direction of treatment effect across the x-axis. The size of the p-value is indicated by the size of the central black circle. The number of events is represented by the proportion of outer segments (or donuts) that are shaded, with different colours for each treatment arm. Each event takes up a 'row' along the y-axis. BMJ graphics designer (Will Stahl Timmins) proposed this as an idea but suggests needs further refinement.

Figure A5.5: Dot plot

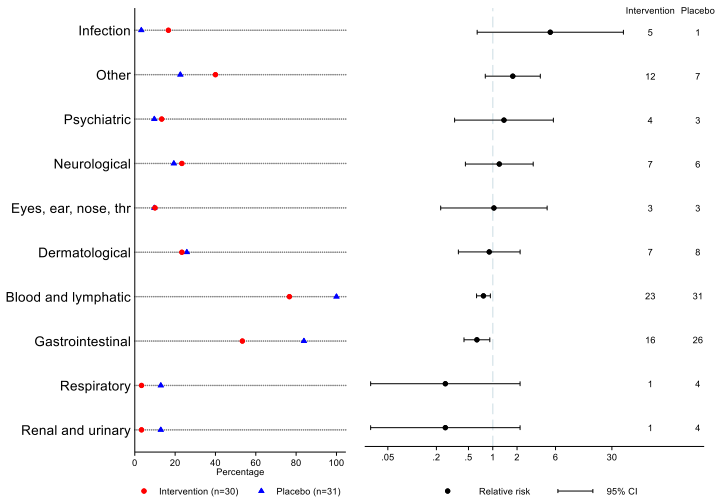


Figure description: On the left we display the percentages with each event by treatment arm. The space between the markers on the left gives a visual indication of the absolute risk differences for each event. In the center a relative difference is displayed, in this example the relative risk is displayed with the corresponding 95% CI. On the far right we have adapted this plot to display the number of participants with at least one event by treatment arm. *Original plot first proposed in: Amit O, Heiberger RM and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. Pharmaceutical Statistics 2008; 7: 20-35.*

Figure A5.6 Bar chart

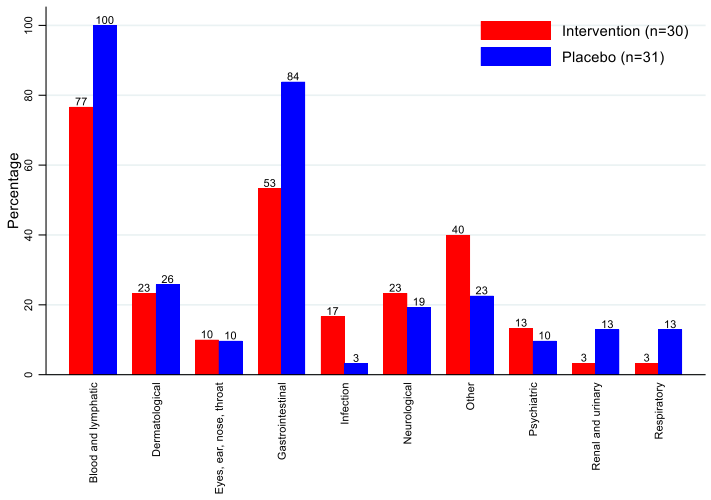


Figure description: Each bar represents the percentage of participants that experience an event at least once. Colour of the bars indicates treatment arm. Includes percentage labels at the top of each bar. Event names labelled along the x-axis.

Figure A5.7 Tendril plot

Figure 2. A tendril plot of all AEs in the RE^{2}SPOND trial. Each MedDRA PT is represented by a line (tendril) and each point is an event. Since time runs along each tendril, it is the shape that carries the important information, rather than the *x* and *y* coordinates. An event on the roflumilast treatment arm will tilt tendril direction to the left, and an event on the placebo arm will tilt tendril direction to the right. Point size is indicative of the total number of events for the type of AE in both treatment arms in the trial. The FDR adjusted Pearson's chi-squared *P*-value in each point is mapped onto a continuous color gradient.

Figure description: Each AE term is represented by a line (or tendril). Each point on the line indicates the occurrence of an event. The distance from the origin indicates the time the event occurred. The direction or tilt of the line is used to indicate the treatment arm the event occurred in i.e. the line takes a unit tilt to the left for an event in the intervention arm and a unit tilt to the right for an event in the control arm. The colour of the points along the lines indicate the size of the p-value. Reprinted from: *Karpefors, M. and J. Weatherall (2018). "The Tendril Plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials." Journal of the American Medical Informatics Association* ***25****(8): 1069-1073 with permission of Oxford University Press.*

Figure A5.8 Heat map

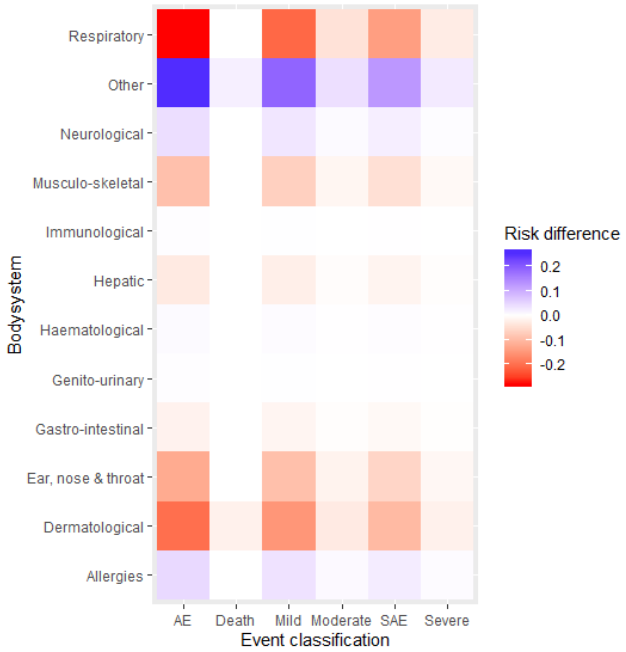


Figure description: Individual event names are displayed along the y-axis. Different event classifications such as severity ratings are displayed across the x-axis. Each x/y-axis grid position/square is coloured to represent a treatment effect for a unique event and classification. The colour of the squares/grid is used to indicate the direction of the (standardised) treatment effect. Colour saturation of the squares/grid is used to indicate size of effect for each AE. *Original plot first proposed in: Zink, R. C., et al. (2018). "Sources of Safety Data and Statistical Strategies for Design and Analysis: Clinical Trials." Therapeutic Innovation & Regulatory Science* ***52****(2): 141-158.*

Figure A5.9a Level plot - Originally proposed for categories of abnormal blood tests

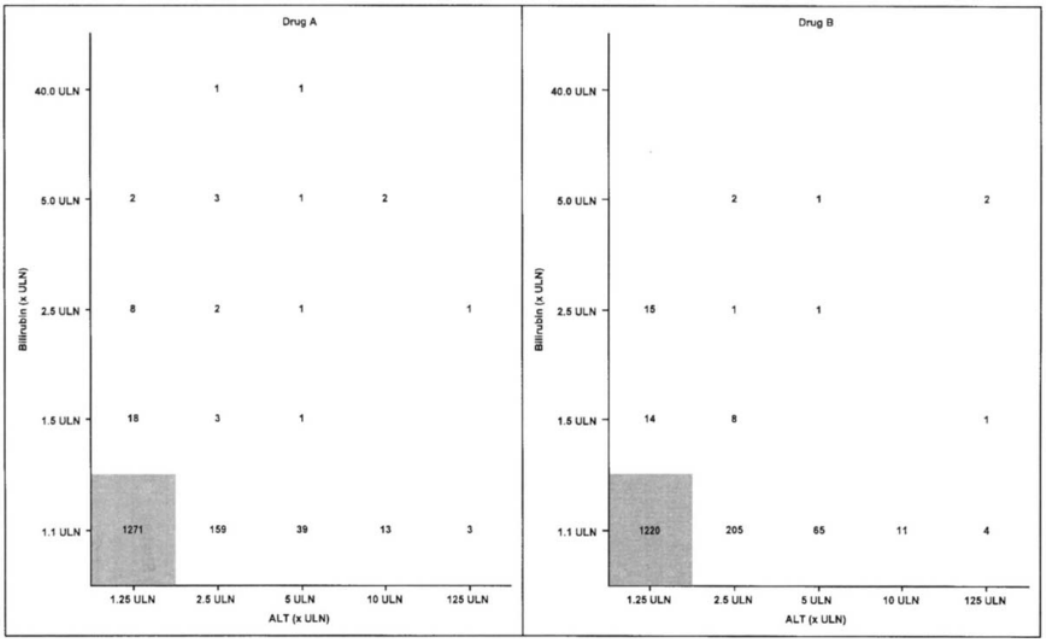


FIGURE 7. A simultaneous grade change in ALT and bilirubin.

Figure description: Displays categories of two different blood tests on the x and y-axes. Displays counts of participants in the intersection of categories.

Reprinted from: Chuang-Stein, C., et al. (2001). "Recent Advancements in the Analysis and Presentation of Safety Data." Drug Information Journal 35(2): 377-397 under the terms of the Creative Commons CC BY License

Figure A5.9b Level plot - An adaption similar to this from Ballarini et al. could provide potentially useful modifications for the heat map

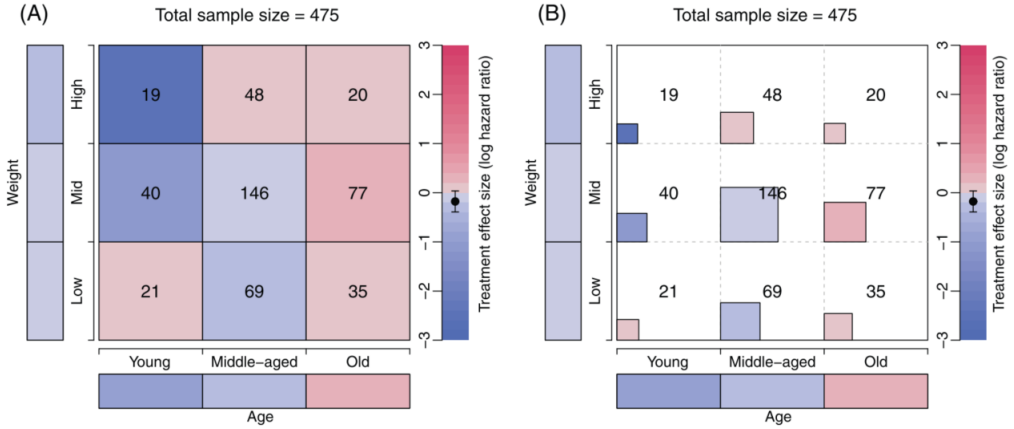
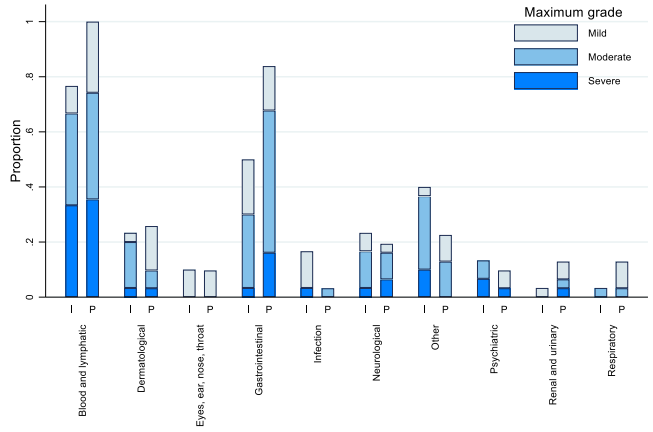


FIGURE A2 Level plots of treatment effect in terms of the log-hazard ratio across mutually disjoint subgroups defined by *age* and *weight* categorised in three levels. The cells on the bottom and the left margins correspond to the marginal subgroups defined by the levels of *age* and *weight*. In B, the area of each square inside the cells is proportional to the sample sizes, which are also displayed in the middle of the cells

Reprinted from: Ballarini, NM, Chiu, Y-D, König, F, Posch, M, Jaki, T. A critical review of graphics for subgroup analyses in clinical trials. Pharmaceutical Statistics. 2020; 1– 20, https://doi.org/10.1002/pst.2012 under the terms of the Creative Commons CC BY License.

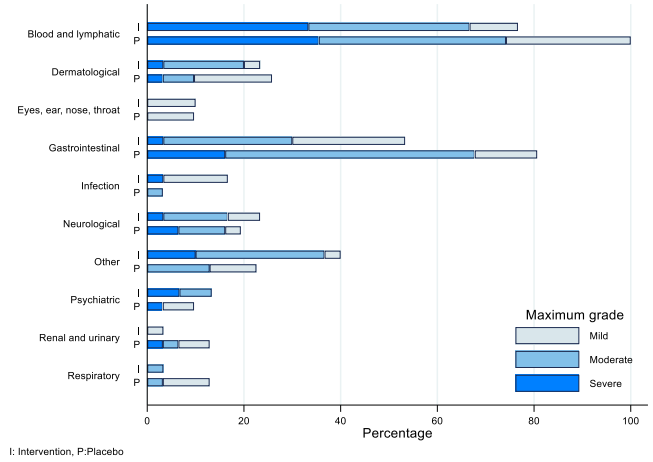
Figure A5.10a Stacked bar charts of maximum severity



I: Intervention, P:Placebo

Figure description: Each bar represents the proportion of participants with an event by maximum grade i.e. if a participant had the same event twice, once classified as mild and once as moderate this participant would be counted as experiencing a moderate event. Bars are split by colour gradient to indicate different severity groups and the total bar height tells us the proportion of patients experiencing that event at least once. The most severe category is displayed first to allow ease of comparison for the most harmful/burdensome events.

Figure A5.10b: Horizontal stacked bar charts of maximum severity



I: Intervention, P:Placebo

Figure description: As previous but with horizontal stacked bars

Figure A5.10c Pyramid stacked bar chart

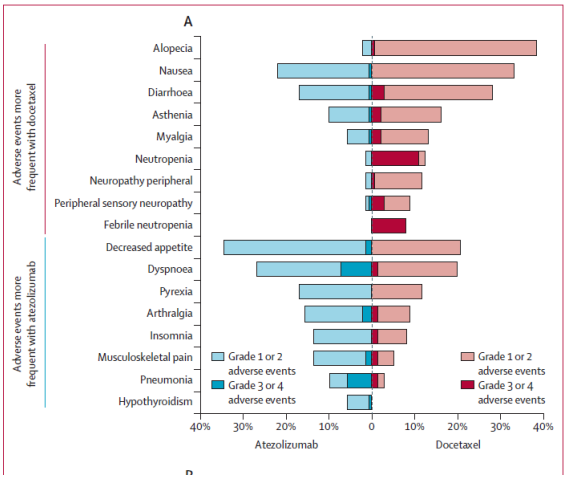


Figure description: As previous with pyramid style bars. The bars in the top segment of the plot indicates events more frequent in the intervention arm and bars in the bottom segment indicate events more frequent in the control arm. *Reprinted from: Fehrenbacher, L., et al. (2016). "Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial." The Lancet **387**(10030): 1837-1846 with permission from Elsevier.*

Figure A5.11a: Stacked bar charts of counts

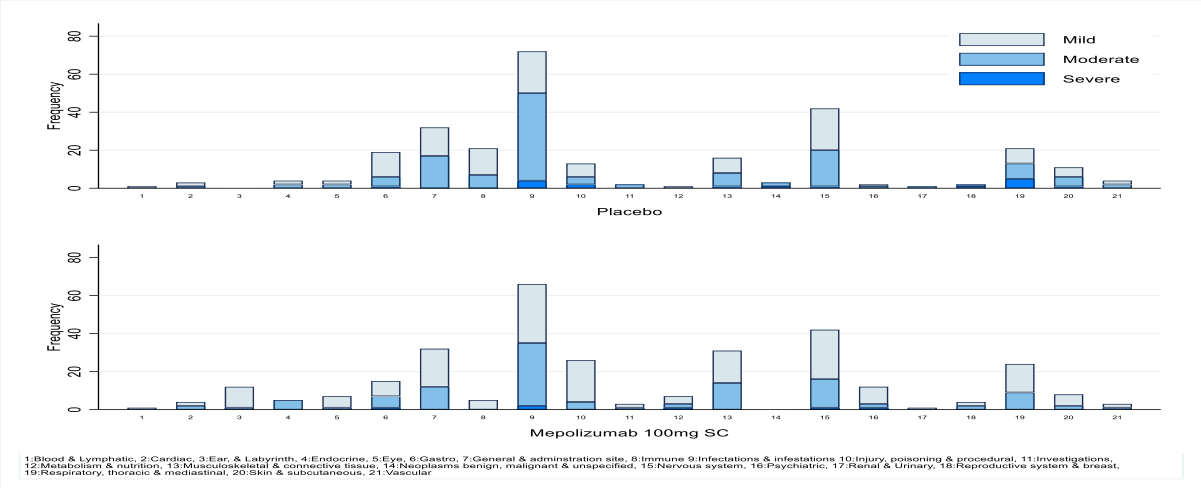


Figure description: Separate plot for each treatment arm. Each bar represents an event (with event names in the legend at the very bottom of the plot). Bar height indicates total counts (multiple events per participant). Bars are stacked by severity. Most severe category placed at the bottom for ease of comparison. *Original plot first proposed in: Chuang-Stein, C. and H. A. Xia (2013). "The practice of pre-marketing safety assessment in drug development." Journal of Biopharmaceutical Statistics **23**(1): 3-25.*

Figure A5.11b Stacked bar chart with count labels

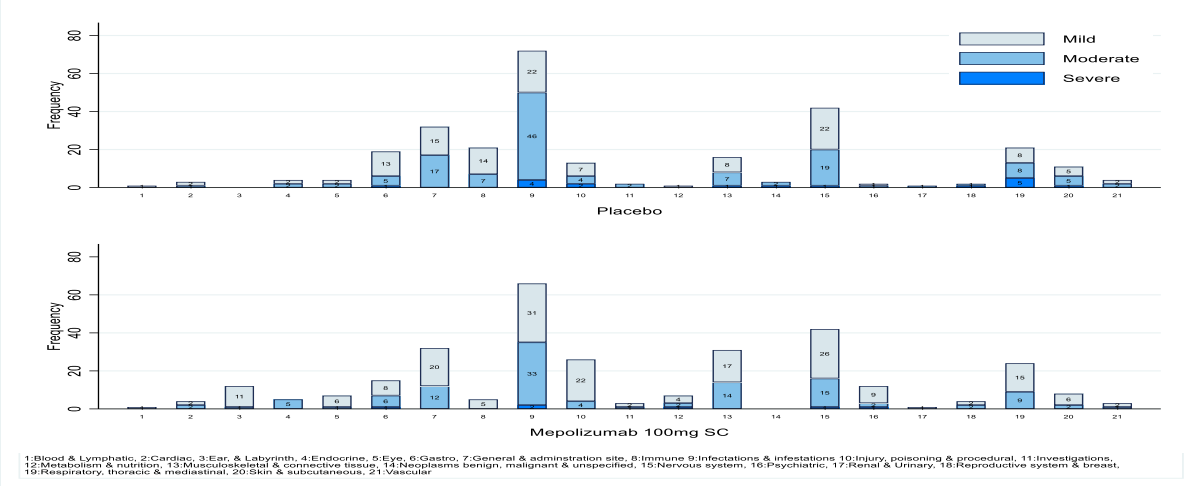


Figure description: As above but includes labels for counts of events.

Figure A5.11c: Horizontal stacked bar chart of counts

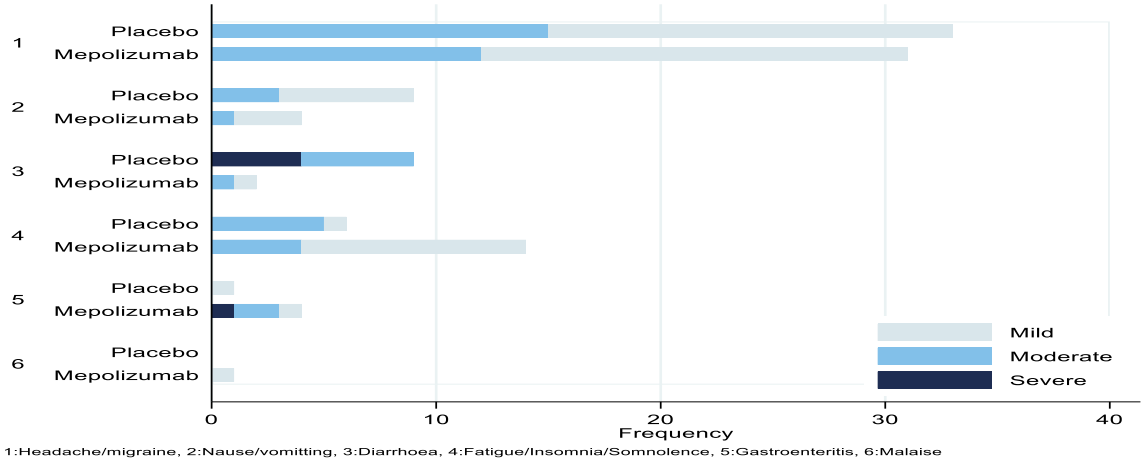


Figure description: As previous but with horizontal stacked bars

The following have not been specifically proposed for AE data analysis but have been suggested as potentially useful:

Figure A5.12: Star plot

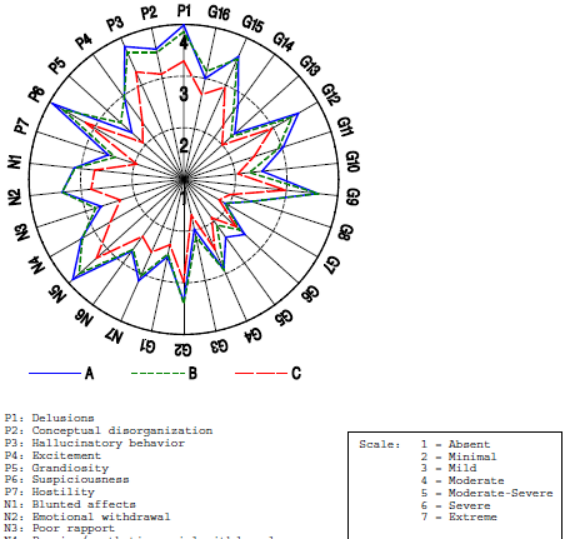


Figure description: Displays mean values for multiple clinical results (in this example each of the 30 PANSS items rated on a Likert scale). The coloured lines represent different treatment arms. Concentric reference lines included to help read off values. Could be a**dapted to** present mean grade for each AE by treatment arm. *Thanks to Steven Julious at Sheffield University for flagging this plot. Reprinted from: Squassante et al. Simple graphical methods of displaying multiple clinical results. Pharmaceut. Statist. 2006; 5: 51–60 with permission from John Wiley & Son.*

Figure A5.13: Alluvial plot

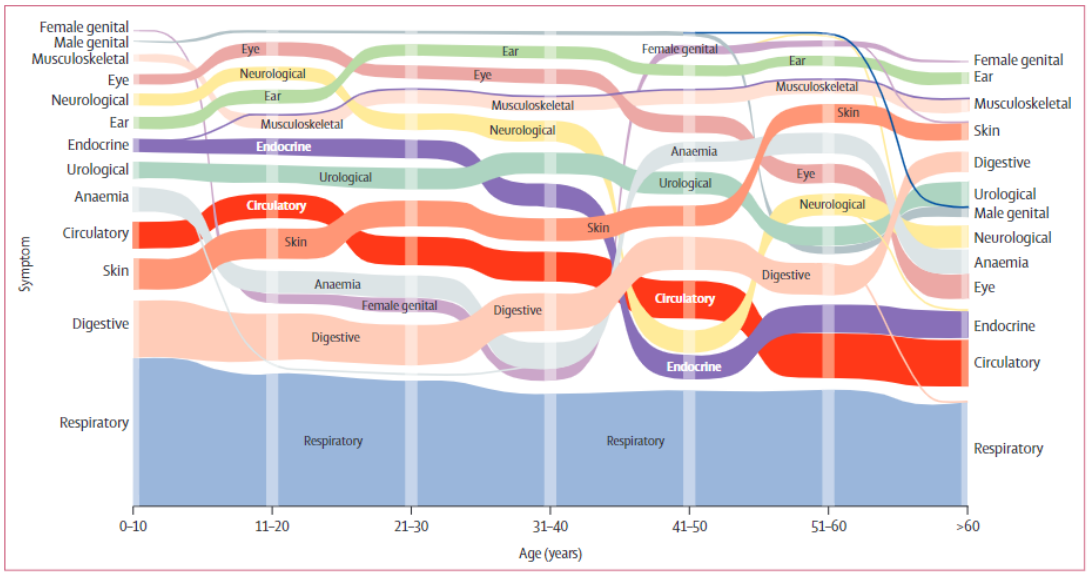


Figure 4: **Alluvial graph of disease or symptom prevalence across decades of life**
Width of bands corresponds to relative proportion of symptoms or medical conditions, with the y axis organised to have the most prevalent conditions closest to the x axis. Associations between symptoms or conditions are represented as offshoots that connect systems. The light shaded vertical bars correspond to the decade-wise groupings used in analysis..

