MODELLING OPTIMAL USE OF TESTS FOR MONITORING DISEASE PROGRESSION AND RECURRENCE

by

ALICE JAYNE SITCH

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Institute of Applied Health Research
College of Medical and Dental Sciences
University of Birmingham
January 2019

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Background

Monitoring to identify disease recurrence or progression is common, often with limited evidence to support the tests used, subsequent decisions, frequency and duration of monitoring.

Aims

To develop methods for designing evidence-based monitoring strategies and estimating measurement error, a key consideration in selecting monitoring tests.

Methods

To investigate studies of measurement error: frameworks were identified; design, analysis and reporting of studies were reviewed; a case study was analysed; and, simulation studies were performed to evaluate varying sample size and outlier detection methods. To develop methods for designing monitoring strategies the methods literature was reviewed and simulation models were developed and validated.

Results

Biological variability studies are often poorly designed and reported. Studies are frequently small and may not produce valid results; the required precision of estimates can inform the sample size. Outlier detection can negatively bias variability estimates; methods should be used with caution, with interpretation allowing for potential bias. Modelling monitoring data requires knowledge of the natural history of disease, test performance and measurement error; such evaluation enables selection of evidence-based monitoring strategies prior to full-scale investigation.

Conclusions

Poor monitoring tests can be identified early using small-scale studies and monitoring strategies should be optimised prior to full evaluation.

For my mother

lumen scientiaque

Acknowledgements

First and foremost, I thank Jon Deeks not just for the support and supervision throughout my studies but for the many opportunities in my career. I am also very grateful for the support and guidance from my other supervisors, Jac Dinnes and Sue Mallett.

I am grateful to the members of the Biomarker Pipeline Programme Grant methodology work stream (Jon Deeks, Doug Altman, Jenny Hewison, Jac Dinnes, Paul Baxter, Roberta Longo, Chris McCabe, Julie Parkes and Walter Gregory) for the valuable advice on the monitoring simulation project. I am also thankful to the ELUCIDATE trial management team and the study participants.

I thank the eGFR-C study team and participants for contributing their time and efforts to the eGFR-C substudy included in this thesis. I particularly thank Ed Lamb, the chief investigator, and Ceri Rowe for their advice.

I acknowledge funding from NIHR PGfAR for the Biomarker Pipeline programme grant; from NIHR HTA for the eGFR-C study; and support from the NIHR Birmingham BRC.

I thank De Gruyter for providing access to the journal of Clinical Chemistry and Laboratory Medicine, making the review of biological variability studies possible.

I thank my colleagues and friends Karla Hemming, James Martin, Yemisi Takwoingi and Siân Taylor-Phillips. Thank you for the useful discussions, encouragement and comments on my plots. Also many thanks go to Natasha Maguire for administrative and moral support.

Finally, thank you to my husband, David, whose support is outweighed only by his ability to distract; thank you for making me laugh every day.

Contents

1 Introduction **1**

1.1 Evidence based monitoring 1

1.1.1 Definition of monitoring 2

1.1.2 Monitoring of disease progression and recurrence 2

1.1.3 Monitoring data 5

1.1.4 Measurement error 5

1.1.5 Biological variability 6

1.1.6 Issues in the field 6

1.2 Design and evaluation of biological variability studies 7

1.2.1 Aims of biological variability studies 7

1.2.2 Variability 8

1.3 Designing and evaluating monitoring strategies 11

1.3.1 Strategy design 11

1.3.2 Method for identifying monitoring strategies 13

1.3.3 Assessing monitoring strategies 14

1.3.4 Monitoring studies 16

1.3.5 Outcomes 18

1.3.6 Impact on patients 18

1.3.7 Relationship to screening 19
1.4 Research questions and thesis outline 21
1.4.1 Research questions 21
1.4.2 Thesis outline 21

2 Introduction to assessment of biological variability 25
2.1 Introduction 26
2.2 Methods 27
2.3 Results 28
2.3.1 Design of biological variability studies 29
2.3.2 Analysis of biological variability studies 34
2.4 Summary and conclusions 53

3 A review of biological variability studies 55
3.1 Introduction 56
3.2 Aims and objectives 57
3.3 Methods 57
3.3.1 Searches used to identify studies reporting biological variability studies 57
3.3.2 Eligibility criteria 59
3.3.3 Review of selected studies and data extraction 60
3.4 Results 62
3.4.1 What is the current state of the field? What are the aims of these studies and in which tests and test areas are they seen? 62
3.4.2 How are studies assessing biological variability of tests designed? . . . 66
3.4.3 How are studies assessing biological variability of tests analysed? . . . 74
3.4.4 How are studies assessing biological variability of tests reported? . . . 76
3.4.5 What are the differences between studies assessing biological variability of laboratory, imaging and physiological tests? 81
3.5 Discussion 82
3.5.1 What is the current state of the field? What are the aims of these studies and in which tests and test areas are they seen? 82

3.5.2 How are studies assessing biological variability of tests designed? . . . 82
3.5.3 How are studies assessing biological variability of tests analysed? . . . 83
3.5.4 How are studies assessing biological variability of tests reported? . . . 84
3.5.5 What are the differences between studies assessing biological variability of laboratory, imaging and physiological tests? . . . 85
3.5.6 Limitations . . . 85
3.5.7 Further work . . . 86
3.6 Conclusions . . . 86

4 Analysis of biological variability studies 89
4.1 Introduction . . . 90
4.2 Aims and objectives . . . 91
4.3 Methods . . . 91
4.3.1 Eligibility criteria . . . 91
4.3.2 Study design . . . 92
4.3.3 Sample size . . . 93
4.4 Analysis . . . 93
4.4.1 Sensitivity analyses . . . 94
4.5 Results . . . 95
4.5.1 Study population and completeness of data . . . 95
4.5.2 Analyses using standard laboratory based biological variability methods 96
4.5.3 Sample size . . . 101
4.5.4 Sensitivity analyses to investigate the impact of methods of analysing standard laboratory based biological variability methods . . . 101
4.6 Discussion . . . 107
4.6.1 Sample size . . . 107
4.6.2 Normality and data transformation . . . 108
4.6.3 Outlier detection and removal . . . 109
4.6.4 Limitations . . . 109
4.7 Conclusion and recommendations . . . 110

5 Sample size guidance and justification for studies of biological variation 113

5.1 Introduction . . . 114

5.2 Aims and objectives . . . 116

5.3 Methods . . . 117

5.3.1 Number of simulations . . . 117

5.3.2 Input values . . . 119

5.3.3 Data simulation . . . 119

5.3.4 Analysis . . . 123

5.3.5 Results for each simulation . . . 123

5.3.6 Repeated data simulations and analyses . . . 124

5.3.7 Simulation inputs . . . 125

5.4 Results . . . 126

5.4.1 Number of participants, observations and assessments . . . 126

5.4.2 Analytical, within-individual and between-individual variability . . . 131

5.4.3 Sensitivity analyses . . . 140

5.5 Discussion . . . 140

5.5.1 Limitations . . . 145

5.6 Conclusions . . . 145

6 The impact of outlier detection and removal on studies of biological variability 147

6.1 Introduction . . . 148

6.2 Aims and objectives . . . 150

6.3 Methods . . . 151

6.3.1 Data simulation . . . 151

6.3.2 Outlier detection methods . . . 152

6.3.3 Data analysis . . . 157

6.3.4 Repeated simulations . . . 158

6.4 Results . . . 158

6.4.1 Simulation without outliers . . . 159

6.4.2 Simulation of outlying data . . . 170

6.4.3 Comparison of