Figure description: Shows how the proportions experiencing an event (for multiple events) change over time. Could be adapted to show changes in severity categories over time for a single event, would be tricky to do this for multiple events. Would need to produce a separate plot for each arm to make a comparison between treatment arms. T*hanks to Marianna Nodale at Cambridge University Hospital for suggesting this image.* *Reprinted from: Salvi S, Apte K, Madas S, et al. Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: a 1-day point prevalence study (the POSEIDON study). Lancet Glob Health. 2015;3(12):e776-e784. doi:10.1016/S2214-109X(15)00152-7 under the terms of the Creative Commons CC BY NC ND License*

Appendix A5.3: Plots for consideration - multiple time-to-event outcomes

Figure A5.14 Matrix of Nelson-Aalen cumulative hazards

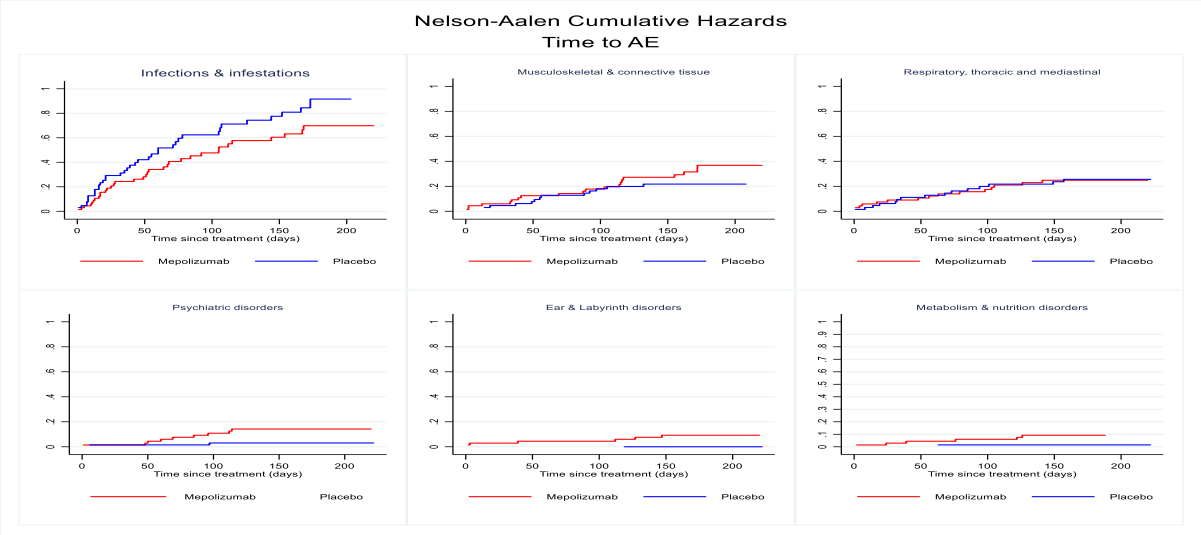


Figure description: Cumulative hazards by treatment arm for multiple events. Treatment arms represented by different colours. Alternative time-to-event plots such a matrix of Kaplan-Meier plots.

Figure A5.15: Alternative survival plot 1 proposed by BMJ graphic designer

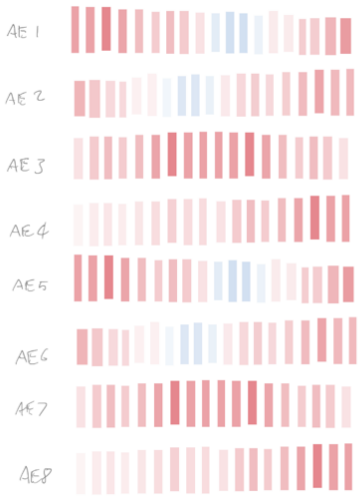


Figure description: Individual AEs displayed along the y-axis. X-axis represents time. Colour used to indicate direction of treatment effect. Colour saturation/intensity is used to reflect size of "difference" between arms. This "difference" is a difference in the cumulative number of events over time. Blocks of colour represent the difference for each unit of time for each event. Needs further development before it could be recommended.

Figure A5.16 Alternative survival plot 2 proposed by BMJ graphic designer for time to withdrawal

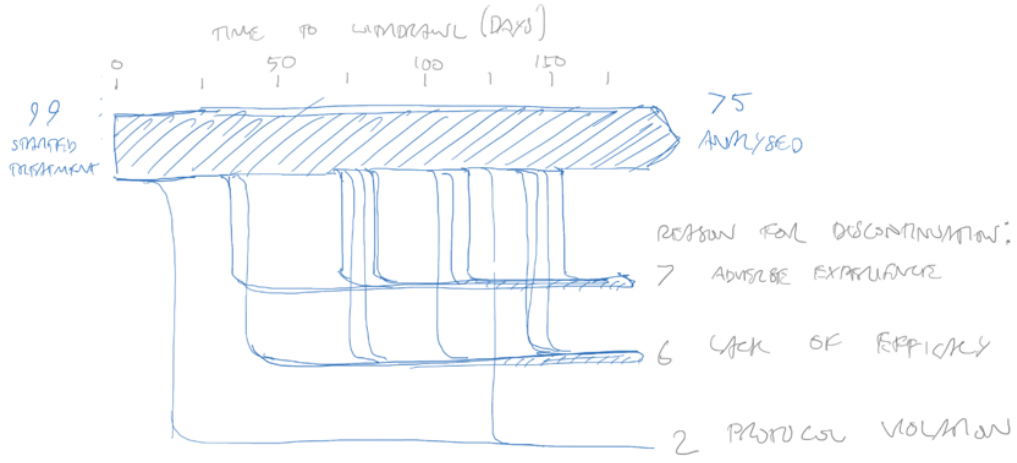


Figure description: Commonly known as a Sankey diagram. Represents the time each participant withdraws with an arrow out of the main horizontal arrow. Arrows flow into different states to reflect the reasons for withdrawals. Would need one image per treatment arm or an adaption to incorporate each treatment arm into the one image.

Figure A5.17 Bar chart of median time to event

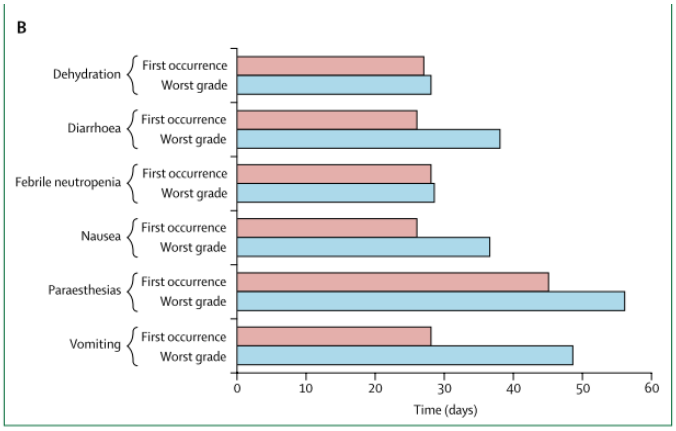


Figure 3: **Time-to-event analyses for adverse events using the Toxicity over Time package**
(A) Time to grade 2 or worse diarrhoea in patients given FOLFOX and IROX in NCCTG N9741.[36] (B) Median time to first occurrence and worst grade toxic effects in patients given IROX in NCCTG N9741.[36] The figures capture the time profile of adverse events from these regimens. IROX=irinotecan and oxaliplatin. FOLFOX=leucovorin, fluorouracil, and oxaliplatin.

Figure description: Horizontal bar graph of median time to first (and worst grade) event. Height/length of each bar represents the median time to event. Different events are displayed along the y-axis. Time is displayed on the x-axis. Use separate bars for each treatment arm instead of first and worst event. **Caution:** This is taken from a publication in the Lancet Oncology but we think it could be very misleading since: it doesn't account for censoring or show how the denominator changes over time; and it doesn't include any information on the number of participants that have these events. *Reprinted from: Thanarajasingam, G., et al. (2018). "Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies." The Lancet Haematology **5**(11): e563-e598 with permission from Elsevier.*

Figure A5.18 Alternative to the bar chart of median time to event proposed by BMJ graph designer

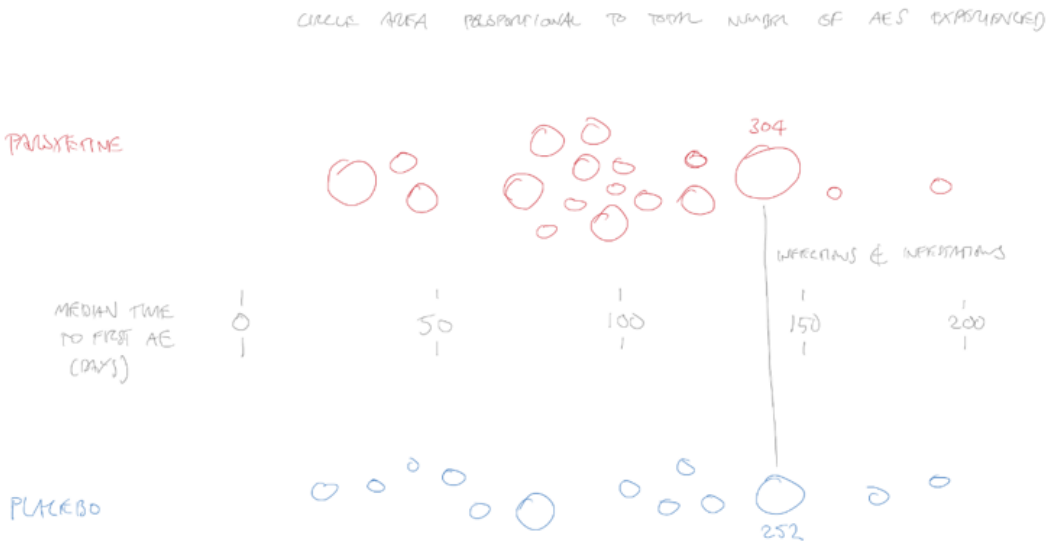


Figure description: Colour used to indicate treatment arm. Position along the x-axis is used to indicate median time of occurrence. Each circle represents an event. The circle area is proportional to the number of AEs experienced.

Appendix A5.4: Plots for consideration - Single binary outcomes

Figure A5.19: Bar chart of frequency of counts

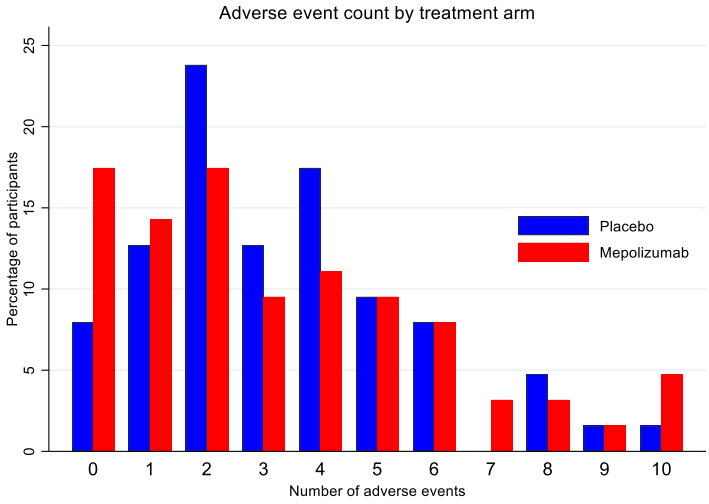


Figure description: Bars represent percentage of participants with 0, 1, 2 etc. events by treatment arm. Treatment arms represented with different colours. Percentage of participants with each number of events within treatment arms.

Appendix A5.5: Plots for consideration - Single continuous outcomes

Figure A5.20: Empirical distribution of maximum change

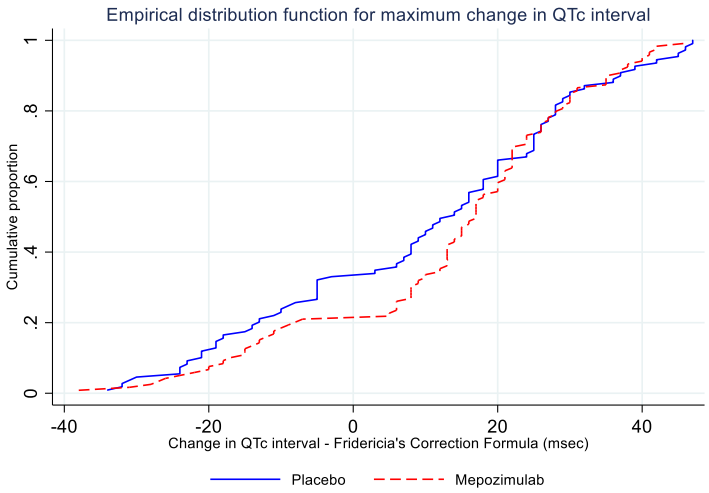


Figure description: Displays the cumulative proportion of participants on the y-axis with a change in QTc less than or equal to the corresponding value on the x-axis. Displays maximum change for each participant. Treatment arms displayed in different colours. *Original plot first proposed in: Amit, O., et al. (2008). "Graphical approaches to the analysis of safety data from clinical trials." Pharmaceutical Statistics **7**(1): 20-35.*

Figure A5.21: Line graph of change values

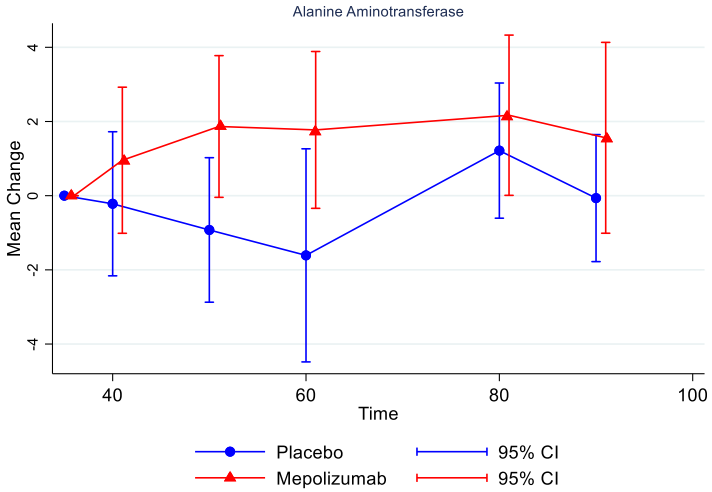


Figure description: Line graph of mean change from baseline with corresponding 95% CI. Treatment arms displayed in different colours. Could present alternative summary statistics such as median changes or mixed effect model estimates.

Figure A5.22: Box plot of change values

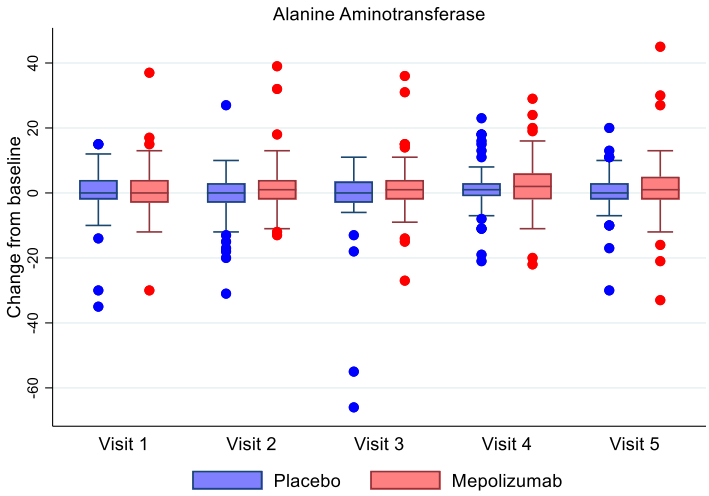


Figure description: Box plot of change from baseline across visits by treatment arm. Treatment arms displayed in different colours.

Figure A5.23: Violin plot of change values

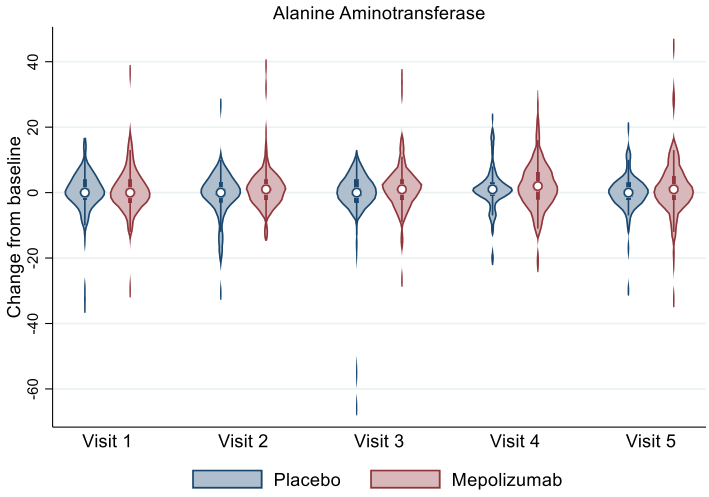


Figure description: Violin plot of change from baseline. Incorporates summary statistics as well as indicating density or frequency of values. Treatment arms displayed in different colours.

Figure A5.24: Histogram of change from baseline - stacked

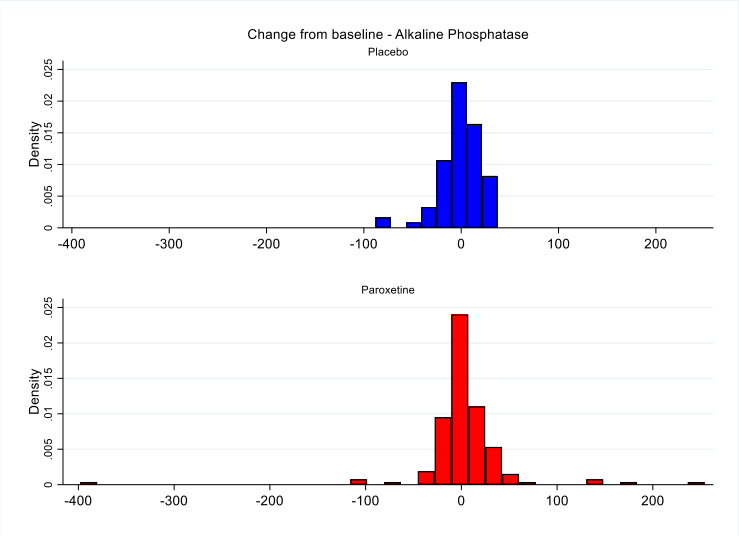


Figure description: Histogram of change from baseline. Plot for each treatment arm stacked above each other

Figure A5.25: Histogram of change from baseline - overlaid

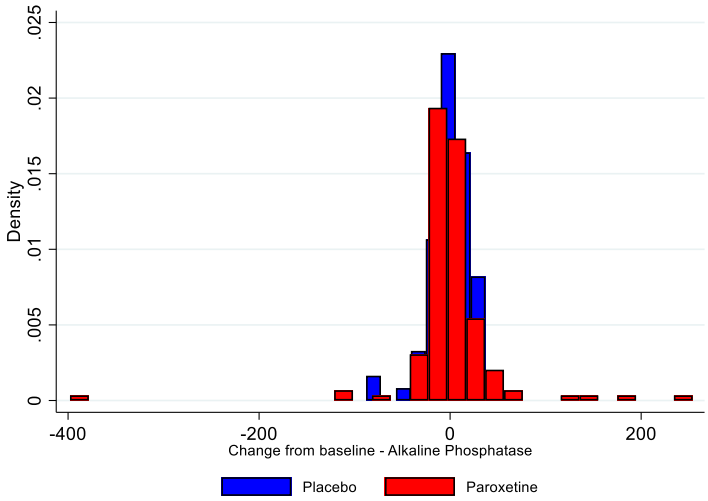


Figure description: As above but with treatment arms overlaid on top of each other.

Figure A5.26 Line graph of raw values

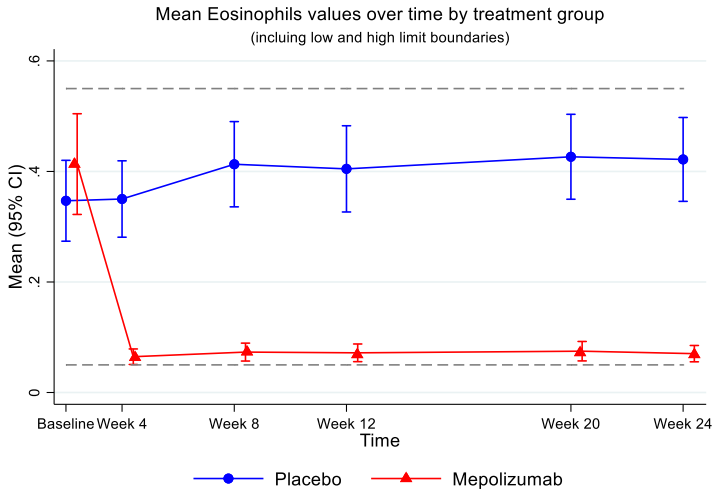


Figure description: Line graph of mean values at each visit with 95% CI. Treatment arms displayed in different colours. Includes reference/normal range as dashed lines. Could use mixed effects models to produce standard errors/95% confidence intervals. Instead of laboratory values could display mean grade of events over time.

Figure A5.27: Box plot of raw values

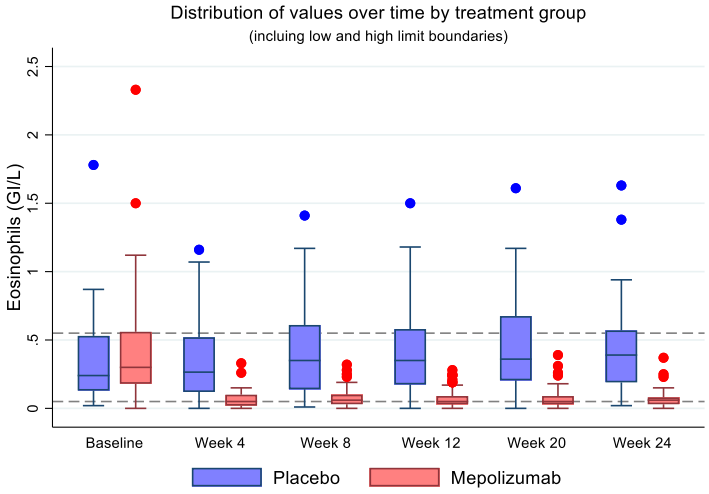


Figure description: Box plots of values at each visit. Treatment arms displayed in different colours. Includes reference/normal range.

Figure A5.28: Violin plot of raw values

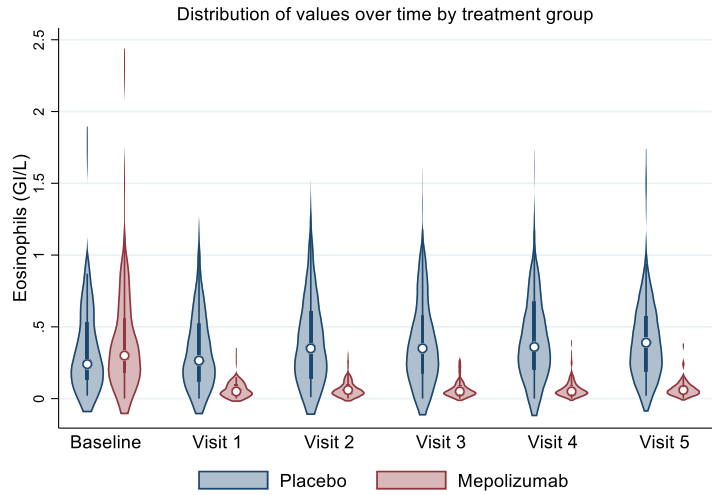


Figure description: Violin plots of values at each visit. Incorporates summary statistics as well as indicating density or frequency of values. Treatment arms displayed in different colour

Figure A5.29: Delta plot

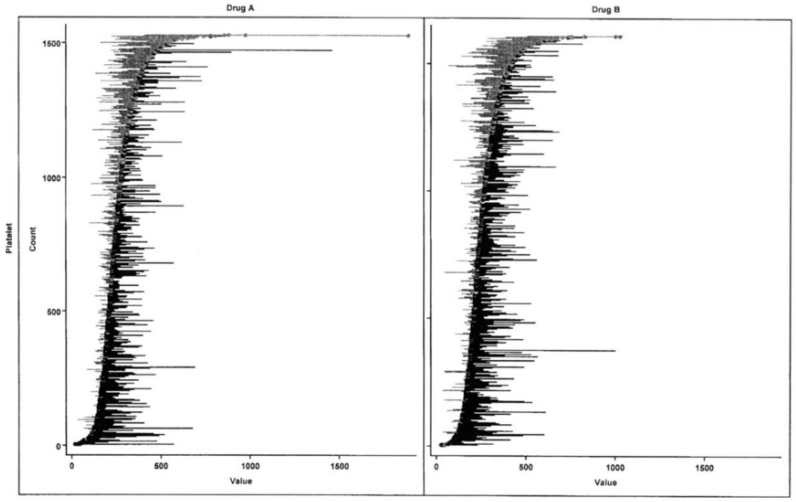


FIGURE 2a. Side-by-side Delta plots of baseline and last follow-up platelet values.

Figure description: Displays individual participant changes. The ends of each line indicate baseline and last follow-up values read from the x-axis for individual participants. Arranged according to baseline values. Y-axis tracks cumulative number of lines/participants. **Caution**: We don't find this plot very intuitive/helpful but included for comprehension. *Reprinted from: Chuang-Stein, C., et al. (2001). "Recent Advancements in the Analysis and Presentation of Safety Data." Drug Information Journal 35(2): 377-397 under the terms of the Creative Commons CC BY License.*

Appendix A5.6: Plots for consideration - single time-to-event outcomes

Figure A5.30: Stacked bar chart

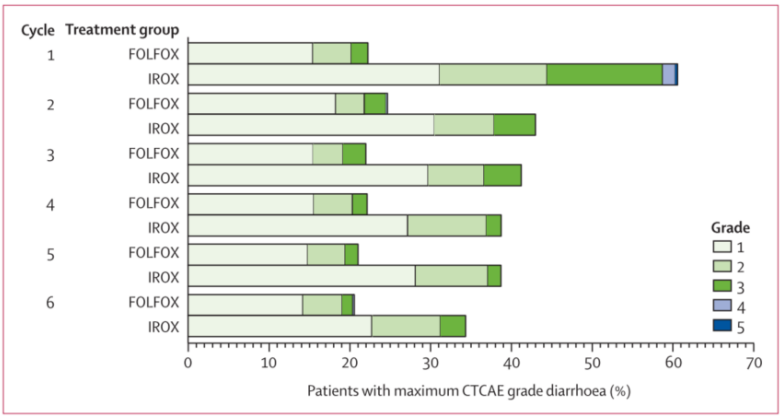


Figure 1: **Incidence of diarrhoea in patients given FOLFOX and IROX in NCCTG N9741 by drug cycle and adverse event grade**
FOLFOX=5-fluorouracil plus oxaliplatin. IROX=irinotecan plus oxaliplatin. CTCAE=Common Terminology Criteria for Adverse Events.

Figure description: Percentage of participants with an event stacked by maximum severity in each treatment cycle. Each bar represents the treatment cycle that the event occurred in and treatment arm. Could be adapted so that instead of treatment cycles could use visits or time periods. *Image taken from: Thanarajasingam, G., et al. (2016). "Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254." Lancet Oncology **17**(5): 663-670 with permission with permission from Elsevier.*

Figure A5.31: Histogram of counts over time

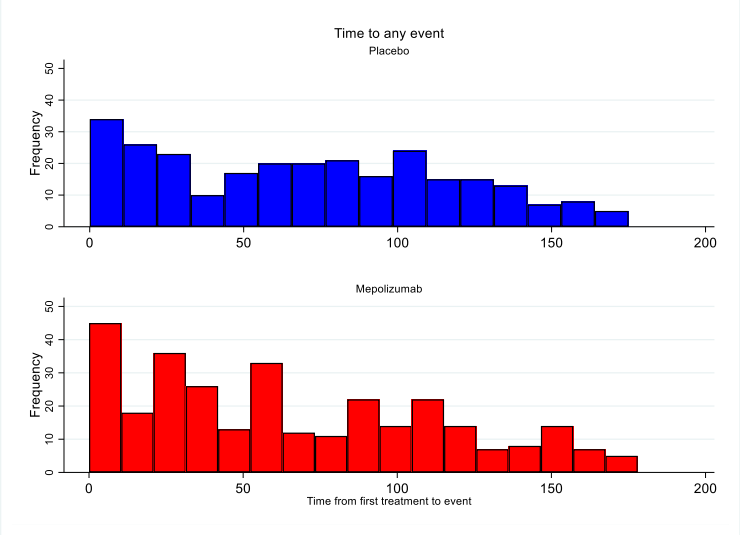


Figure description: Histogram of time of events (rather than categorising into arbitrary time periods). Not just looking at time-to-first event or maximum event, includes time of every event.

Figure A5.32 Nelson-Aalen cumulative hazards

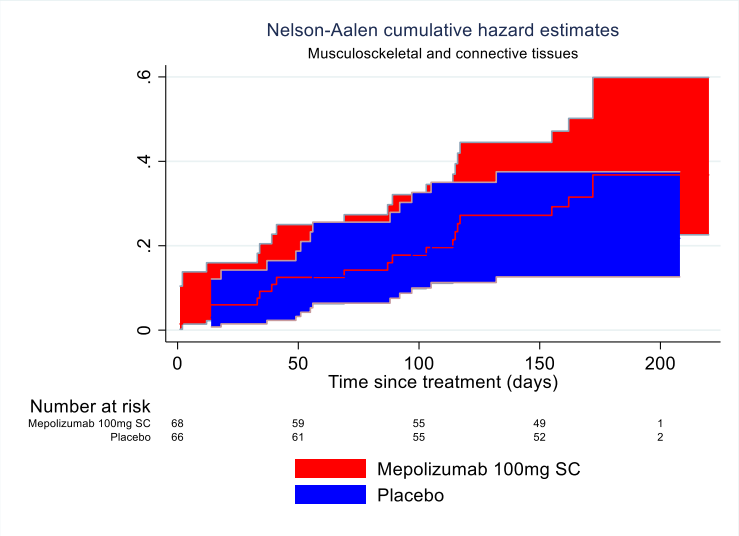


Figure description: Cumulative hazards by treatment arm with 95% confidence intervals and a table of numbers at risk

Figure A5.33 Kaplan-Meier

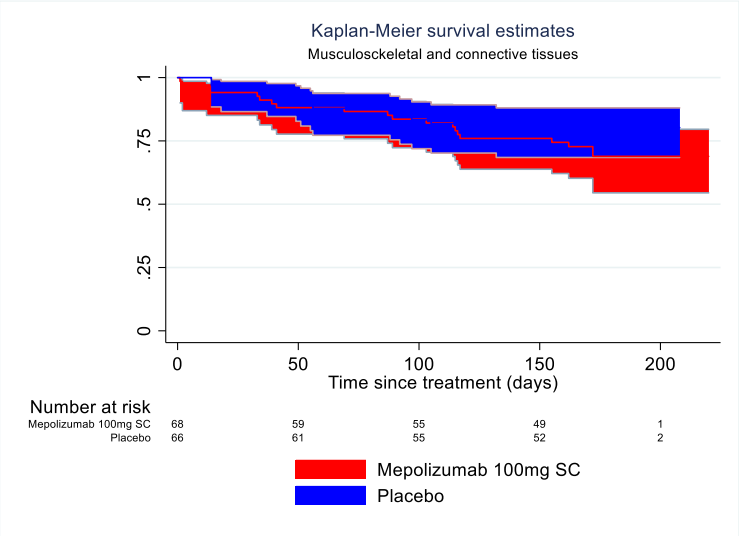


Figure description: Kaplan-Meier plot of survival estimates by treatment arm with 95% confidence intervals and at risk table. Treatments arms displayed in different colours.

Figure A5.34: Mean cumulative function

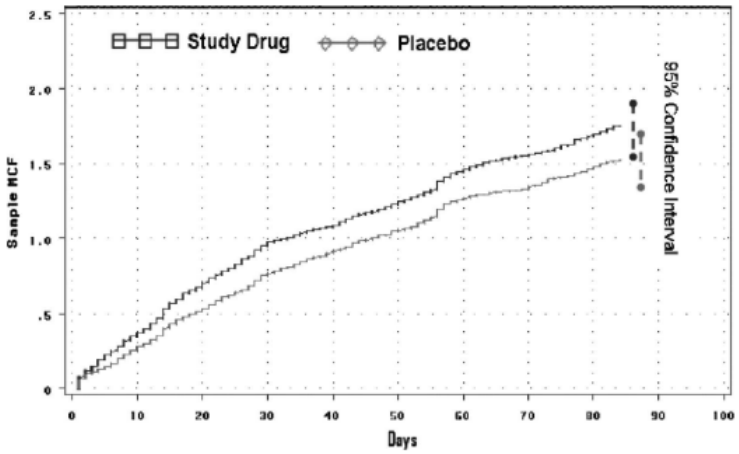


Figure 2 MCF of all AEs.

Figure description: Displays the mean cumulative function (MCF) as a function of time by treatment arm. Includes 95% CI for final time point. The MCF is a non-parametric estimate of the mean cumulative number of events per participant. Includes repeated events per participant. *Reprinted from: Siddiqui, O. (2009). "Statistical methods to analyze adverse events data of randomized clinical trials." Journal of Biopharmaceutical Statistics **19**(5): 889-899 with permission from Taylor & Francis.*

Figure A5.35: Mean cumulative duration

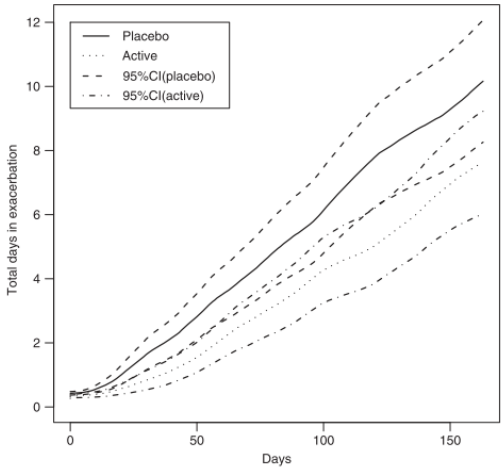


Figure 1 Mean duration of exacerbation with 95% CI based on the robust variance estimate for recurrent pulmonary exacerbations in patients with fibrosis treated with placebo and rhDNase (Fuchs et al., 1994).

Figure description: Displays the mean cumulative duration (MCD) as a function of time by treatment arm. The MCD is a non-parametric estimate of the mean cumulative duration of events per participant. Accounts for repeated occurrence of an event in a participant. Includes 95% confidence interval bands across follow-up. *Reprinted from: Wang, J. and G. Quartey (2012). "Nonparametric estimation for cumulative duration of adverse events." Biometrical Journal **54**(1): 61-74 with permission from John Wiley & Sons.*

Appendix A5.7: Plots for consideration - multiple continuous outcomes

Figure A5.36: Scatterplot matrix

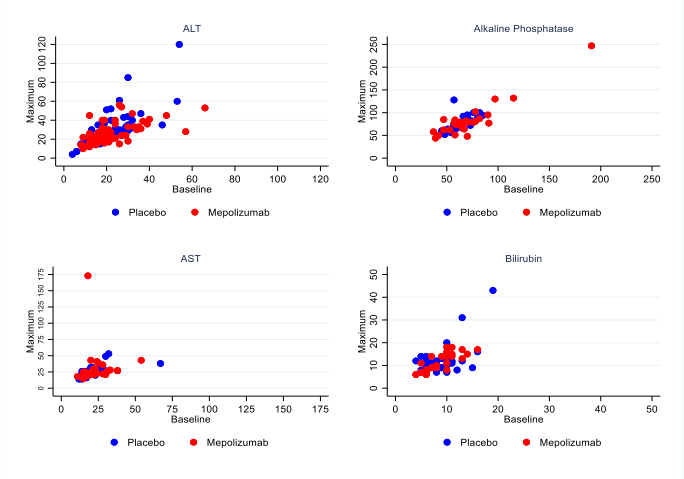


Figure description: Matrix of scatterplots. Plots baseline laboratory values against the maximum on treatment values. Treatment arms represented by different colours. **Note**: need to consider feasibility of including this in a journal article. Need to consider where, if anywhere, we would advise using such an image.

Figure A5.37: E-dish plot

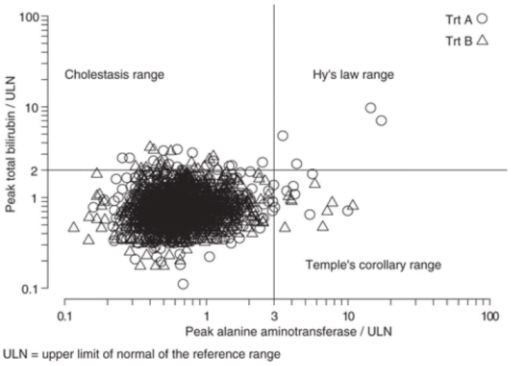


Figure 2 e-DISH-like plot. Notes: Plot of peak bilirubin (/ULN) *vs* peak ALT (/ULN). This figure shows peak values for total bilirubin and aminotransferases by treatment groups. Significant elevations of aminotransferases ('Temple's Corollary range') and, especially, abnormalities in the 'Hy's Law Range' should be carefully analyzed as potential signals for drug-induced liver injury [11]. Please see Refs [12] and [13] for further information on the use of this plot

Figure description: Specific scatterplot for maximum ALT, AST & Bilirubin values. Plots peak bilirubin vs peak ALT or AST. **Note**: Again, we need to consider where, if anywhere, we would advise using such an image. Perhaps better suited to monitoring of ongoing trials. *Reprinted from: Xia HA, Crowe BJ, Schriver RC, Oster M, Hall DB. Planning and core analyses for periodic aggregate safety data reviews. Clin Trials. 2011;8(2):175-182. doi:10.1177/1740774510395635 with permission from Sage Publishing.*

Figure A5.38: Vector plot

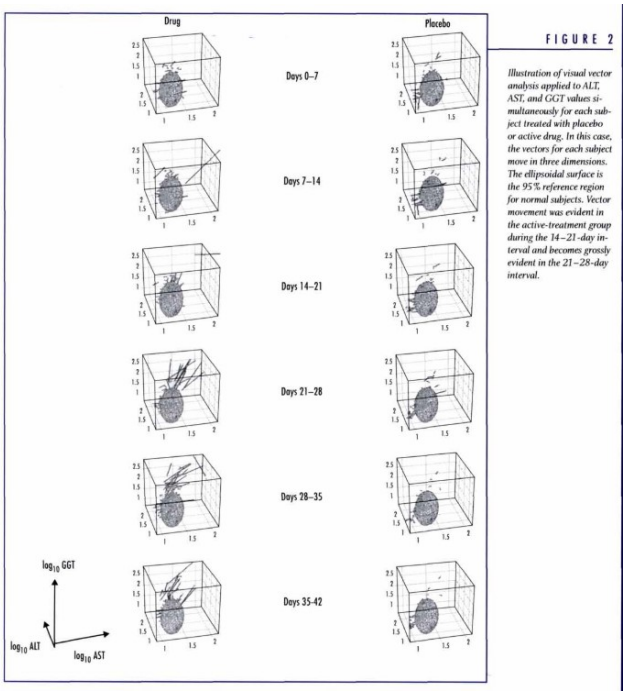


FIGURE 2

Illustration of visual vector analysis applied to ALT, AST, and GGT values simultaneously for each subject treated with placebo or active drug. In this case, the vectors for each subject move in three dimensions. The ellipsoidal surface is the 95% reference region for normal subjects. Vector movement was evident in the active-treatment group during the 14–21-day interval and becomes grossly evident in the 21–28-day interval.

Figure description: Simultaneously displays individual participant changes across three laboratory values. Grey circle indicates the 95% reference range of values for 'normal' subjects. **Caution**: 3D images don't work as static so recommend we don't explore this image any further. *Image taken from: Trost, D. C. and J. W. Freston (2008). "Vector Analysis to Detect Hepatotoxicity Signals in Drug Development." Therapeutic Innovation & Regulatory Science 42(1): 27-34 under the terms of the Creative Commons CC BY License.*

Appendix A5.8: Questions clinicians were asked to consider for each plot during interviews

Question 1: What do you think of this plot? Do you like or dislike?

Question 2: Is anything unclear in this plot that requires further explanation?

Question 3: Are there any advantages of using this plot that you think should be incorporated into the recommendations?

Question 4: Are there any disadvantages of using this plot that can be incorporated into the limitations?

Question 5: Does this plot require any modifications?

Appendix A5.9: Table and figures summarising initial appraisals of all plots by outcome type

Table A5.1a: Plots suitable for **Multiple Binary Outcomes** – summary of scores

Appraisal criteria	Volcano		Alternative volcano 1		Alternative volcano 2		Alternative volcano 3		Dot plot		Bar	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	21	4.0 (1.0) 4 (1, 5)	21	4.0 (0.8) 4 (3, 5)	21	2.3 (1.2) 2 (1, 5)	15	3.2 (1.3) 4 (1, 5)	21	4.3 (1.1) 5 (1, 5)	21	2.8 (1.1) 3 (1, 5)
2.Direction of effect	21	4.1 (1.1) 4 (1, 5)	21	4.0 (1.2) 4 (1, 5)	20	2.9 (1.4) 3 (1, 5)	14	2.8 (1.7) 3 (1, 5)	21	4.5 (0.8) 5 (2, 5)	21	4.0 (1.1) 4 (1, 5)
3.Uncertainty	21	1.8 (1.1) 1 (1, 5)	21	1.8 (0.7) 2 (1, 3)	21	2.0 (1.0) 2 (1, 5)	14	1.3 (0.5) 1 (1, 2)	21	4.4 (1.0) 5 (1, 5)	21	1.3 (0.6) 1 (1, 3)
4.Supplementary data needed	21	2.1 (1.1) 2 (1, 4)	21	2.5 (1.1) 2 (1, 5)	21	2.0 (1.1) 2 (1, 5)	13	2.2 (1.1) 2 (1, 4)	21	3.8 (1.2) 4 (1, 5)	21	2.6 (1.1) 2 (1, 5)
5.Understandable	21	2.9 (0.8) 3 (2, 4)	21	3.0 (1.1) 3 (1, 5)	21	2.3 (1.0) 2 (1, 4)	14	1.5 (0.7) 1 (1, 3)	21	4.2 (0.8) 4 (2, 5)	21	4.8 (0.4) 5 (4, 5)
6.Understandable non-stats	21	2.3 (0.9) 2 (1, 4)	21	2.9 (1.0) 3 (1, 4)	21	2.3 (0.9) 2 (1, 4)	14	1.4 (0.6) 1 (1, 3)	21	4.0 (0.9) 4 (2, 5)	20	4.7 (0.6) 5 (3, 5)
7.Multi-arm studies	21	2.0 (0.7) 2 (1, 3)	21	2.1 (0.9) 2 (1, 4)	21	3.8 (1.0) 4 (1, 5)	14	1.9 (1.0) 2 (1, 4)	21	3.0 (1.0) 3 (1, 5)	21	4.7 (0.7) 5 (2, 5)
8.Limits numbers	21	3.2 (0.8) 3 (2, 5)	21	3.9 (0.8) 4 (2, 5)	21	3.7 (1.1) 4 (1, 5)	14	3.1 (1.2) 3 (1, 5)	21	3.7 (0.8) 4 (2, 5)	21	3.9 (0.7) 4 (3, 5)
9.Overall score*	**21**	**19.2 (4.5)** **19 (11, 27)**	**21**	**20.3 (4.0)** **20 (12, 28)**	**21**	**17.3 (5.0)** **17 (7, 28)**	**21**	**9.6 (7.6)** **11 (0, 23)**	**21**	**28.1 (5.0)** **29 (15, 34)**	**21**	**24.6 (3.2)** **26 (18, 30)**
10. Suitable for publication	21	3.3 (1.1) 3 (1, 5)	21	3.6 (1.0) 4 (1, 5)	19	2.5 (1.4) 2 (1, 5)	14	1.7 (1.1) 1 (1, 4)	20	4.3 (0.9) 5 (2, 5)	20	3.8 (1.0) 4 (2, 5)
11. Suitable for final report	20	3.5 (1.1) 4 (1, 5)	20	3.7 (1.0) 4 (1, 5)	19	2.5 (1.4) 2 (1, 5)	14	1.8 (1.3) 1 (1, 5)	20	4.3 (0.7) 5 (3, 5)	20	4.0 (0.9) 4 (2, 5)
12. Suitable for interim analysis	20	3.2 (1.2) 3 (1, 5)	20	3.5 (1.1) 4 (1, 5)	19	2.4 (1.3) 2 (1, 5)	14	1.5 (0.9) 1 (1, 4)	20	4.3 (0.7) 4 (3, 5)	20	4.1 (0.7) 4 (3, 5)
13.Exploratory analysis	20	3.7 (0.6) 4 (3, 5)	20	3.4 (0.8) 4 (2, 5)	19	2.8 (0.9) 3 (1, 4)	14	1.9 (1.1) 2 (1, 4)	19	3.8 (0.8) 4 (2, 5)	19	3.7 (1.1) 4 (1, 5)
14.Explanatory analysis	20	3.1 (0.7) 3 (2, 4)	20	3.3 (0.9) 3 (2, 5)	19	2.5 (1.0) 3 (1, 4)	14	1.9 (1.1) 2 (1, 4)	19	4.1 (0.8) 4 (3, 5)	19	3.5 (0.9) 3 (2, 5)
Ranking	**18**	**5.6 (2.1)** **5 (2, 10)**	**18**	**4.8 (1.8)** **5 (2, 9)**	**18**	**6.6 (2.9)** **7 (1, 12)**	**14**	**9.8 (2.3)** **11 (4, 12)**	**20**	**1.6 (1.6)** **1 (1, 7)**	**17**	**3.8 (1.9)** **4 (1, 9)**

*** Overall score is the sum total of questions 1-7**

Table A5.1b: Plots suitable for **Multiple Binary Outcomes** – summary of scores

Question	**Tendril**		**Heat map**		**Stacked bar chart**		**Stacked bar chart - counts**		**Star**		**Alluvial**	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	21	1.3 (0.6) 1 (1, 3)	20	3.0 (0.7) 3 (2, 4)	21	3.1 (1.2) 3 (1, 5)	12	2.4 (1.1) 3 (1, 4)	21	2.0 (0.9) 2 (1, 4)	21	1.4 (0.6) 1 (1, 3)
2.Direction of effect	21	1.6 (1.0) 1 (1, 5)	20	3.6 (1.0) 4 (1, 5)	21	4.1 (0.9) 4 (1, 5)	12	3.3 (1.2) 4 (1, 5)	21	2.7 (1.0) 3 (1, 4)	21	1.3 (0.6) 1 (1, 3)
3.Uncertainty	21	1.2 (0.5) 1 (1, 3)	20	1.3 (0.7) 1 (1, 4)	21	1.4 (0.6) 1 (1, 3)	12	1.3 (0.5) 1 (1, 2)	20	1.1 (0.4) 1 (1, 2)	20	1.1 (0.4) 1 (1, 2)
4.Supplementary data needed	21	1.6 (1.2) 1 (1, 5)	20	2.1 (0.8) 2 (1, 4)	21	3.2 (0.9) 3 (2, 5)	12	3.0 (1.0) 3 (2, 5)	20	1.9 (1.1) 2 (1, 5)	20	1.9 (1.4) 1 (1, 5)
5.Understandable	21	1.2 (0.5) 1 (1, 3)	20	3.1 (1.0) 3 (2, 5)	21	4.7 (0.5) 5 (4, 5)	12	4.4 (0.8) 5 (3, 5)	20	2.3 (0.9) 2 (1, 4)	20	2.1 (1.0) 2 (1, 4)
6.Understandable non-stats	21	1.0 (0.2) 1 (1, 2)	20	2.9 (1.0) 3 (2, 5)	21	4.5 (0.6) 5 (3, 5)	12	4.1 (1.0) 4 (2, 5)	20	1.9 (0.7) 2 (1, 3)	20	1.9 (0.9) 2 (1, 4)
7.Multi-arm studies	21	1.4 (0.7) 1 (1, 3)	20	1.9 (0.9) 2 (1, 4)	21	4.1 (1.2) 5 (1, 5)	12	4.3 (1.2) 5 (2, 5)	20	3.9 (1.4) 4 (1, 5)	20	1.6 (0.7) 1 (1, 3)
8.Limits numbers	21	3.3 (1.2) 4 (1, 5)	19	3.6 (1.0) 4 (1, 5)	21	3.5 (0.7) 3 (2, 5)	11	3.5 (0.7) 3 (3, 5)	19	2.9 (1.0) 3 (1, 5)	20	2.5 (1.1) 3 (1, 4)
9.Overall score*	**21**	**9.3 (2.4)** **9 (7, 14)**	**21**	**17.1 (5.1)** **17 (0, 24)**	**21**	**25.1 (2.8)** **25 (19, 29)**	**21**	**13.0 (12.0)** **15 (0, 28)**	**21**	**15.4 (5.0)** **16 (2, 23)**	**21**	**10.9 (3.5)** **11 (2, 21)**
10. Suitable for publication	20	1.5 (0.9) 1 (1, 4)	20	2.8 (0.9) 3 (1, 4)	21	4.0 (0.9) 4 (2, 5)	11	3.5 (1.1) 4 (2, 5)	20	2.2 (1.2) 2 (1, 5)	19	1.7 (1.1) 1 (1, 4)
11. Suitable for final report	20	1.8 (1.1) 1 (1, 4)	20	2.7 (0.9) 3 (1, 4)	21	4.0 (0.8) 4 (2, 5)	11	3.6 (1.0) 4 (2, 5)	20	2.4 (1.3) 2 (1, 5)	19	2.0 (1.1) 2 (1, 4)
12. Suitable for interim analysis	20	2.1 (1.4) 2 (1, 5)	20	2.5 (0.9) 3 (1, 4)	21	4.1 (0.8) 4 (2, 5)	11	3.6 (1.0) 4 (2, 5)	20	2.3 (1.2) 2 (1, 5)	19	2.1 (1.2) 2 (1, 4)
13.Exploratory analysis	20	3.0 (1.2) 3 (1, 5)	19	3.1 (1.0) 3 (1, 4)	20	4.0 (0.9) 4 (2, 5)	11	3.5 (1.1) 4 (2, 5)	19	2.9 (1.1) 3 (1, 5)	19	2.8 (1.5) 3 (1, 5)
14.Explanatory analysis	19	1.7 (1.0) 1 (1, 5)	19	2.6 (1.0) 3 (1, 5)	20	4.0 (1.1) 4 (1, 5)	11	3.3 (1.3) 3 (1, 5)	19	2.2 (1.1) 2 (1, 5)	18	1.7 (0.8) 2 (1, 3)
Ranking	**18**	**10.2 (2.3)** **11 (2, 12)**	**17**	**7.5 (1.8)** **8 (3, 10)**	**19**	**2.4 (1.1)** **2 (1, 5)**	**11**	**4.9 (2.3)** **4 (2, 10)**	**17**	**8.0 (2.3)** **8 (3, 12)**	**17**	**9.9 (2.6)** **11 (3, 12)**

*** Overall score is the sum total of questions 1-7**

Table A5.2: Plots suitable for **Single Binary Outcomes** – summary of scores

Question	**Bar chart**	
	n	Mean (SD) Median (Min, Max)
1.Effect size	23	2.1 (0.9)
		2 (1, 4)
2.Direction of effect	23	2.3 (1.0)
		2 (1, 4)
3.Uncertainty	23	1.1 (0.3)
		1 (1, 2)
4.Supplementary data needed	23	2.5 (1.2)
		2 (1, 4)
5.Understandable	23	4.0 (1.2)
		4 (1, 5)
6.Understandable non-stats	23	3.9 (0.9)
		4 (2, 5)
7.Multi-arm studies	23	3.9 (0.9)
		4 (2, 5)
8.Limits numbers	23	3.2 (1.2)
		3 (1, 5)
9.Overall score	**23**	**19.8 (3.3)**
		20 (11, 26)
10. Suitable for publication	22	3.1 (0.9)
		3 (1, 4)
11. Suitable for final report	22	3.4 (1.0)
		4 (1, 4)
12. Suitable for interim analysis	22	3.5 (1.0)
		4 (1, 5)
13.Exploratory analysis	22	3.1 (1.0)
		3 (1, 5)
14.Explanatory analysis	22	3.0 (1.0)
		3 (1, 5)
Ranking	**6**	**1.0 (0.0)**
		1 (1, 1)

*** Overall score is the sum total of questions 1-7**

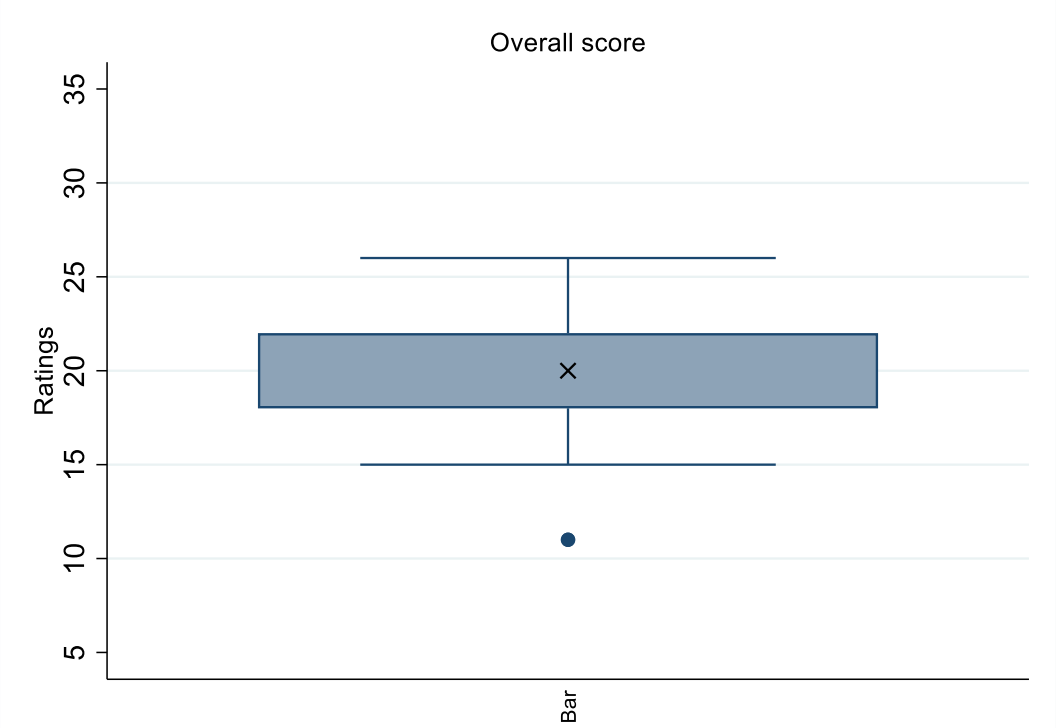


Figure A5.39: Box plot of overall scores, ordered by highest to lowest mean values (higher scores indicate better performance). Note: X indicates median values.

Table A5.3: Plots suitable for **Multiple Time-to-Event Outcomes** – summary of scores

Question	**Matrix of cumulative hazards**		**Bar chart**		**Alternative bar chart**		**Alternative survival plot 1**		**Alternative survival plot 2**	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	21	3.4 (1.0) 4 (2, 5)	21	2.4 (1.2) 2 (1, 4)	21	2.3 (1.3) 2 (1, 4)	21	2.3 (0.8) 2 (1, 4)	18	1.6 (0.8) 1 (1, 4)
2.Direction of effect	21	3.9 (0.8) 4 (2, 5)	21	3.0 (1.3) 3 (1, 5)	21	2.8 (1.1) 3 (1, 4)	21	2.9 (1.2) 3 (1, 5)	18	1.4 (0.6) 1 (1, 3)
3.Uncertainty	21	2.4 (1.5) 2 (1, 5)	21	1.4 (0.8) 1 (1, 4)	21	1.5 (0.8) 1 (1, 3)	21	1.1 (0.3) 1 (1, 2)	18	1.2 (0.4) 1 (1, 2)
4.Supplementary data needed	20	2.5 (1.1) 2 (1, 5)	21	2.1 (1.1) 2 (1, 4)	21	1.9 (0.8) 2 (1, 3)	21	2.0 (1.1) 2 (1, 5)	18	1.7 (1.1 1 (1, 5)
5.Understandable	21	4.1 (0.7) 4 (3, 5)	21	3.2 (1.1) 3 (1, 5)	21	2.7 (1.4) 3 (1, 5)	21	2.2 (1.0) 2 (1, 5)	18	2.5 (1.2) 3 (1, 4)
6.Understandable non-stats	20	3.2 (1.0) 3 (1, 5)	21	3.0 (1.1) 3 (1, 5)	21	2.3 (1.2) 2 (1, 5)	21	2.0 (1.0) 2 (1, 5)	18	2.1 (1.2) 2 (1, 4)
7.Multi-arm studies	21	4.5 (0.6) 5 (3, 5)	21	3.6 (1.3) 4 (1, 5)	21	3.7 (1.4) 4 (1, 5)	21	1.9 (0.9) 2 (1, 4)	18	2.4 (1.2) 3 (1, 5)
8.Limits numbers	21	2.3 (1.1) 2 (1, 5)	21	3.0 (1.0) 3 (1, 4)	19	3.8 (1.3) 4 (1, 5)	21	3.2 (1.1) 3 (1, 5)	18	2.3 (1.0) 3 (1, 4)
9.Overall score	**21**	**24.0 (4.8)** **23 (18, 35)**	**21**	**19.0 (5.4)** **20 (7, 28)**	**21**	**17.5 (6.6)** **20 (7, 29)**	**21**	**14.6 (4.2)** **15 (7, 22)**	**21**	**11.2 (6.1)** **13 (0, 20)**
10. Suitable for publication	20	3.4 (1.1) 4 (1, 5)	20	2.2 (1.1) 2 (1, 4)	20	2.2 (1.4) 2 (1, 5)	20	1.8 (0.9) 2 (1, 4)	18	1.9 (1.0) 2 (1, 4)
11. Suitable for final report	20	3.8 (1.0) 4 (1, 5)	20	2.3 (1.3) 2 (1, 4)	20	2.3 (1.3) 2 (1, 5)	20	1.8 (0.9) 2 (1, 4)	18	1.9 (1.0) 2 (1, 4)
12. Suitable for interim analysis	20	3.8 (1.0) 4 (1, 5)	20	2.3 (1.4) 2 (1, 5)	20	2.3 (1.3) 2 (1, 5)	20	2.2 (1.4) 2 (1, 5)	18	2.1 (1.0) 2 (1, 4)
13.Exploratory analysis	20	3.7 (1.1) 4 (1, 5)	20	2.5 (1.3) 3 (1, 4)	20	2.6 (1.3) 3 (1, 5)	20	3.1 (1.3) 3 (1, 5)	17	2.4 (1.4) 2 (1, 5)
14.Explanatory analysis	20	3.5 (1.1) 4 (1, 5)	20	2.2 (1.3) 2 (1, 4)	20	2.4 (1.2) 3 (1, 4)	20	2.4 (1.0) 2 (1, 4)	17	2.2 (1.1) 2 (1, 4)
Ranking	**18**	**1.3 (0.6)** **1 (1, 3)**	**16**	**3.4 (1.3)** **4 (1, 5)**	**17**	**3.5 (1.3)** **3 (1, 5)**	**16**	**3.1 (1.1)** **3 (2, 5)**	**15**	**4.5 (0.7)** **5 (3, 5)**

*** Overall score is the sum total of questions 1-7**

Table A5.4: Plots suitable for **Single TTE Outcomes** – summary of scores

Question	**Cumulative Hazard**		**Kaplan Meier**		**Mean Cumulative Function**		**Mean Cumulative Duration**		**Stacked bar chart over time**		**Histogram of counts over time**	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	23	3.1 (1.2) 3 (1, 5)	23	3.2 (1.2) 4 (1, 5)	23	3.0 (1.2) 3 (1, 5)	23	2.6 (1.2) 3 (1, 4)	23	2.2 (1.3) 2 (1, 5)	23	2.0 (1.1) 2 (1, 5)
2.Direction of effect	23	3.8 (1.1) 4 (1, 5)	23	3.8 (1.1) 4 (1, 5)	23	3.7 (1.1) 4 (1, 5)	23	3.2 (1.2) 3 (1, 5)	23	2.7 (1.3) 3 (1, 5)	23	2.5 (1.2) 2 (1, 4)
3.Uncertainty	23	4.0 (1.1) 4 (1, 5)	23	4.0 (1.0) 4 (1, 5)	23	3.6 (0.9) 4 (1, 5)	23	3.7 (1.0) 4 (1, 5)	23	1.2 (0.5) 1 (1, 3)	23	1.1 (0.5) 1 (1, 3)
4.Supplementary data needed	23	3.7 (0.9) 4 (2, 5)	23	3.8 (0.9) 4 (2, 5)	23	2.9 (1.0) 3 (1, 4)	23	2.2 (1.1) 2 (1, 4)	23	2.6 (1.2) 3 (1, 5)	23	2.6 (1.3) 3 (1, 4)
5.Understandable	23	4.2 (0.7) 4 (3, 5)	23	4.3 (0.7) 4 (3, 5)	23	3.2 (1.1) 3 (1, 5)	23	2.8 (1.1) 3 (1, 5)	23	3.4 (1.2) 4 (1, 5)	23	3.7 (1.1) 4 (1, 5)
6.Understandable non-stats	23	3.1 (0.9) 3 (1, 5)	23	3.5 (0.8) 4 (2, 5)	23	2.7 (1.1) 3 (1, 4)	23	2.3 (1.0) 2 (1, 4)	23	3.2 (1.1) 3 (1, 5)	23	3.4 (1.2) 3 (1, 5)
7.Multi-arm studies	21	3.9 (0.5) 4 (3, 5)	21	3.9 (0.5) 4 (3, 5)	21	3.8 (0.9) 4 (1, 5)	21	3.5 (0.9) 4 (1, 5)	21	3.9 (0.7) 4 (3, 5)	21	3.4 (0.9) 4 (1, 5)
8.Limits numbers	23	3.1 (1.4) 3 (1, 5)	23	3.1 (1.5) 3 (1, 5)	23	3.4 (1.5) 4 (1, 5)	23	3.2 (1.5) 3 (1, 5)	22	2.8 (1.2) 3 (1, 5)	23	3.5 (1.5) 4 (1, 5)
9.Overall score	**21**	**26.1 (3.8)** **27 (18, 32)**	**21**	**26.7 (3.6)** **27 (18, 32)**	**21**	**23.2 (5.6)** **24 (7, 32)**	**21**	**21.0 (5.1)** **22 (7, 28)**	**21**	**19.8 (5.6)** **18 (11, 32)**	**21**	**19.1 (5.2)** **19 (7, 31)**
10. Suitable for publication	22	3.6 (1.1) 4 (2, 5)	22	4.0 (0.8) 4 (3, 5)	22	3.5 (1.1) 4 (1, 5)	22	3.0 (1.0) 3 (1, 4)	22	2.9 (1.2) 3 (1, 5)	22	2.0 (1.0) 2 (1, 4)
11. Suitable for final report	22	4.0 (1.0) 4 (2, 5)	22	4.2 (0.7) 4 (3, 5)	22	3.6 (1.1) 4 (1, 5)	22	3.2 (1.1) 4 (1, 5)	22	3.0 (1.2) 3 (1, 5)	22	2.4 (1.0) 2 (1, 4)
12. Suitable for interim analysis	22	4.0 (1.0) 4 (2, 5)	22	4.2 (0.8) 4 (2, 5)	22	3.3 (1.1) 3 (1, 5)	22	3.0 (1.0) 3 (1, 5)	22	2.8 (1.4) 3 (1, 5)	22	2.9 (1.3) 3 (1, 5)
13.Exploratory analysis	22	4.0 (0.8) 4 (2, 5)	23	4.0 (1.0) 4 (1, 5)	23	3.6 (0.9) 4 (1, 5)	23	3.5 (0.9) 4 (1, 5)	23	3.2 (1.2) 3 (1, 5)	23	3.3 (1.3) 4 (1, 5)
14.Explanatory analysis	22	4.0 (1.0) 4 (2, 5)	23	4.2 (0.7) 4 (3, 5)	23	3.5 (1.0) 4(1, 5)	23	3.0 (1.0) 3 (1, 5)	23	2.7 (1.2) 3 (1, 5)	23	2.3 (1.1) 2 (1, 4)
Ranking	**20**	**2.5 (1.5)** **2 (1, 6)**	**20**	**1.9 (1.1)** **2 (1, 5)**	**19**	**2.9 (1.4)** **3 (1, 5)**	**20**	**4.3 (1.6)** **5 (1, 6)**	**16**	**4.3 (1.8)** **5 (1, 7)**	**18**	**4.8 (1.4)** **5 (2, 7)**

*** Overall score is the sum total of questions 1-7**

Table A5.5: Plots suitable for **Multiple Continuous Outcomes** – summary of scores

Question	**Scatterplot matrix**		**E-dish**	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	22	2.5 (1.1)	20	1.9 (0.9)
		2 (1, 5)		2 (1, 4)
2.Direction of effect	22	3.0 (1.1)	20	2.5 (1.1)
		3 (1, 5)		2 (1, 5)
3.Uncertainty	22	1.7 (0.8)	20	1.6 (0.8)
		2 (1, 3)		1 (1, 3)
4.Supplementary data needed	22	3.2 (1.1)	20	2.4 (1.1)
		3 (1, 5)		2 (1, 5)
5.Understandable	22	4.5 (0.8)	20	3.8 (1.3)
		5 (2, 5)		4 (1, 5)
6.Understandable non-stats	22	4.3 (0.6)	20	3.6 (1.0)
		4 (3, 5)		4 (1, 5)
7.Multi-arm studies	20	2.8 (1.4)	18	2.2 (1.3)
		3 (1, 5)		2 (1, 5)
8.Limits numbers	22	2.9 (1.2)	19	2.1 (1.2)
		3 (1, 5)		2 (1, 5)
9.Overall score	**20**	**22.8 (4.3)**	18	18.6 (5.4)
		23 (16, 31)		18 (9, 30)
10. Suitable for publication	20	2.8 (1.3)	19	2.7 (1.2)
		3 (1, 5)		3 (1, 5)
11. Suitable for final report	20	3.4 (1.3)	19	3.2 (1.2)
		4 (1, 5)		3 (1, 5)
12. Suitable for interim analysis	20	4.0 (1.0)	19	3.4 (1.2)
		4 (1, 5)		4 (1, 5)
13.Exploratory analysis	20	4.2 (0.8)	19	3.7 (1.1)
		4 (3, 5)		4 (1, 5)
14.Explanatory analysis	20	3.0 (1.1)	19	2.7 (1.2)
		3 (1, 5)		3 (1, 5)
Ranking	**18**	**1.2 (0.4)**	**18**	**1.9 (0.6)**
		1 (1, 2)		**2 (1, 3)**

* Overall score is the sum total of questions 1-7

Excludes summary for vector plots.

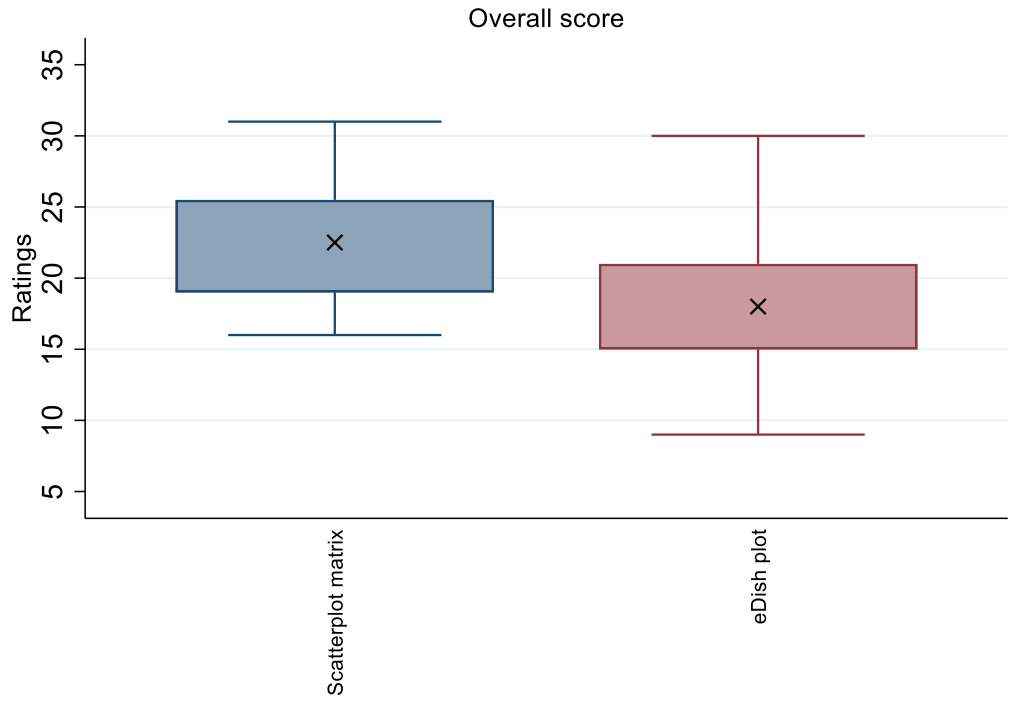


Figure A5.40: Box plot of overall scores, ordered by highest to lowest mean values (higher scores indicate better performance). Note: X indicates median values. Excludes summary for vector plots.

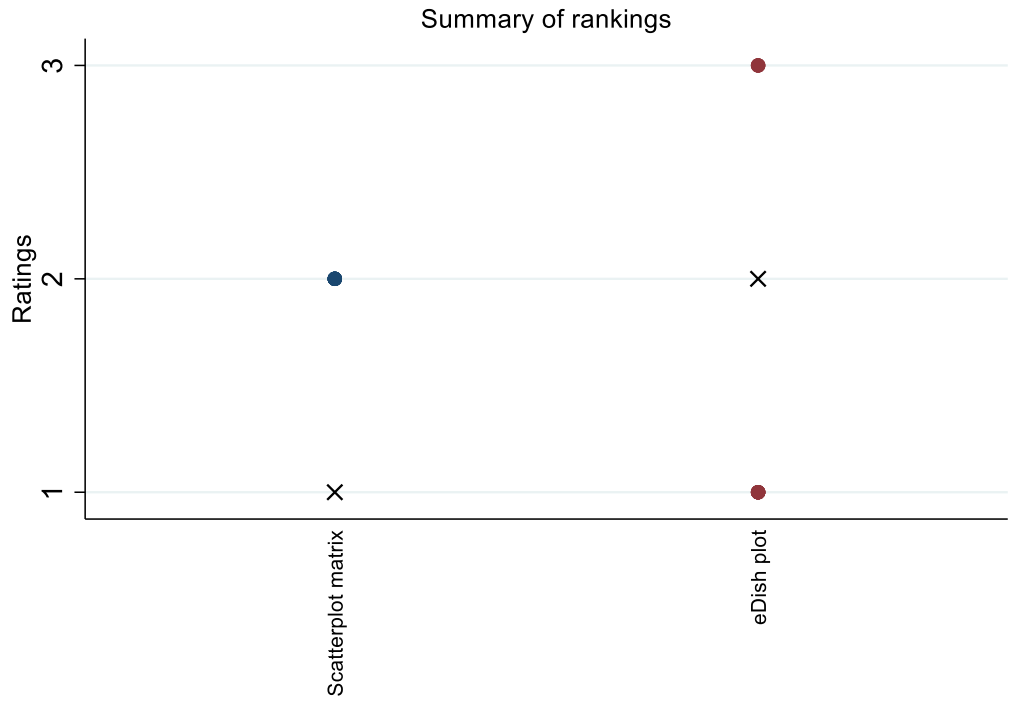


Figure A5.41: Box plot of rankings ordered by best to worst mean rank (lower ranking indicates preferred plot). Note: X indicates median ranks. Excludes summary for vector plots.

Table A5.6a: Plots suitable for **Single Continuous Outcomes** – summary of scores

Question	Empirical distribution of max change		Histogram of max change		Delta plot		Line graph - change		Boxplot - change		Violin plot - change	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	22	2.2 (1.2) 2 (1, 4)	22	2.5 (1.2) 2 (1, 5)	23	1.7 (1.1) 1 (1, 5)	22	3.4 (1.4) 4 (1, 5)	22	2.7 (1.3) 3 (1, 5)	22	2.5 (1.3) 2 (1, 5)
2.Direction of effect	22	3.1 (1.2) 3 (1, 5)	22	3.0 (1.1) 3 (1, 4)	23	1.6 (0.8) 1 (1, 4)	22	4.0 (0.9) 4 (2, 5)	22	3.1 (1.1) 3 (1, 5)	22	3.0 (1.1) 3 (1, 5)
3.Uncertainty	22	1.3 (0.5) 1 (1, 2)	22	1.7 (1.0) 1 (1, 4)	23	1.2 (0.4) 1 (1, 2)	22	3.9 (0.9) 4 (2, 5)	22	3.3 (1.3) 4 (1, 5)	22	3.0 (1.1) 3 (1, 5)
4.Supplementary data needed	22	2.5 (1.1) 3 (1, 4)	22	3.5 (0.9) 4 (2, 5)	23	1.7 (0.8) 2 (1, 3)	22	3.6 (1.0) 4 (2, 5)	22	3.4 (0.9) 4 (2, 5)	22	3.1 (0.8) 3 (2, 4)
5.Understandable	21	3.0 (0.9) 3 (1, 4)	22	4.5 (0.7) 5 (3, 5)	23	1.7 (0.8) 2 (1, 4)	22	4.4 (0.7) 4 (3, 5)	22	4.5 (0.7) 5 (3, 5)	22	3.8 (0.7) 4 (3, 5)
6.Understandable non-stats	22	2.4 (0.9) 3 (1, 4)	22	4.2 (0.6) 4 (3, 5)	23	1.4 (0.6) 1 (1, 3)	22	4.1 (0.8) 4 (3, 5)	22	3.9 (0.8) 4 (2, 5)	22	2.9 (0.9) 3 (2, 5)
7.Multi-arm studies	21	3.9 (0.9) 4 (1, 5)	20	3.2 (0.9) 3 (2, 5)	21	2.4 (1.1) 2 (1, 4)	20	4.0 (0.6) 4 (3, 5)	20	4.0 (0.8) 4 (2, 5)	19	3.7 (0.6) 4 (3, 5)
8.Limits numbers	22	2.5 (1.6) 3 (1, 5)	20	2.4 (1.5) 2 (1, 5)	22	2.0 (1.2) 2 (1, 5)	22	2.6 (1.4) 3 (1, 5)	22	2.7 (1.4) 3 (1, 5)	22	2.7 (1.4) 3 (1, 5)
9.Overall score	**22**	**17.7 (5.9)** **19 (0, 26)**	**21**	**21.9 (5.8)** **21 (0, 28)**	**21**	**12.0 (3.8)** **11 (7, 20)**	**21**	**26.3 (7.3)** **28 (0, 35)**	**21**	**23.7 (6.3)** **25 (0, 34)**	**21**	**20.3 (5.4)** **21 (0, 26)**
10. Suitable for publication	22	2.9 (1.2) 3 (1, 5)	21	3.2 (1.0) 3 (1, 5)	21	1.6 (0.6) 2 (1, 3)	21	4.0 (0.7) 4 (3, 5)	21	3.6 (1.0) 4 (1, 5)	21	3.4 (0.9) 3 (1, 5)
11. Suitable for final report	22	3.1 (1.1) 4 (1, 5)	21	3.7 (1.0) 4 (1, 5)	21	1.8 (0.9) 2 (1, 4)	21	4.0 (0.7) 4 (3, 5)	21	3.6 (1.1) 4 (1, 5)	21	3.4 (1.0) 3 (1, 5)
12. Suitable for interim analysis	22	3.3 (1.2) 4 (1, 5)	21	3.6 (1.1) 4 (1, 5)	21	2.0 (1.1) 2 (1, 5)	21	3.9 (0.8) 4 (3, 5)	21	3.8 (1.0) 4 (1, 5)	21	3.5 (0.9) 3 (1, 5)
13.Exploratory analysis	22	3.6 (1.1) 4 (1, 5)	21	3.7 (0.9) 4 (2, 5)	21	2.4 (1.1) 2 (1, 5)	21	4.0 (0.8) 4 (2, 5)	21	3.9 (1.0) 4 (1, 5)	21	3.7 (0.7) 4 (2, 5)
14.Explanatory analysis	22	2.9 (1.2) 3 (1, 5)	21	3.4 (1.1) 3 (2, 5)	21	1.8 (0.8) 2 (1, 3)	21	4.0 (0.7) 4 (2, 5)	21	3.7 (1.1) 4 (1, 5)	21	3.5 (0.8) 3 (2, 5)
Ranking	**17**	**6.8 (1.9)** **8 (3, 9)**	**17**	**4.9 (2.1)** **5 (1, 8)**	**19**	**8.1 (1.8)** **9 (2, 9)**	**20**	**2.0 (1.2)** **2 (1, 4)**	**18**	**3.6 (2.1)** **4 (1, 7)**	**18**	**3.8 (1.8)** **4 (1, 7)**

*** Overall score is the sum total of questions 1-7**

Table A5.6b: Plots suitable for **Single Continuous Outcomes** – summary of scores

Question	**Line graph - raw**		**Box plot - raw**		**Violin - raw**	
	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)	n	Mean (SD) Median (Min, Max)
1.Effect size	22	3.0 (1.4)	22	2.6 (1.3)	22	2.4 (1.3)
		3 (1, 5)		2 (1, 5)		2 (1, 5)
2.Direction of effect	22	3.7 (1.1)	22	3.0 (1.0)	22	2.9 (1.0)
		4 (1, 5)		3 (1, 5)		3 (1, 4)
3.Uncertainty	22	3.6 (1.0)	22	3.3 (1.3)	22	3.0 (1.1)
		4 (1, 5)		4 (1, 5)		3 (1, 5)
4.Supplementary data needed	22	3.4 (1.1)	22	3.2 (1.0)	22	3.0 (1.0)
		3 (2, 5)		3 (2, 5)		3 (1, 5)
5.Understandable	22	4.5 (0.6)	22	4.4 (0.8)	22	3.8 (0.9)
		5 (3, 5)		5 (2, 5)		4 (2, 5)
6.Understandable non-stats	22	4.2 (0.7)	22	3.9 (0.9)	22	2.9 (1.0)
		4 (3, 5)		4 (2, 5)		3 (1, 5)
7.Multi-arm studies	20	4.0 (0.6)	20	3.9 (0.7)	19	3.6 (0.6)
		4 (3, 5)		4 (2, 5)		4 (3, 5)
8.Limits numbers	22	2.6 (1.4)	22	2.7 (1.4)	22	2.6 (1.4)
		2 (1, 5)		3 (1, 5)		3 (1, 5)
9.Overall score	**21**	**25.3 (7.4)**	**21**	**23.1 (6.4)**	**21**	**20.1 (5.4)**
		25 (0, 35)		**23 (0, 34)**		**21 (0, 26)**
10. Suitable for publication	21	3.8 (1.0)	21	3.4 (1.1)	21	3.2 (1.0)
		4 (1, 5)		3 (1, 5)		3 (1, 5)
11. Suitable for final report	21	3.9 (1.0)	21	3.6 (1.0)	21	3.4 (1.0)
		4 (1, 5)		4 (1, 5)		3 (1, 5)
12. Suitable for interim analysis	21	3.8 (1.0)	21	3.6 (1.0)	21	3.3 (1.0)
		4 (1, 5)		4 (1, 5)		3 (1, 5)
13.Exploratory analysis	21	3.9 (0.8)	21	3.8 (0.9)	20	3.6 (0.7)
		4 (2, 5)		4 (1, 5)		4 (2, 5)
14.Explanatory analysis	21	3.8 (1.0)	21	3.5 (1.1)	20	3.3 (0.9)
		4 (2, 5)		3 (1, 5)		3 (2, 5)
Ranking	**18**	**3.2 (1.8)**	**18**	**4.7 (2.2)**	**18**	**4.7 (2.3)**
		3 (1, 8)		**5 (1, 8)**		**5 (1, 8)**

*** Overall score is the sum total of questions 1-7**

Appendix A5.10: Free text comments summarised for each plot

Included plots considered for multiple binary outcomes

i. Dot plot

Proposals to the dot plot included adding numerical raw data via either a data table on the right-hand side of the plot or labelling data points on the left-hand side of the plot in order to enrich information presented and to provide an alternative to the typical frequency tables presented in publications. Concerns were raised about the inclusion of confidence intervals in this plot as this could encourage use as a proxy for hypothesis tests but discussions indicated that this could be caveated by including a caution to avoid such interpretation in the recommendations for use.

ii. Stacked bar charts

Preference was for stacked bar charts of percentages with at least one event and inclusion of bar labels of frequencies or counts of events. Imposing a meaning to the order of bars was also advocated.

Excluded plots considered for multiple binary outcomes

i. Bar chart

Suggested amendments included labelling bars with frequencies/counts instead of percentages, to order the bars in a meaningful way, to present as dots instead of bars and to incorporate confidence intervals or standard errors. General comments concluded that whilst this is a simple plot that can convey a clear message, it doesn't include key information that is presented in other examples such as the dot plot or stacked bar chart.

ii. Volcano plot

Comments on amendments focused on removal of repeated information and to instead use colour saturation to display severity or some other facet instead of the size of the p-value which is also displayed on y-axis; to incorporate the number of events/participants or proportions onto the plot (probably only suitable if a small number of events); and the possibility of displaying alternative data on x and y axis. Criticisms of this plot included the inefficient use of space as the two axes are strongly correlated, that the log and p-value is confusing for most clinicians and as such should be avoided, the strong focus on p-values, the large redundancy of information, reliance on colour and overlap/crowding of labels if lots of events.

iii. Alternative volcano 1

This first draft alternative was considered as an alternative to the volcano plot and several possible adaptions were discussed including possible incorporation of information on confidence intervals possibly with lines, colour saturation or parentheses and incorporation of information on severity possibly through colour shading. It scored higher than the volcano in terms of overall score and ranking.

iv. Alternative volcano 2

Amendments suggested for the second alternative included changing the x-axis to an estimate of treatment effect such as the risk-ratio or risk difference and instead using colour to represent the size of the p-value. Comments also indicated that this could be easily adapted to multi-arm trials.

v. Heat map

Proposed amendments included annotating the plot with event counts or using colour sharing of a proportion of the box to represent total number of events. Criticisms focused on failure to properly convey uncertainty, the colour dependence, and that light shading appears to convey that there is no information, but in fact there may be a lot of events that are equally balanced between treatment arms.

vi. Star plot

Comments on the star plot focused on the difficulty in interpretation and the limits in the information it conveys.

vii. Alluvial plot

Comments suggested that there could be potential utility in this plot to explore AE over time but would need separate plots for each treatment arm.

viii. Tendril plot

Comments suggested that this plot was too complicated and similar information could be presented more simply using Cartesian coordinates rather than polar.

ix. Alternative volcano 3

Limited comments on this plot as discussions indicated that it needed more design considerations before could be fairly appraised.

Included plots considered for single binary outcomes

i. Bar chart of counts

A boxplot or dot plot of a summary measure of count data was suggested to replace the bar chart as a means to summarise count data, however, there was not whole group support for this idea. There was variation in preferences for layout of the bar chart with some preferring side-by-side plots for each treatment group and others preferring plots stacked one above the other for each treatment group. Discussions highlighted that some participants question the need for this plot, for example, with one participant commenting, “is aggregation of data like this helpful?”, and others felt there could be difficulty in interpreting these plots. Discussions concluded that this plot might only be useful for summaries of serious events or pre-specified events.

Excluded plots considered for multiple time-to-event outcomes

i. Matrix of Kaplan-Meier

Discussions indicated that the matrix of Kaplan-Meier plots required incorporation of confidence bands and tables of numbers at risk as per the individual Kaplan-Meier plots. Participants indicated that this plot would be useful to detect disproportionalities for pre-specified events but that the number of events looked at would need to be limited to be useful. To avoid encouraging performance of many hypothesis tests it was also highlighted that it should be clearly specified that this plot should be used as a way to display risk over time to help identify disproportionalities and raise signals for ADRs. Alternatives that incorporate information on recurrent events are still needed.

ii. Alternative survival plot 1

Proposed as an alternative to the matrix of survival plots. Amendments considered including using lines instead of blocks of colour or a line over time that rises above the x-axis for an event in one treatment arm and below for an event in the other treatment arm. General comments indicated with refinement this plot could be useful but concern about how it would account for censoring and incorporate uncertainty.

iii. Alternative survival plot 2 – Sankey diagram

Proposed as an alternative to the matrix of survival plots. Comments indicated that this might be useful for profiling one specific AE of interest in one arm, but otherwise was too nuanced and dense.

iv. Bar chart of median time-to-event

This plot caused a lot of confusion. Participants predominantly thought it showed median time to events which is not always achieved (i.e. if <50% have event). However, it in fact showed median time of events amongst those with events, which present its own problems such as ambiguity around numbers with events.

v. Alternative to bar chart of median time data

This plot was proposed as an alternative to the bar chart of median time to events plot but was deemed to only be feasible in an interactive setting and was given little further consideration.

Included plots considered for single time-to-event outcomes

i. Kaplan-Meier

Amendments discussed for the Kaplan-Meier plot included: incorporating an extended risk table including the number of participants that remain 'at risk', the cumulative number that have been censored and the cumulative number that have experienced an event at each discrete time point; providing a clear definition of what 'survival' means in the context of analysing harm outcomes in the recommendations; and incorporating a between group comparison. This latter point prompted discussions and a suggestion to consider the survival ratio plot proposed by Newell et al.[246] Survival ratio plots were not formally considered via appraisals but were incorporated into discussions for consideration. Given the context of use in this thesis I instead refer to the survival ratio plot as the event-free ratio plot. Discussions revealed some concerns about use of time-to-event plots in this setting and concluded that recommendations should caution users to bear in mind the consequences of competing risks (this is discussed below in section 5.5.5).

ii. Mean cumulative function

Proposals for the mean cumulative function included adding confidence interval bands and at risk tables. Discussions covered whether grouping all events together in this plot should be encouraged or that instead the recommendation should be for use when analysing pre-specified events of interest. Participants endorsed the latter, recommending use to account for recurrent events. Discussions also indicated that recommendation should include clear text descriptions explaining the interpretation of this plot given its novelty and clarifying that it adequately accounts for censoring.

Excluded plots considered for single time-to-event outcomes

i. Stacked bar charts of counts

The main discussions around this plot were highlighting that it wasn't really a time-to-event plot and that it in fact conveyed information on binary outcomes and the time at which they occurred.

ii. Histogram of counts over time

General comments indicated that this could be used to given an idea of overall burden to participants but that there were many limitations. Including: not accounting for censoring and its inability to convey information on uncertainty. There were also concerns about grouping all events together.

iii. Mean cumulative duration

The main concerns with this plot were that it is conditional on participants having the event, so it only provides a fair comparison if event rates are similar between arms.

Included plots considered for multiple continuous outcomes

i. Scatterplot matrix

Discussed amendments to the scatterplot matrix included ways to help ease the problems created by overlapping points, inclusion of reference lines and labels for outliers.

Excluded plots considered for multiple continuous outcomes

i. E-dish plot

The group decided that this plot looked at the relationship between two 'harm' outcomes rather than summarising harms so wasn't given any further consideration.

ii. Vector plot

Discussions indicated that this plot would only be feasible in an interactive setting so wasn't given any further consideration.

Included plots considered for single continuous outcomes

i. Line chart

Discussions focused on the appropriate statistic to display on the line chart as well as advocating for inclusion of tables with numbers at risk at the bottom of the plot that are typically seen on time-to-event plots such as the Kaplan-Meier plot.

ii. Violin plot

A proposal to remove the duplication of information in the 'mirrored' distributions of the violin plot was discussed and an indication of a preference for violin plots over box plots was voiced.

iii. Histogram/Kernel density plot

Discussions indicated that participants wished to see this information presented graphically but would prefer to see it displayed in kernel density plots instead of histograms, which would overcome the problems of overlap encountered in the histogram.

Excluded plots considered for single continuous outcomes

i. Empirical distribution

General comments indicated this could be useful in situations when there were not many events and displaying maximum values is indicative of the whole variable distribution, otherwise it would be more useful to display statistics such as mean or medians on an alternative plot.

ii. Box plot

Comments on the boxplot were varied. Many liked it but indicated a preference for the line or violin plots. Comments indicated that the box plot becomes crowded easily, hence making it harder to pick out messages and differences between treatment arms.

iii. Delta plot

Comments on this plot were predominantly about the difficulties in making comparison and how it should be interpreted and as such was given little further consideration.

Appendix A5.11: Tables summarising Mentimeter votes to decide which plots to take forward and amendments

Table A5.7: Decisions for **Multiple Binary Outcomes**

Question		n	%
Should we **keep** the dot plot?	Yes	19	100
	No	0	0
Should we **keep** the stacked bar chart?	Yes	18	95
	No	1	5
Should we **keep** the bar chart?	Yes	17	81
	No	4	19
Should we **exclude** the tendril plot?	Yes	20	100
	No	0	0
Should we **exclude** the alluvial plot?	Yes	17	94
	No	1	6
Should we **exclude** alternative volcano 3?	Yes	18	100
	No	0	0
Should we **keep** the volcano plot?	Yes	11	65
	No	6	35
Should we **keep** alternative volcano 1?	Yes	8	47
	No	9	53
Should we **keep** alternative volcano 2?	Yes	10	50
	No	10	50
Should we keep the **heat** map?	Yes	3	16
	No	16	84
Should we keep the **star** plot?	Yes	2	10
	No	18	90

Table A5.8: Decisions about amendments for the recommend plots for **Multiple Binary Outcomes**

		n	%
Are we happy to recommend the dot plot as it is unedited?	Yes	14	67
	No	7	33
Do we want to add in counts and number of participants into the data table?	Yes	12	60
	No	8	40
Are we happy to recommend the stacked bar chart as it is unedited?	Yes	16	80
	No	4	20
Do we want to recommend the volcano in light of possible alternative?	Yes	7	35
	No	13	65
Do we want to recommend the alternative volcano 2 instead?	Yes	7	35
	No	13	65

Table A5.9: Decisions about plots in the **Single Binary Outcome** setting

		n	%
Is it helpful to use a plot in this setting?	Yes	14	67
	No	7	33
Would you like to see this in bar chart?	Yes	15	75
	No	5	25
Would you prefer the data to be presented by bars or dots?	Bars	15	79
	Dots	4	21
Should we present as two separate charts one above the other aligned vertically?	Yes	9	50
	No	9	50
Should we present as two separate charts one above the other aligned vertically?	Context specific e.g. only 2 arms then horizontal, >2 arms then stacked	11	58
	Stacked one above the other	7	37
	Horizontal - side by side	1	5

Table A5.10: Decisions for plots in the **Multiple Time-to-Event** setting

		n	%
Should we recommend any of these plots?	Matrix of multiple KM	8	40
	Bar chart of median time-to-event	0	0
	Heat map/alternative survival plot 1	2	10
	None of these	10	50

Acronyms: KM – Kaplan-Meier

Table A5.11: Decisions for plots to recommend in the **Single Time-to-Event** setting

		n	%
Should we recommend KM or Cumulative Hazard plots?	Cumulative hazard	3	17
	Kaplan-Meier	15	83
What should the table at the bottom of the KM plot contain?	No table	1	6
	Minimum - at risk table only (by arm)	4	24
	Full table as per KMUNICATE	12	71
Should we recommend the survival ratio plot as an alternative to the KM?	Yes	12	67
	No	6	33
Should we recommend the MCF plot for displaying information on repeated events?	Yes	15	88
	No	2	12
Should the table at the bottom of the MCF plot only contain the number at risk (by arm)	Yes	17	94
	No	1	6

Acronyms: KM – Kaplan-Meier; MCF – Mean Cumulative Function

Table A5.12: Decisions for plots in the **Multiple Continuous Outcome** setting

		n	%
Should we recommend the scatterplot matrix?	Yes	16	94
	No	1	6

Table A5.13: Decisions for plots in the **Single Continuous Outcome** setting

		n	%
Should we recommend a version of the line chart?	Yes	17	94
	No	1	6
Should we recommend a version of the boxplot?	Yes	9	53
	No	8	47
Should we recommend a version of the violin plot?	Yes	12	67
	No	6	33
Should we recommend a version of the histogram?	Kernel density	11	61
	Histogram	0	0
	Neither	7	39

Appendix A7.1: Power and false positive rates of the Fisher's exact test and Chi-squared test

Table A7.1: Power of Fisher's exact test and Chi-squared test to detect a signal for an ADR by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Fisher's exact		**Chi-squared test**	
		N=45,000			
Sample size		n	n/N	n	n/N
200	Power	4,631	**0.10**	5,648	**0.13**
	Model fail	0		0	
400	Power	9,584	**0.21**	10,833	**0.24**
	Model fail	0		0	
800	Power	15,320	**0.34**	15,971	**0.35**
	Model fail	0		0	
1000	Power	17,179	**0.38**	17,776	**0.40**
	Model fail	0		0	
2000	Power	23,661	**0.53**	24,255	**0.54**
	Model fail	0		0	
5000	Power	32,487	**0.72**	32,775	**0.73**
	Model fail	0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence

Table A7.2: False positive rate for the Fisher's exact test and Chi-squared test by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Fisher's exact		**Chi-squared test**	
		N=45,000			
Sample size		n	n/N	n	n/N
200	False positive	975	**0.02**	1,545	**0.03**
	Model fail	0		0	
400	False positive	1,605	**0.04**	2,715	**0.06**
	Model fail	0		0	
800	False positive	1,575	**0.04**	2,415	**0.05**
	Model fail	0		0	
1000	False positive	1,320	**0.03**	1,875	**0.04**
	Model fail	0		0	
2000	False positive	2,160	**0.05**	2,415	**0.05**
	Model fail	0		0	
5000	False positive	2,040	**0.05**	2,250	**0.05**
	Model fail	0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence

Appendix A7.2: Power and false positive rates of the of the alternative Weibull survival models

Table A7.3: Power of variations of the Weibull survival model to detect signals for an ADR by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

Model		Weibull model with no treatment covariate or ancillary parameter		Weibull model with treatment covariate				Weibull model with treatment group ancillary parameter		Weibull model with treatment covariate and treatment group ancillary parameter			
Parameter to flag signal		Overall shape parameter		Treatment group covariate parameter		Overall shape parameter		Treatment group shape parameter		Treatment group covariate parameter		**Treatment group shape parameter***	
		N=45,000											
Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
200	Power	8,594	**0.19**	4,920	**0.11**	8,589	**0.19**	7,113	**0.17**	4,810	**0.11**	7,007	**0.16**
	Model fail	0		0		0		2437		2		2,437	
400	Power	12,697	**0.28**	9,500	**0.21**	12,727	**0.28**	10,504	**0.24**	9,380	**0.21**	10,413	**0.23**
	Model fail	0		0		0		314		0		314	
800	Power	16,766	**0.37**	15,464	**0.34**	16,801	**0.37**	14,025	**0.31**	15,307	**0.34**	13,957	**0.31**
	Model fail	0		0		0		0		0		0	
1000	Power	17,489	**0.39**	17,388	**0.39**	17,505	**0.39**	15,013	**0.33**	17,261	**0.38**	14,964	**0.33**
	Model fail	0		0		0		0		0		0	
2000	Power	22,039	**0.49**	24,012	**0.53**	22,041	**0.49**	19,302	**0.43**	23,867	**0.53**	19,296	**0.43**
	Model fail	0		0		0		0		0		0	
5000	Power	28,164	**0.63**	32,625	**0.73**	28,065	**0.62**	25,237	**0.56**	32,490	**0.72**	25,186	**0.56**
	Model fail	0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
*Primary result presented for the Weibull survival model with ancillary parameter

Table A7.4: False positive rate for variations of the Weibull survival model by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

Model		Weibull model with no treatment covariate or ancillary parameter		Weibull model with treatment covariate				Weibull model with treatment group ancillary parameter		Weibull model with treatment covariate and treatment group ancillary parameter			
Parameter to flag signal		Overall shape parameter		Treatment group covariate parameter		Overall shape parameter		Treatment group shape parameter		Treatment group covariate parameter		**Treatment group shape parameter***	
Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
200	False positive	23,190	**0.52**	930	**0.02**	23,310	**0.52**	6,525	**0.15**	2,100	**0.05**	7,230	**0.18**
	Model fail	0		0		0		1635		3900		4,005	
400	False positive	28,920	**0.64**	1,185	**0.03**	28,845	**0.64**	7,095	**0.16**	2,160	**0.05**	7,500	**0.17**
	Model fail	0		0		0		390		915		960	
800	False positive	34,815	**0.77**	1,395	**0.03**	34,830	**0.77**	7,470	**0.17**	2,280	**0.05**	7,950	**0.18**
	Model fail	0		0		0		15		45		45	
1000	False positive	36,000	**0.80**	1,365	**0.03**	35,970	**0.80**	6,690	**0.15**	2,460	**0.05**	7,350	**0.16**
	Model fail	0		0		0		0		0		0	
2000	False positive	40,545	**0.90**	2,355	**0.05**	40,590	**0.90**	6,585	**0.15**	3,000	**0.07**	7,290	**0.16**
	Model fail	0		0		0		0		0		0	
5000	False positive	44,205	**0.98**	1,980	**0.04**	44,205	**0.98**	6,255	**0.14**	2,865	**0.06**	6,945	**0.15**
	Model fail	0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
*Primary result presented for the Weibull survival model with ancillary parameter

Appendix A7.3: Sample sizes required by each of the other investigated tests to achieve 80% power and the specific power of each test

Table A7.5: Sample size required by each test to achieve ≥80% power across scenarios

Time		Day 1 ± 0.5 day			Month 1 ± 2 weeks			Month 3 ± 2 weeks			Month 6 ± 2 weeks			Month 11 ± 2 weeks		
		Increased AE rate (%)														
Model/test	AE background rate	25	50	100	25	50	100	25	50	100	25	50	100	25	50	100
		Sample size														
Weibull model with Ancillary parameter	1%	-	5000	2000	-	-	-	-	-	-	-	-	-	-	-	-
	5%	2000	800	400	-	5000	2000	-	-	-	-	-	-	-	5000	2000
	10%	800	400	200	5000	2000	800	-	-	-	-	-	-	-	2000	800
Double-Weibull model with Ancillary parameter	1%	-	5000	2000	-	-	-	-	-	-	-	-	-	-	-	-
	5%	2000	800	400	-	5000	2000	-	-	-	-	-	5000	-	5000	2000
	10%	800	400	200	5000	2000	1000	-	-	2000	-	5000	2000	-	5000	800
Cox PH model with GT test for disproportionality	1%	-	-	5000	-	-	5000	-	-	-	-	-	-	-	-	5000
	5%	-	2000	800	-	5000	1000	-	-	5000	-	-	-	-	5000	1000
	10%	5000	1000	400	5000	2000	800	-	5000	2000	-	-	-	5000	2000	800
Double-Cox PH model with GT test for disproportionality	1%	-	-	5000	-	-	5000	-	-	-	-	-	-	-	-	-
	5%	5000	2000	800	-	5000	1000	-	-	5000	-	-	5000	-	5000	2000
	10%	2000	800	400	5000	2000	800	-	5000	2000	-	5000	1000	5000	2000	800
Combined test	1%	-	5000	2000	-	-	5000	-	-	5000	-	-	5000	-	-	5000
	5%	5000	800	400	5000	2000	800	-	5000	800	-	5000	1000	-	5000	1000
	10%	2000	400	200	2000	800	200	5000	2000	400	-	2000	400	-	2000	800
Fisher's Exact test	1%	-	-	5000	-	-	5000	-	-	5000	-	-	5000	-	-	5000
	5%	-	5000	800	-	5000	800	-	5000	800	-	5000	800	-	5000	800
	10%	5000	2000	400	5000	2000	400	5000	2000	400	5000	2000	400	5000	2000	400
Beta-Binomial model	1%	-	-	5000	-	-	5000	-	-	5000	2000	-	5000	-	-	5000
	5%	-	2000	800	-	2000	800	-	2000	800	-	2000	800	-	2000	800
	10%	5000	800	200	5000	800	400	5000	800	200	5000	800	200	5000	800	200
Gamma-Poisson model	1%	-	-	5000	-	-	5000	-	-	5000	-	-	5000	-	-	5000
	5%	-	2000	800	-	2000	800	-	2000	800	-	2000	800	-	2000	800
	10%	5000	1000	400	5000	1000	400	5000	1000	400	5000	1000	400	5000	1000	400

Note: Dash (-) indicates that 80% power not achieved across the sample sizes explored (n= 200, 400, 800, 1000, 2000, 5000). Acronyms: PH – proportional hazards; GT - Grambsch-Therneau

Table A7.6: **Power** by scenario for the **Weibull survival model with ancillary parameter** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	6.7 (N=749)	8.9 (N=944)	17.2	17.6	28.5	74.7	8.2 (N=753)	21.4 (N=998)	40.3	36.1	74.9	**98.2**	21.2 (N=993)	40.0	60.5	70.4	**94.5**	**100**
	5%	16.9	28.3	67.4	77.1	**97.9**	**100**	37.3	72.0	**96.3**	**98.5**	**100**	**100**	69.9	**93.6**	**99.9**	**100**	**100**	**100**
	10%	30.6	69.9	**94.3**	**96.8**	**100**	**100**	74.1	**95.3**	**100**	**100**	**100**	**100**	**93.7**	**99.9**	**100**	**100**	**100**	**100**
Month 1	**1%**	8.3 (N=773)	7.9 (N=929)	7.3	5.2	6.7	15.0	6.4 (N=778)	7.8	10.4	10.9	19.6	34.3	9.1 (N=993)	13.8	17.7	21.9	35.5	71.9
	5%	6.7	6.9	13.8	16.4	28.8	61.1	10.1	19.6	27.0	34.3	65.3	**96.6**	18.5	33.4	59.9	69.1	**93.7**	**100**
	10%	8.3	13.3	23.6	26.9	53.0	**92.7**	17.7	28.1	56.2	65.9	**92.0**	**100**	32.8	56.9	**85.7**	**92.2**	**100**	**100**
Month 3	**1%**	8.8 (N=761)	9.4 (N=944)	4.2	3.1	5.2	4.4	7.8 (N=755)	6.2	4.8	5.7	5.0	5.2	5.8	6.9	7.0	5.2	7.0	10.6
	5%	5.2	4.0	4.5	5.8	6.7	8.4	4.8	5.9	4.8	5.7	8.0	16.0	5.8	6.9	8.6	9.6	15.4	30.0
	10%	4.4	4.8	4.4	5.4	7.7	11.5	4.7	5.8	7.2	7.8	11.5	26.8	5.9	7.7	10.7	12.7	20.5	48.3
Month 6	**1%**	6.9 (N=739)	10.0 (N=932)	5.3	5.0	5.3	7.2	7.3 (N=754)	5.2	4.2	3.1	5.4	7.3	3.1	6.2	5.8	4.4	5.5	8.2
	5%	6.3	6.0	6.0	5.0	6.8	8.3	5.7	4.8	4.4	6.5	9.2	17.4	5.4	6.6	8.6	10.6	16.5	39.0
	10%	5.7	4.8	6.6	6.3	8.1	14.3	6.4	6.0	9.8	9.9	16.6	36.0	7.1	9.0	16.0	20.6	36.5	76.6
Month 11	**1%**	6.7 (N=744)	10.0 (N=939)	11.8	9.5	8.3	11.0	7.3 (N=771)	13.9	14.9	12.9	17.3	28.8	16.4	23.4	22.0	23.8	39.7	66.8
	5%	9.2	9.1	11.2	11.4	19.4	39.5	11.9	17.5	25.0	31.8	51.8	**87.1**	26.6	38.7	57.6	69.4	**92.1**	**99.9**
	10%	7.1	12.3	18.0	20.0	34.7	66.2	13.8	27.2	47.9	53.9	**79.4**	**99.3**	36.4	58.6	**86.9**	**92.0**	**99.6**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.7: **Power** by scenario for the **double-Weibull survival model with ancillary parameter** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	5.6 (N=752)	10.7 (N=944)	20.6	19.0	30.5	75.4	6.6 (N=753)	24.7 (N=998)	38.8	36.4	73.8	**97.8**	22.9 (N=993)	42.5	59.7	66.1	**92.3**	**100**
	5%	18.8	28.6	66.1	77.9	**98.3**	**100**	39.1	70.5	**95.3**	**98.1**	**100**	**100**	69.6	**91.5**	**99.9**	**100**	**100**	**100**
	10%	30.9	68.8	**95.4**	**97.4**	**100**	**100**	72.8	**94.1**	**100**	**100**	**100**	**100**	**92.4**	**99.7**	**100**	**100**	**100**	**100**
Month 1	**1%**	8.9 (N=774)	9.7 (N=929)	8.7	8.3	6.7	11.5	7.8	10.6	11.5	11.3	16.2	26.8	11.98 (N=993)	15.5	18.0	19.3	28.6	61.2
	5%	8.4	7.0	11.6	11.8	22.7	49.5	11.6	15.2	22.2	27.6	56.9	**94.4**	18.9	28.3	50.1	56.9	**88.1**	**100**
	10%	6.9	10.5	17.2	20.1	43.6	**88.3**	15.2	22.8	46.8	55.4	**86.8**	**100**	26.8	47.9	77.3	**87.5**	**99.9**	**100**
Month 3	**1%**	8.0 (N=763)	11.8 (N=944)	5.7	5.1	4.4	3.8	6.7 (N=757)	9.2	5.4	6.3	4.7	4.8	8.9	10.3	7.7	5.0	6.2	9.5
	5%	6.7	3.6	3.5	4.9	5.9	9.3	5.9	3.5	5.2	4.8	9.0	20.1	6.1	6.8	9.0	10.8	18.1	45.3
	10%	4.4	4.4	4.2	5.7	8.1	17.2	5.0	6.1	7.0	9.4	17.1	43.5	5.6	9.2	13.7	18.6	35.1	**86.2**
Month 6	**1%**	7.8 (N=740)	11.2 (N=932)	10.6	10.6	9.1	9.7	8.2 (N=754)	14.4	14.1	11.4	12.5	15.8	9.5	18.0	20.6	19.2	23.3	35.0
	5%	11.4	8.5	9.2	7.7	9.0	16.7	13.5	10.6	13.2	15.8	23.3	51.1	20.2	23.7	30.8	38.7	58.9	**92.6**
	10%	8.9	9.4	10.6	10.7	16.8	34.6	13.8	15.6	22.9	25.0	43.9	**82.0**	22.3	30.7	53.5	63.4	**87.8**	**99.8**
Month 11	**1%**	6.2 (N=745)	11.8 (N=939)	11.9	10.5	7.5	8.4	8.2 (N=771)	13.4	15.0	13.3	14.3	21.9	14.6	21.5	19.8	20.4	32.1	59.7
	5%	11.6	8.7	8.9	8.3	13.2	30.4	14.2	14.8	19.7	25.7	43.5	**83.3**	22.3	32.1	49.5	61.7	**89.3**	**99.8**
	10%	6.0	8.5	13.3	15.8	25.8	57.4	10.9	20.7	36.5	45.8	73.0	**98.9**	28.7	51.5	**83.4**	**89.8**	**99.3**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.8: **Power** by scenario for the **Cox proportional hazards model and the Grambsch-Therneau test for disproportionality** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	0.0 (N=998)	2.2 (N=998)	5.0	6.1	7.9	17.2	0.0 (N=993)	4.9 (N=993)	11.9	10.1	22.7	47.3	0.0 (N=984)	11.5 (N=998)	21.8	25.2	51.8	**90.0**
	5%	6.7	8.0	14.5	15.9	34.6	77.3	9.3	21.0	41.4	49.7	**79.6**	**99.6**	29.6	49.9	**81.5**	**88.3**	**99.8**	**100**
	10%	7.4	16.4	28.5	35.2	66.8	**96.3**	22.2	40.5	69.4	**79.5**	**98.7**	**100**	48.2	**79.0**	**97.8**	**99.8**	**100**	**100**
Month 1	**1%**	0.0 (N=989)	2.0 (N=999)	5.8	6.5	7.3	15.5	0.0 (N=986)	6.4	8.4	8.0	18.0	36.4	0.0 (N=979)	11.85 (N=996)	20.3	25.1	44.1	**81.1**
	5%	5.6	7.1	13.2	15.6	29.0	59.8	8.2	18.4	31.7	37.3	69.9	**97.3**	22.8	39.8	70.6	**79.7**	**97.9**	**100**
	10%	7.5	11.3	22.2	24.6	52.7	**90.2**	16.8	29.7	58.4	68.4	**92.2**	**100**	39.3	65.9	**92.2**	**96.4**	**100**	**100**
Month 3	**1%**	0.0 (N=992)	2.2 (N=993)	4.5	4.4	6.9	8.0	0.0 (N=990)	5.3 (N=999)	6.9	8.1	11.9	21.0	0.0 (N=983)	9.2 (N=998)	14.7 (N=999)	15.4	23.3	51.2
	5%	6.2	5.4	8.4	9.8	16.7	29.6	7.7	11.3	16.5	20.6	37.1	74.8	13.7	24.6	41.7	45.9	72.6	**97.6**
	10%	5.5	8.5	10.8	14.4	24.7	49.4	10.4	17.5	27.5	34.0	62.1	**94.3**	23.7	38.0	61.3	71.3	**93.9**	**100**
Month 6	**1%**	0.0 (N=990)	2.1 (N=994)	5.2	5.6	5.5	6.7	0.0 (N=985)	4.9 (N=999)	5.9	4.2	6.4	7.5	0.0 (N=985)	9.3 (N=997)	8.3	8.8	8.4	10.3
	5%	6.2	6.5	5.6	5.0	6.4	6.0	6.7	5.3	5.7	4.3	6.4	6.0	10.1	8.7	7.1	7.4	8.2	10.0
	10%	5.3	5.4	6.0	5.6	6.1	6.0	6.2	5.8	6.4	6.5	7.3	8.1	8.7	9.1	8.9	8.7	9.6	9.4
Month 11	**1%**	0.0 (N=991)	2.5 (N=998)	5.4	5.4	8.1	13.2	0.0 (N=986)	5.92 (N=997)	10.3	8.2	17.7	37.3	0.0 (N=978)	12.36 (N=995)	17.4	21.9	42.2	**80.2**
	5%	5.7	7.8	13.5	14.5	28.3	63.4	9.6	18.5	32.6	40.6	72.4	**98.1**	24.7	42.4	69.4	**82.4**	**98.5**	**100**
	10%	6.7	12.4	25.1	29.1	52.2	**93.6**	16.3	32.5	63.2	73.4	**95.8**	**100**	41.1	73.3	**96.4**	**98.8**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.9: **Power** by scenario for the **double-Cox proportional hazards model and the Grambsch-Therneau test for disproportionality** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	0.0 (N=988)	0.0	3.2	3.0	8.8	21.2	0.0 (N=993)	0.0 (N=999)	7.1	8.2	28.8	57.8	0.0 (N=993)	3.5 (N=999)	25.7	32.9	60.7	**96.1**
	5%	3.7	6.3	19.3	21.6	47.5	**94.0**	7.1	23.5	52.2	62.9	**92.2**	**100**	33.6	59.6	**91.1**	**95.3**	**100**	**100**
	10%	7.9	19.3	40.1	48.5	**84.2**	**99.5**	24.9	53.4	**84.7**	**92.6**	**99.9**	**100**	57.9	**88.3**	**99.8**	**100**	**100**	**100**
Month 1	**1%**	0.0 (N=989)	0.0	2.7	4.3	8.0	15.7	0.0 (N=986)	0.0	6.5	6.4	20.0	41.2	0.0 (N=989)	3.7 (N=998)	22.8	27.9	47.9	**85.7**
	5%	2.8	7.9	15.7	16.9	32.9	66.7	6.8	20.8	34.9	41.4	75.7	**99.4**	23.8	45.9	75.2	**83.6**	**98.7**	**100**
	10%	6.6	13.9	23.1	30.7	60.6	**96.5**	18.4	34.2	65.5	73.9	**95.9**	**100**	42.4	71.3	**95.5**	**98.8**	**100**	**100**
Month 3	**1%**	0.0 (N=992)	0.0 (N=998)	2.0	2.2	6.0	6.3	0.0 (N=990)	0.0	5.6	5.9	10.8	15.5	0.0 (N=990)	3.6	12.5	15.2	23.6	43.8
	5%	3.5	5.2	6.4	8.9	13.9	21.9	5.4	9.2	14.5	16.1	29.6	66.6	14.7	23.8	34.2	39.2	67.7	**96.6**
	10%	4.6	8.5	9.6	9.5	17.4	42.2	9.9	14.9	21.8	28.3	52.7	**91.7**	22.2	35.2	55.3	66.2	**91.6**	**99.8**
Month 6	**1%**	0.0 (N=990)	0.0 (N=996)	2.8	3.3	5.1	9.9	0.0 (N=985)	0.0	4.2	3.8	8.3	18.2	0.0 (N=991)	1.7 (N=997)	7.9	11.0	21.3	42.7
	5%	3.9	6.8	9.1	8.2	13.1	28.9	5.6	9.3	13.6	17.0	32.6	75.8	11.6	22.0	36.3	44.8	76.4	**99.5**
	10%	5.8	7.1	11.2	12.6	25.2	58.6	11.2	14.7	28.6	35.4	66.4	**97.7**	19.9	34.3	66.8	**79.5**	**98.1**	**100**
Month 11	**1%**	0.0 (N=991)	0.0	2.2	3.3	6.7	9.0	0.0 (N=986)	0.0 (N=999)	5.0	4.2	12.9	26.8	0.0 (N=987)	4.1 (N=998)	10.3	13.5	32.3	72.7
	5%	3.1	6.6	10.3	11.4	20.0	53.6	4.6	12.8	24.0	32.0	61.2	**97.1**	15.5	31.0	59.1	74.7	**97.6**	**100**
	10%	4.4	10.5	18.6	21.7	41.4	**86.1**	10.6	24.8	52.4	63.5	**92.6**	**100**	28.6	62.1	**95.0**	**97.6**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.10: **Power** by scenario for the **combined test** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	0.0 (N=412)	10.9	4.7	4.7	5.0	22.7	0.0 (N=437)	6.5	5.1	6.2	25.1	**93.4**	4.2 (N=984)	5.8	15.1	28.7	**97.9**	**100**
	5%	5.0	5.3	19.5	24.2	67.0	**99.7**	8.4	25.3	**87.9**	**94.5**	**100**	**100**	32.3	**98.6**	**100**	**100**	**100**	**100**
	10%	6.0	21.6	54.2	67.2	**99.3**	**100**	30.5	**90.1**	**100**	**100**	**100**	**100**	**99.0**	**100**	**100**	**100**	**100**	**100**
Month 1	**1%**	0.0 (N=431)	14.0	3.6	4.5	4.7	18.2	0.0 (N=384)	6.0	4.3	6.9	23.3	63.3	3.8 (N=979)	6.9	16.6	23.3	72.1	**99.4**
	5%	4.1	6.9	14.9	19.5	43.3	**90.8**	6.4	21.4	54.4	64.5	**97.7**	**100**	29.1	77.0	**98.9**	**99.8**	**100**	**100**
	10%	5.7	17.5	34.7	41.4	**83.0**	**99.9**	27.1	56.5	**92.8**	**98.8**	**100**	**100**	**81.0**	**99.4**	**100**	**100**	**100**	**100**
Month 3	**1%**	0.0 (N=411)	11.3	4.9	3.0	6.5	12.8	0.0 (N=431)	6.5	5.0	5.9	15.8	36.5	3.7	7.4	16.7	23.4	51.8	**93.2**
	5%	5.0	4.7	11.0	14.2	24.6	60.3	7.3	15.8	31.4	39.4	74.8	**99.7**	28.1	53.8	**87.4**	**95.4**	**99.9**	**100**
	10%	5.7	11.4	20.1	25.2	52.0	**89.2**	18.0	37.6	64.2	76.5	**98.5**	**100**	60.0	**92.1**	**99.7**	**100**	**100**	**100**
Month 6	**1%**	0.0 (N=424)	12.7	3.3	5.1	3.7	10.8	0.0 (N=399)	6.1	4.3	5.0	13.6	26.9	3.4	5.8	15.5	17.5	39.0	**82.5**
	5%	3.5	5.3	9.0	12.0	18.2	44.3	5.6	13.5	26.0	29.6	60.6	**95.8**	21.2	42.2	76.9	**85.7**	**100**	**100**
	10%	5.8	9.7	15.8	17.4	36.9	76.2	14.8	25.8	54.2	62.3	**91.5**	**100**	46.3	**80.2**	**99.2**	**99.9**	**100**	**100**
Month 11	**1%**	0.0 (N=422)	12.3	4.4	5.3	4.9	10.7	0.0 (N=439)	6.7	4.7	7.0	12.2	26.3	4.2	6.9	13.6	17.5	38.0	**83.5**
	5%	4.0	5.6	7.1	9.7	15.6	37.4	5.0	11.3	23.9	26.1	57.5	**94.7**	17.2	40.1	73.7	**82.1**	**99.3**	**100**
	10%	4.7	8.5	14.5	16.6	34.4	67.5	11.8	22.9	45.3	55.7	**88.6**	**100**	40.6	74.8	**97.0**	**99.7**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.11: **Power** by scenario for the **Fisher's exact test** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						50%						100%					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	0.0	0.9	4.6	3.1	3.2	10.2	0.0	1.4	4.0	5.9	11.9	27.5	0.0	5.8	12.6	17.4	38.1	**88.0**
	5%	2.9	4.9	9.9	9.4	18.3	48.4	7.6	10.6	27.1	30.6	63.2	**96.6**	17.3	38.8	**80.3**	**88.7**	**99.3**	**100**
	10%	5.0	8.8	16.5	20.3	42.5	**79.2**	11.5	30.0	53.8	66.9	**94.7**	**100**	49.5	**84.8**	**98.7**	**99.9**	**100**	**100**
Month 1	**1%**	0.0	1.6	3.2	2.8	2.8	9.9	0.0	2.3	3.7	6.5	13.9	27.9	0.0	6.9	14.0	14.6	40.5	**84.2**
	5%	2.4	5.7	8.5	10.2	18.0	48.0	5.4	11.1	25.7	34.4	64.8	**96.8**	17.7	44.5	**81.2**	**86.7**	**99.4**	**100**
	10%	4.7	9.3	16.8	19.4	40.4	**80.7**	14.5	31.4	54.6	66.4	**93.8**	**100**	48.4	**82.4**	**98.8**	**99.6**	**100**	**100**
Month 3	**1%**	0.0	1.4	4.9	1.8	3.8	11.2	0.0	2.5	3.8	5.4	12.9	29.7	0.0	7.4	13.2	18.9	42.4	**88.1**
	5%	3.0	4.4	9.5	9.7	18.3	48.0	6.2	11.5	23.6	35.4	60.4	**97.5**	19.1	41.4	**78.6**	**88.8**	**99.3**	**100**
	10%	5.0	8.8	16.5	20.3	42.5	**79.2**	11.5	30.0	53.8	66.9	**94.7**	**100**	49.5	**84.8**	**98.7**	**99.9**	**100**	**100**
Month 6	**1%**	0.0	1.1	3.3	3.0	2.4	10.9	0.0	2.6	3.4	5.0	13.6	27.7	0.0	5.8	14.7	15.8	39.5	**86.2**
	5%	2.7	5.4	9.3	11.7	18.5	48.8	5.2	12.7	26.4	33.5	63.1	**96.5**	18.3	40.8	**78.3**	**87.7**	**100**	**100**
	10%	5.1	9.2	17.7	19.0	41.0	**79.3**	12.5	29.0	56.4	65.7	**92.8**	**100**	49.6	**83.3**	**99.6**	**99.7**	**100**	**100**
Month 11	**1%**	0.0	1.5	4.3	3.8	3.2	11.1	0.0	2.0	4.2	6.8	13.2	28.1	0.0	6.9	14.0	17.8	41.7	**88.2**
	5%	2.1	5.6	7.9	10.7	16.9	44.8	4.4	11.9	26.9	33.2	64.7	**96.8**	16.0	43.6	**80.0**	**87.5**	**99.2**	**100**
	10%	5.0	8.8	16.5	20.3	42.5	**79.2**	11.5	30.0	53.8	66.9	**94.7**	**100**	49.5	**84.8**	**98.7**	**99.9**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.12: **Power** by scenario for the Bayesian **beta-binomial model** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						**50%**						**100%**					
Sample size		**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**	**200**	**400**	**800**	**1000**	**2000**	**5000**
Time	**AE background rate**	**Power**																	
Day 1	**1%**	24.9	5.9	14.6	17.7	17.1	31.9	25.6	22.0	23.0	19.6	37.0	62.0	26.2	32.3	42.3	47.5	73.0	**97.4**
	5%	15.8	17.7	26.5	29.4	45.4	75.0	22.5	37.1	54.9	61.7	**86.3**	**99.6**	51.9	74.8	**96.4**	**98.1**	**100**	**100**
	10%	18.8	29.9	43.9	45.8	68.5	**94.2**	41.6	61.8	**83.1**	**88.2**	**99.2**	**100** (n=999)	**79.3**	**96.6**	**99.9**	**100**	**100**	**100**
Month 1	**1%**	22.8	6.8	14.2	15.8	16.3	28.4	21.6	23.9	20.7	20.6	35.8	60.0	24.0	31.5	42.6	47.0	72.9	**97.8**
	5%	14.5	19.5	26.1	30.2	42.3	73.7	22.5	34.6	54.9	63.9	**87.8**	**99.7**	51.0	74.3	**94.9**	**97.4**	**100**	**100**
	10%	20.4	29.6	43.5	47.7	69.0	**93.5**	41.0	57.9	**82.8**	**88.7**	**99.0**	**100**	77.9	**95.5**	**99.9**	**100.0**	**100**	**100**
Month 3	**1%**	24.9	5.9	14.6	17.7	17.1	31.9	22.0	25.3	24.0	19.1	36.8	63.5	26.2	32.3	42.3	47.5	73.0	**97.4**
	5%	15.8	17.7	26.5	29.4	45.4	75.0	23.9	37.7	57.0	62.3	**86.6**	**99.7**	51.9	74.8	**96.4**	**98.1**	**100**	**100**
	10%	18.8	29.9	43.9	45.8	68.5	**94.2**	40.6	60.1	**81.4**	**89.4**	**98.6**	**100**	**79.3**	**96.6**	**99.9**	**100**	**100**	**100**
Month 6	**1%**	24.9	31.2	58.4	64.7	**82.7**	**99.9**	23.4	23.8	22.9	20.8	35.3	60.9	26.1	31.6	37.6	50.3	72.5	**97.0**
	5%	16.1	16.8	28.6	30.3	44.5	71.8	22.5	36.8	55.8	60.9	**87.1**	**99.2**	50.8	78.1	**96.0**	**98.7**	**99.0**	**100**
	10%	18.9	29.0	48.2	48.4	68.0	**93.7**	38.9	58.7	**83.0**	**87.5**	**99.0**	**100**	**80.8**	**96.0**	**99.9**	**100**	**100.0**	**100**
Month 11	**1%**	24.9	5.9	14.6	17.7	17.1	31.9	23.2	22.2	22.7	19.1	35.6	60.8	26.2	32.3	42.3	47.5	73.0	**97.4**
	5%	15.8	17.7	26.5	29.4	45.4	75.0	25.2	34.6	55.8	60.6	**86.9**	**99.4**	51.9	74.8	**96.4**	**98.1**	**100**	**100**
	10%	18.8	29.9	43.9	45.8	68.5	**94.2**	40.2	58.3	**82.1**	**87.0**	**99.0**	**100**	**79.3**	**96.6**	**99.9**	**100**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Table A7.13: **Power** by scenario for the Bayesian **gamma-Poisson model** (N=1000)

		Percentage increase in background event rate due to ADR																	
		25%						50%						100%					
Sample size		200	400	800	1000	2000	5000	200	400	800	1000	2000	5000	200	400	800	1000	2000	5000
Time	AE background rate	Power																	
Day 1	1%	26.1	5.9	14.2	18.2	13.8	33.8	22.9	30.8	14.6	17.8	42.3	51.1	26.0	32.8	37.4	64.2	74.0	**96.5**
	5%	16.4	14.1	30.3	34.4	44.4	72.4	15.0	41.1	57.6	55.6	**86.4**	**98.7**	62.5	72.0	**93.1**	**97.9**	**99.9**	**100**
	10%	12.3	31.3	36.5	46.9	62.9	**91.6**	39.4	55.9	77.5	**87.5**	**98.3**	**100**	75.4	**91.2**	**99.8**	**100**	**100**	**100**
Month 1	1%	24.2	6.5	16.1	15.7	13.2	36.3	23.9	33.0	13.8	17.0	40.7	58.1	24.1	32.2	35.9	60.9	73.7	**96.8**
	5%	19.3	14.6	33.2	34.8	45.9	72.1	17.5	41.3	57.2	54.1	**84.7**	**99.3**	62.3	71.2	**93.0**	**96.6**	**100**	**100**
	10%	14.4	32.9	35.2	45.3	62.8	**91.5**	41.3	54.3	78.3	**87.8**	**98.1**	**100**	74.3	**90.9**	**99.9**	**100**	**100**	**100**
Month 3	1%	26.9	6.3	13.7	17.1	13.8	34.2	26.6	31.5	14.2	17.8	40.5	55.7	26.8	28.1	38.6	61.0	73.8	**97.9**
	5%	17.8	12.4	33.8	33.8	42.7	72.5	17.3	40.1	53.9	57.2	**86.1**	**99.3**	60.9	73.6	**91.8**	**95.8**	**100**	**100**
	10%	12.3	31.3	36.5	46.9	62.9	**91.6**	39.4	55.9	77.5	**87.5**	**98.3**	**100**	75.4	**91.2**	**99.8**	**100.0**	**100**	**100**
Month 6	1%	25.5	6.2	14.6	15.4	12.2	32.0	25.4	31.0	15.0	17.6	41.1	53.3	23.6	31.9	36.0	62.3	75.7	**97.1**
	5%	16.6	14.8	31.0	33.1	48.3	71.2	15.3	39.3	60.6	54.3	**85.4**	**99.5**	62.5	74.4	**93.3**	**96.7**	**100**	**100**
	10%	13.4	30.1	39.0	47.9	64.7	**90.8**	40.5	54.4	78.1	**86.7**	**98.1**	**100**	74.9	**91.9**	**99.8**	**100**	**100**	**100**
Month 11	1%	25.7	4.9	14.2	17.1	14.4	35.5	23.8	29.1	15.6	17.3	41.8	53.2	27.2	33.8	37.2	60.5	75.4	**97.0**
	5%	18.1	12.6	32.9	33.2	48.8	69.4	18.9	39.5	54.8	56.9	**88.0**	**99.0**	60.4	76.9	**92.3**	**97.4**	**100**	**100**
	10%	12.3	31.3	36.5	46.9	62.9	**91.6**	39.4	55.9	77.5	**87.5**	**98.3**	**100**	75.4	**91.2**	**99.8**	**100**	**100**	**100**

Note: Red, bolded figures indicate scenarios where power exceeds 80%. Where model failures occurred and N ≠ 1000, complete N is presented in brackets after the figure for power.

Appendix A7.4: False positives across scenarios

A7.14: False positives for each test by sample size and increases in background events rates of 25%, 50% & 100% over background event rates of 1%, 5% & 10% at day 1, month 1, 3, 6 & 11

			Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality**		**Double-Cox PH model with GT test for disproportionality**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
			N=15000															
	Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
Over percentage increases	200	False positive	2,410	**0.18**	2,070	**0.15**	510	**0.03**	400	**0.03**	455	**0.03**	325	**0.02**	2,225	**0.15**	2,195	**0.15**
		Model fail	1335		1265		75		75		40		0		0		0	
	400	False positive	2,500	**0.17**	2,120	**0.14**	580	**0.04**	465	**0.03**	1,055	**0.07**	535	**0.04**	1,300	**0.09**	1,315	**0.09**
		Model fail	320		305		10		5		0		0		0		0	
	800	False positive	2,650	**0.18**	2,265	**0.15**	810	**0.05**	640	**0.04**	805	**0.05**	525	**0.04**	1,660	**0.11**	1,585	**0.11**
		Model fail	15		15		0		0		0		0		0		0	
	1000	False positive	2,450	**0.16**	1,945	**0.13**	735	**0.05**	535	**0.04**	575	**0.04**	440	**0.03**	1,605	**0.11**	1,320	**0.09**
		Model fail	0		0		0		0		0		0		0		0	
	2000	False positive	2,430	**0.16**	1,920	**0.13**	800	**0.05**	675	**0.05**	860	**0.06**	720	**0.05**	1,690	**0.11**	1,615	**0.11**
		Model fail	0		0		0		0		0		0		5		0	
	5000	False positive	2,315	**0.15**	1,860	**0.12**	785	**0.05**	805	**0.05**	700	**0.05**	680	**0.05**	1,485	**0.10**	1,410	**0.09**
		Model fail	0		0		0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: PH – proportional hazards; GT - Grambsch-Therneau
Note: The rates of false positives do not change with varying increases in event rates and are presented for information only

Table A.15: False positives for each test by sample and time over: background rates of 1%, 5% & 10% & increases in background rates due to ADRs of 25%, 50% & 100%

			Weibull model with ancillary parameter		**Double-Weibull model with ancillary parameter**		**Cox PH model with GT test for disproportionality**		**Double-Cox PH model with GT test for disproportionality**		**Combined test**		**Fisher's exact test**		**Bayesian beta-binomial model**		**Bayesian gamma-Poisson model**	
			N=9000															
Time	Sample size		n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N	n	n/N
Across all scenarios	200	False positive	1,446	**0.18**	1,242	**0.15**	306	**0.03**	240	**0.03**	273	**0.03**	195	**0.02**	1,335	**0.15**	1,317	**0.15**
		Model fail	801		759		45		45		24		0		0		0	
	400	False positive	1,500	**0.17**	1,272	**0.14**	348	**0.04**	279	**0.03**	633	**0.07**	321	**0.04**	780	**0.09**	789	**0.09**
		Model fail	192		183		6		3		0		0		0		0	
	800	False positive	1,590	**0.18**	1,359	**0.15**	486	**0.05**	384	**0.04**	483	**0.05**	315	**0.04**	996	**0.11**	951	**0.11**
		Model fail	9		9		0		0		0		0		0		0	
	1000	False positive	1,470	**0.16**	1,167	**0.13**	441	**0.05**	321	**0.04**	345	**0.04**	264	**0.03**	963	**0.11**	792	**0.09**
		Model fail	0		0		0		0		0		0		0		0	
	2000	False positive	1,458	**0.16**	1,152	**0.13**	480	**0.05**	405	**0.05**	516	**0.06**	432	**0.05**	1,014	**0.11**	969	**0.11**
		Model fail	0		0		0		0		0		0		3		0	
	5000	False positive	1,389	**0.15**	1,116	**0.12**	471	**0.05**	483	**0.05**	420	**0.05**	408	**0.05**	891	**0.10**	846	**0.09**
		Model fail	0		0		0		0		0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Acronyms: PH – proportional hazards; GT - Grambsch-Therneau
Note: The rates of false positives do not change with the time of increase and are presented for information only

Appendix A7.5: Power of Bayesian methods with varying thresholds of risk to detect

Table A7.16: **Power of beta-binomial model** to detect signals of varying sizes sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Beta-Binomial (P(RR>1.0) ≥ 0.9)		**Beta-Binomial** (P(RR>1.25) ≥ 0.9)		**Beta-Binomial** (P(RR>1.5) ≥ 0.9)		**Beta-Binomial** (P(RR>2.0) ≥ 0.9)	
		N=45,000							
Sample size		n	n/N	n	n/N	n	n/N	n	n/N
200	Power	15,136	**0.34**	9,255	**0.21**	5,354	**0.12**	1,630	**0.04**
	Model fail	0				0		0	
400	Power	19,062	**0.42**	11,407	**0.25**	5,744	**0.13**	1,786	**0.04**
	Model fail	1		1		1		1	
800	Power	24,648	**0.55**	15,358	**0.34**	7,803	**0.17**	1,579	**0.04**
	Model fail	0		0		0		0	
1000	Power	25,954	**0.58**	15,921	**0.35**	8,527	**0.19**	1,348	**0.03**
	Model fail	0		0		0		0	
2000	Power	31,900	**0.71**	19,560	**0.43**	10,996	**0.24**	1,296	**0.03**
	Model fail	1		1		1		1	
5000	Power	38,560	**0.86**	24,956	**0.55**	13,947	**0.31**	1,334	**0.03**
	Model fail	1		1		1		1	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
Note: P(RR> δ) ≥ 0.9 is the probability that a predefined 'tolerable risk ratio' (δ) is crossed where $\delta = 1.0, 1.25, 1.5$ & 2.0 have been investigated

Table A7.17: **False positive rate for the beta-binomial model** by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Beta-Binomial (P(RR>1.0) ≥ 0.9**)**		**Beta-Binomial (**P(RR>1.25) ≥ 0.9**)**		**Beta-Binomial (**P(RR>1.5) ≥ 0.9**)**		**Beta-Binomial (**P(RR>2.0) ≥ 0.9**)**	
		N=45,000							
Sample size		n	n/N	n	n/N	n	n/N	n	n/N
200	False positive	6,675	**0.15**	3,315	**0.07**	885	**0.02**	90	**0.002**
	Model fail	0		0		0		0	
400	False positive	3,900	**0.09**	1,785	**0.04**	1,275	**0.03**	1,005	**0.02**
	Model fail	0		0		0		0	
800	False positive	4,980	**0.11**	1,185	**0.03**	735	**0.02**	270	**0.01**
	Model fail	0		0		0		0	
1000	False positive	4,815	**0.11**	1,260	**0.03**	600	**0.01**	150	**0.003**
	Model fail	0		0		0		0	
2000	False positive	5,070	**0.11**	705	**0.02**	255	**0.01**	75	**0.002**
	Model fail	15		15		15		15	
5000	False positive	4,455	**0.10**	360	**0.01**	75	**0.002**	0	**0.00**
	Model fail	0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Note: P(RR> δ) ≥ 0.9 is the probability that a predefined 'tolerable risk ratio' (δ) is crossed where $\delta = 1.0, 1.25, 1.5$ & 2.0 have been investigated

Table A7.18: **Power of gamma-Poisson model** to detect signals of varying sizes sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Gamma-Poisson (P(IRR>1.0) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>1.25) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>1.5) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>2.0) ≥ 0.9**)**	
		N=45,000							
Sample size		n	n/N	n	n/N	n	n/N	n	n/N
200	Power	14,996	**0.33**	7,268	**0.16**	4,322	**0.10**	828	**0.02**
	Model fail	0		0		0		0	
400	Power	18,716	**0.42**	9,954	**0.22**	5,192	**0.12**	1,692	**0.04**
	Model fail	1		1		1		1	
800	Power	23,114	**0.51**	15,081	**0.34**	7,890	**0.18**	1,363	**0.03**
	Model fail	3		3		3		3	
1000	Power	25,823	**0.57**	15,528	**0.35**	7,824	**0.17**	1,056	**0.02**
	Model fail	3		3		3		3	
2000	Power	31,143	**0.69**	18,458	**0.41**	10,441	**0.23**	1,092	**0.02**
	Model fail	0		0		0		0	
5000	Power	37,389	**0.83**	23,744	**0.53**	13,077	**0.29**	1,089	**0.02**
	Model fail	1		1		1		1	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios with a signal = power
Model fail indicate either failure to estimate parameters or non-convergence
Note: P(IRR> δ) ≥ 0.9 is the probability that a predefined 'tolerable incident rate ratio' (δ) is crossed where $\delta = 1.0, 1.25, 1.5$ & 2.0 have been investigated

Table A7.19: **False positive rate for the gamma-Poisson model** by sample size over: AE background rates of 1%, 5% & 10%, with increases in background rate of 25%, 50% & 100% due to ADRs, at day 1, month 1, 3, 6 & 11 (relative to a 12 month trial)

		Gamma-Poisson (P(IRR>1.0) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>1.25) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>1.5) ≥ 0.9**)**		**Gamma-Poisson (**P(IRR>2.0) ≥ 0.9**)**	
		N=45,000							
Sample size		n	n/N	n	n/N	n	n/N	n	n/N
200	False positive	6,585	**0.15**	1,530	**0.03**	1,020	**0.02**	135	**0.003**
	Model fail	0		0		0		0	
400	False positive	3,945	**0.09**	2,115	**0.05**	1,380	**0.03**	1,005	**0.02**
	Model fail	0		0		0		0	
800	False positive	4,755	**0.11**	810	**0.02**	525	**0.01**	45	**0.001**
	Model fail	0		0		0		0	
1000	False positive	3,960	**0.09**	1,215	**0.03**	975	**0.02**	150	**0.003**
	Model fail	0		0		0		0	
2000	False positive	4,845	**0.11**	1,245	**0.03**	405	**0.01**	90	**0.002**
	Model fail	0		0		0		0	
5000	False positive	4,230	**0.09**	255	**0.01**	45	**0.001**	0	**0.00**
	Model fail	0		0		0		0	

n: number of simulations indicating a signal; n/N: number of simulations indicating a signal/total number of simulated scenarios without a signal = false positives
Model fail indicate either failure to estimate parameters or non-convergence
Note: P(IRR> δ) ≥ 0.9 is the probability that a predefined 'tolerable incident rate ratio' (δ) is crossed where $\delta = 1.0, 1.25, 1.5$ & 2.0 have been investigated

Appendix A.8: Table of permissions for all reused copyrighted works in this thesis

Thesis page no.	Thesis title (and type)	Name of original work	Source of original work	Copyright holder and contact	Date requested permission	I have permission yes /no	Permission note
45	Table 2.2	Table 1: Characteristics of included studies	Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open **9**(2): e024537).	*© Author(s) (or their employer(s)) 2019. Open access.*	NA	Yes	I am the first author of this published work licensed with a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. See appendix A.9 item (i) for full terms and conditions
48	Table 2.4	Table 2: Collection, assessment and analysis methods reported by studies					
51 and 55	Table 2.6 and table 2.9	Table 3: Summaries of results presented by studies					
84	Figure 3.1	Fig. 1 Flow diagram describing the assessment of sources of evidence	Phillips, R., et al. (2020). "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy." BMC Medical Research Methodology 20(1): 288	*© Author(s). 2020 Open access.*	NA	Yes	I am the first author of this published work licensed with a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/. See appendix A.9 item (ii) for full terms and conditions
86	Figure 3.2	Fig. 2 Taxonomy of methods for adverse event (AE) analysis					
87	Table 3.2	Table 1 Summary level classifications					
88	Table 3.3	Table 2 Article classifications					
111	Figure 3.3	Fig. 4 Visual representations of AE data for case study 2—randomised controlled trial of GDNF in Parkinson's disease.	Cornelius, V., et al. (2020). "Advantages of visualisations to evaluate and communicate adverse event information in	*© Author(s). 2020 Open access.*	NA	Yes	I am the last author on this published work licensed with a Creative Commons Attribution 4.0 International License

			randomised controlled trials." Trials 21(1): 1028				https://creativecommons.org/licenses/by/4.0/. See appendix A.9 item (ii) for full terms and conditions
128	Figure 4.1	Figure A1: Flow diagram of participation	Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open 10(6): e036875	*© Author(s) (or their employer(s)) 2020. Open access.*	NA	Yes	I am the first author of this published work licensed with a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. See appendix A.9 item (iii) for full terms and conditions
130	Figure 4.2	Figure 1 Participant characteristics by sector and overall. CRO, clinical research organisation; CTUs, clinical trials units; pharma, pharmaceuticals.					
131	Table 4.1	Table 1 Participant characteristics by sector and overall					
134	Table 4.2	Table 2 AE information typically presented by sector and overall					
137	Table 4.5	Table A4: Reasons specialist adverse event (AE) methods are not used (of participants aware of such methods)					
138	Table 4.6	Table A5: Classification of participants' comments on the reasons for a lack of use of specialist methods for adverse event (AE) analysis					
143	Table 4.7	Table A6: Influences the analysis performed					

144	Table 4.8	Table A7: Barriers when analysing adverse events (AEs)					
145	Table 4.9	Table A8: Opinions on adverse event (AE) analysis					
148	Table 4.10	Table A9: Concerns about available methods for adverse event (AE) analysis					
149	Table 4.11	Table A10: Solutions to support a change in adverse event (AE) analysis practice					
150	Table 4.12	Table A11: Classification of participants' comments on solutions to support change in adverse event (AE) analysis practices					
153	Table 4.13	Table A12: Classification of participants' general comments raised regarding adverse event (AE) analysis practices					
178 and 386	Figure 5.1 image 11 and figure A5.7	Figure 2. A tendril plot of all AEs in the RE2SPOND trial.	Karpefors, M. and J. Weatherall (2018). "The Tendril Plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials." Journal of the American Medical Informatics Association 25(8): 1069-1073	© Oxford University Press	14 June 2021	Yes	See appendix A.10 item (i) for full terms and conditions

178 and 387	Figure A5.9b	FIGURE A2 Level plots of treatment effect in terms of the log-hazard ratio across mutually disjoint subgroups defined by age and weight categorised in three levels.	Ballarini NM, Chiu Y-D, König F, et al. A critical review of graphics for subgroup analyses in clinical trials. Pharmaceutical Statistics 2020 25 March 2020. DOI: 10.1002/pst.2012	© The Author(s) 2020	NA	Yes	Work licensed under a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. Open access. See appendix A.10 item (ii) for full terms and conditions
388	Figure A5.10c	Figure 5: All-cause and treatment-related adverse events in the safety population. (A) All-cause adverse events that differed by 5% or more between study groups.	Fehrenbacher, L., et al. (2016). "Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial." The Lancet 387(10030): 1837-1846	© Elsevier	14 June 2021	Yes	See appendix A.10 item (iii) for full terms and conditions
178 and 390	Figure 5.1 image 8 and figure A5.12	Figure 5. Star-plot of mean values of the 30 PANSS Items by treatment.	Squassante, L., et al. (2006). "Simple graphical methods of displaying multiple clinical results." Pharm Stat 5(1): 51-60 with permission from John Wiley & Son.	© John Wiley & Sons	14 June 2021	Yes	See appendix A.10 item (iv) for full terms and conditions
178 and 390	Figure 5.1 image 10 and figure A5.13	Figure 4. Alluvial graph of disease or symptom prevalence across decades of life	Salvi S, Apte K, Madas S, et al. Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: a 1-day point prevalence study (the POSEIDON study). Lancet Glob	© The Author(s) 2015	NA	Yes	Work licensed under a Creative Commons CC BY-NC-ND Licensce. https://creativecommons.org/licenses/by-nc-nd/4.0/ Open Access

			Health. 2015;3(12):e776-e784. doi:10.1016/S2214-109X(15)00152-				See appendix A.10 item (v) for full terms and conditions
181 and 400	Figure 5.3 image 2 and figure A5.30	Figure 1: Incidence of diarrhoea in patients given FOLFOX and IROX in NCCTG N9741 by drug cycle and adverse event grade	Thanarajasingam, G., et al. (2016). "Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254." Lancet Oncology 17(5): 663-670	© Elsevier	14 June 2021	Yes	See appendix A.10 item (vi) for full terms and conditions
182 and 392	Figure 5.4 image 3 and figure A5.17	Figure 3: Time-to-event analyses for adverse events using the Toxicity over Time package (B) Median time to first occurrence and worst grade toxic effects in patients given IROX in NCCTG N9741.	Thanarajasingam, G., et al. (2018). "Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies." The Lancet Haematology 5(11): e563-e598	© Elsevier	14 June 2021	Yes	See appendix A.10 item (vii) for full terms and conditions
184 and 402	Figure 5.6 image 3 and figure A5.35	Figure 2 MCF of all AEs.	Siddiqui, O. (2009). "Statistical methods to analyze adverse events data of randomized clinical trials." Journal of Biopharmaceutical Statistics 19(5): 889-899	© Taylor & Francis.	14 June 2021	Yes	Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published. See appendix A.10 item (viii) for full terms and conditions

184 and 402	Figure 5.6 image 4 and figure A5.34	Figure 1 Mean duration of exacerbation with 95% CI based on the robust variance estimate for recurrent pulmonary exacerbations in patients with fibrosis treated with placebo and rhDNase	Wang, J. and G. Quartey (2012). "Nonparametric estimation for cumulative duration of adverse events." Biometrical Journal 54(1): 61-74 with permission from John Wiley & Sons	© John Wiley & Sons.	14 June 2021	Yes	See appendix A.10 item (ix) for full terms and conditions
186 and 403	Figure 5.8 image 2 and figure A5.37	Figure 2 e-DISH-like plot.	Xia HA, Crowe BJ, Schriver RC, Oster M, Hall DB. Planning and core analyses for periodic aggregate safety data reviews. Clin Trials. 2011;8(2):175-182. doi:10.1177/1740774510395635	© Sage Publishing	14 June 2021	Yes	Permission is granted at no cost for use of content in a Master's Thesis and/or Doctoral Dissertation, subject to the following limitations. You may use a single excerpt or up to 3 figures tables. If you use more than those limits, or intend to distribute or sell your Master's Thesis/Doctoral Dissertation to the general public through print or website publication, please return to the previous page and select 'Republish in a Book/Journal' or 'Post on intranet/password-protected website' to complete your request. See appendix A.10 item (x) for full terms and conditions

186 and 404	Figure 5.8 image 3 and figure A5.38	Figure 2	Trost, D. C. and J. W. Freston (2008). "Vector Analysis to Detect Hepatotoxicity Signals in Drug Development." Therapeutic Innovation & Regulatory Science 42(1): 27-34	© 2008 Drug Information Association. Inc.	14 June 2021	Yes	Work licensed under a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. Open access See appendix A.10 item (xi) for full terms and conditions
188 and 399	Figure 5.9 image 9 and A5.29	Figure 2a. Side-by-side Delta plots of baseline and last follow-up platelet values	Chuang-Stein, C., et al. (2001). "Recent Advancements in the Analysis and Presentation of Safety Data." Drug Information Journal 35(2): 377-397	© 2001 Drug Information Association. Inc.	14 June 2021	Yes	Work licensed under a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. Open access See appendix A.10 item (xii) for full terms and conditions
387	Figure A5.9a	Figure 7. A simultaneous grade change in ALT and bilirubin	Chuang-Stein, C., et al. (2001). "Recent Advancements in the Analysis and Presentation of Safety Data." Drug Information Journal 35(2): 377-397	© 2001 Drug Information Association. Inc.	14 June 2021	Yes	Work licensed under a Creative Commons CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/. Open access See appendix A.10 item (xiii) for full terms and conditions

Appendix A.9: Copies of permission documents to republish copyrighted works (my own work)

i) Some of the work presented in chapter two has been published in BMJ Open and has been referenced accordingly (Phillips, R., et al. (2019). "Analysis and reporting of adverse events in randomised controlled trials: a review." BMJ Open **9**(2): e024537). Including tables 2.2, 2.4, 2.6 and 2.9. Copyright information:

© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY. Published by BMJ.This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/

ii) Some of the work presented in chapter three has been published in BMC Medical Research Methodology and Trials and has been referenced accordingly (Phillips, R., et al. (2020). "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy." BMC Medical Research Methodology **20**(1): 288 and Cornelius, V., et al. (2020). "Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials." Trials **21**(1): 1028.). Including figures 3.1, 3.2 and 3.3, and tables 3.2 and 3.3. Copyright information for BMC Medical Research Methodology and Trials:

Open Access *This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data*

iii) Some of the work published in chapter four has been published in BMJ Open and has been referenced accordingly (Phillips, R. and V. Cornelius (2020). "Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry." BMJ Open **10**(6): e036875.) Including figure 4.1 and 4.2, and tables 4.1, 4.2, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12 and 4.13 Copyright information:

© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY. Published by BMJ. https://creativecommons.org/licenses/by/4.0/This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

Appendix A.10: Copies of permission documents to republish third party copyrighted works

Several of the images included in chapter five and the associated appendix have been reprinted from published articles. Copyright information for each image is detailed below:

i) Figure 5.1 image 11 and figure A5.7

Permission sought via Rights Link and terms of use detailed below:

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number 5087571008287
License date Jun 14, 2021
Licensed content publisher Oxford University Press
Licensed content publication Journal of the American Medical Informatics Association
Licensed content title The Tendril Plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials
Licensed content author Karpefors, Martin; Weatherall, James
Licensed content date Mar 21, 2018
Type of Use Thesis/Dissertation
Institution name
Title of your work Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Publisher of your work Imperial College London
Expected publication date Sep 2021
Permissions cost 0.00 GBP
Value added tax 0.00 GBP
Total 0.00 GBP
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Figure 2
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID GB125506730
Total 0.00 GBP
Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring

society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.
9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.
10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:
v1.4

ii) Figure A5.9b

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. You are not required to obtain permission to reuse this article. For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions. Permission is not required for this type of reuse.

iii) Figure A5.10c

Permission sought via Rights Link and terms of use detailed below:
ELSEVIER LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.
License Number 5087581401325
License date Jun 14, 2021
Licensed Content Publisher Elsevier
Licensed Content Publication The Lancet
Licensed Content Title Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial
Licensed Content Author Louis Fehrenbacher,Alexander Spira,Marcus Ballinger,Marcin Kowanetz,Johan Vansteenkiste,Julien Mazieres,Keunchil Park,David Smith,Angel Artal-

Cortes,Conrad Lewanski,Fadi Braiteh,Daniel Waterkamp,Pei He,Wei Zou,Daniel S Chen,Jing Yi,Alan Sandler et al.
Licensed Content Date 30 April–6 May 2016
Licensed Content Volume 387
Licensed Content Issue 10030
Licensed Content Pages 10
Start Page 1837
End Page 1846
Type of Use reuse in a thesis/dissertation
Portion figures/tables/illustrations
Number of figures/tables/illustrations 1
Format electronic
Are you the author of this Elsevier article? No
Will you be translating? No
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Image A Figure 5
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID GB 494 6272 12
Total 0.00 GBP
Terms and Conditions
INTRODUCTION
1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).
GENERAL TERMS
2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. For book authors the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. Posting to a repository: Authors are permitted to post a summary of their chapter only in their institution's repository.

19. Thesis/Dissertation: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted

publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. Other Conditions:

v1.10

iv) Figure 5.1 image 8 and figure A5.12

Permission sought via Rights Link and terms of use detailed below:

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5087590466434
License date	Jun 14, 2021
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Pharmaceutical Statistics
Licensed Content Title	Simple graphical methods of displaying multiple clinical results

Licensed Content Author R. L. Palmer, C. N. Robinson, L. Squassante
Licensed Content Date Feb 24, 2006
Licensed Content Volume 5
Licensed Content Issue 1
Licensed Content Pages 10
Type of use Dissertation/Thesis
Requestor type University/Academic
Format Electronic
Portion Figure/table
Number of figures/tables 1
Will you be translating? No
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Figure 5
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID EU826007151
Total 0.00 GBP
Terms and Conditions
TERMS AND CONDITIONS
This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, and any CONTENT (PDF or image file) purchased as part of your order, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.
- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts, You

may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto
- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.
- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.
- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.
- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.
- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.
- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.
- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.
- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library

http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

v) Figure 5.1 image 10 and figure A5.13

This article is published under the terms of the Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND). For non-commercial purposes you may copy and distribute the article, use portions or extracts from the article in other works, and text or data mine the article, provided you do not alter or modify the article without permission from Elsevier. You may also create adaptations of the article for your own personal use only, but not distribute these to others. You must give appropriate credit to the original work, together with a link to the formal publication through the relevant DOI, and a link to the Creative Commons user license above. If changes are permitted, you must indicate if any changes are made but not in any way that suggests the licensor endorses you or your use of the work.

Permission is not required for this non-commercial use. For commercial use please continue to request permission via Rightslink.

vi) Figure 5.3 image 2 and figure A5.30

Permission sought via Rights Link and terms of use detailed below:

ELSEVIER LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number 5087600327500
License date Jun 14, 2021
Licensed Content Publisher Elsevier
Licensed Content Publication The Lancet Oncology
Licensed Content Title Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254
Licensed Content Author Gita Thanarajasingam,Pamela J Atherton,Paul J Novotny,Charles L Loprinzi,Jeff A Sloan,Axel Grothey
Licensed Content Date May 1, 2016
Licensed Content Volume 17
Licensed Content Issue 5
Licensed Content Pages 8
Start Page 663
End Page 670
Type of Use reuse in a thesis/dissertation
Portion figures/tables/illustrations
Number of figures/tables/illustrations 1
Format electronic
Are you the author of this Elsevier article? No
Will you be translating? No
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Figure 1
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID GB 494 6272 12
Total 0.00 GBP
Terms and Conditions
INTRODUCTION
1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).
GENERAL TERMS
2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.
LIMITED LICENSE
The following terms and conditions apply only to specific license types:
15. Translation: This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. Posting licensed content on any Website: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a

scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. For journal authors: the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).
If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:
- immediately
  - o via their non-commercial person homepage or blog
  - o by updating a preprint in arXiv or RePEc with the accepted manuscript
  - o via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - o directly by providing copies to their students or to research collaborators for their personal use
  - o for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - o via non-commercial hosting platforms such as their institutional repository
  - o via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:
- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

vii) Figure 5.4 image 3 and figure A5.17

Permission sought via Rights Link and terms of use detailed below:

ELSEVIER LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number 5087591363061
License date Jun 14, 2021
Licensed Content Publisher Elsevier
Licensed Content Publication The Lancet Haematology
Licensed Content Title Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies
Licensed Content Author Gita Thanarajasingam,Lori M Minasian,Frederic Baron,Franco Cavalli,R Angelo De Claro,Amylou C Dueck,Tarec C El-Galaly,Neil Everest,Jan Geissler,Christian Gisselbrecht,John Gribben,Mary Horowitz,S Percy Ivy,Caron A Jacobson,Armand Keating et al.
Licensed Content Date Nov 1, 2018
Licensed Content Volume 5
Licensed Content Issue 11
Licensed Content Pages 36
Start Page e563
End Page e598
Type of Use reuse in a thesis/dissertation
Portion figures/tables/illustrations
Number of figures/tables/illustrations 1
Format electronic
Are you the author of this Elsevier article? No
Will you be translating? No
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Image B Figure 3
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID GB 494 6272 12
Total 0.00 GBP

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY

COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.
LIMITED LICENSE
The following terms and conditions apply only to specific license types:
15. Translation: This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. Posting licensed content on any Website: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at

v1.10

viii) Figure 5.6 image 3 and figure A5.35

Thesis/Dissertation Reuse Request

Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published.

ix) Figure 5.6 image 4 and figure A5.34

Permission sought via Rights Link and terms of use detailed below:

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Jun 14, 2021

This Agreement between Imperial College London -- Rachel Phillips ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number 5087600920040
License date Jun 14, 2021
Licensed Content Publisher John Wiley and Sons
Licensed Content Publication Biometrical Journal
Licensed Content Title Nonparametric estimation for cumulative duration of adverse events
Licensed Content Author George Quartey, Jixian Wang
Licensed Content Date Dec 14, 2011
Licensed Content Volume 54
Licensed Content Issue 1
Licensed Content Pages 14
Type of use Dissertation/Thesis
Requestor type University/Academic
Format Electronic
Portion Figure/table
Number of figures/tables 1
Will you be translating? No
Title Examining current practice for the analysis and reporting of harm outcomes in phase II and III pharmacology trials: exploring methods to facilitate improved detection of adverse drug reactions
Institution name Imperial College London
Expected presentation date Sep 2021
Portions Figure 1
Requestor Location Imperial College London
Imperial College London
1st Floor Stadium House
68 Wood Lane
London, W12 7RH
United Kingdom
Attn: Rachel Phillips
Publisher Tax ID EU826007151
Total 0.00 GBP
Terms and Conditions
TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

x. Figure 5.8 image 2 and figure A5.37

Gratis Reuse

Permission is granted at no cost for use of content in a Master's Thesis and/or Doctoral Dissertation, subject to the following limitations. You may use a single excerpt or up to 3 figures tables. If you use more than those limits, or intend to distribute or sell your Master's Thesis/Doctoral Dissertation to the general public through print or website publication, please return to the previous page and select 'Republish in a Book/Journal' or 'Post on intranet/password-protected website' to complete your request.

xi. Figure 5.8 image 3 and figure A5.38

Therapeutic Innovation & Regulatory Science articles are published open access under a CC BY licence (Creative Commons Attribution 4.0 International licence). The CC BY licence is the most open licence available and considered the industry 'gold standard' for open access; it is also preferred by many funders. This licence allows readers to copy and redistribute the material in any medium or format, and to alter, transform, or build upon the material, including for commercial use, providing the original author is credited.

xii. Figure 5.9 image 9 and A5.29

Therapeutic Innovation & Regulatory Science articles are published open access under a CC BY licence (Creative Commons Attribution 4.0 International licence). The CC BY licence is the most open licence available and considered the industry 'gold standard' for open access; it is also preferred by many funders. This licence allows readers to copy and redistribute the material in any medium or format, and to alter, transform, or build upon the material, including for commercial use, providing the original author is credited.

xiii. Figure A5.9a

Therapeutic Innovation & Regulatory Science articles are published open access under a CC BY licence (Creative Commons Attribution 4.0 International licence). The CC BY licence is the most open licence available and considered the industry 'gold standard' for open access; it is also preferred by many funders. This licence allows readers to copy and redistribute the material in any medium or format, and to alter, transform, or build upon the material, including for commercial use, providing the original author is credited.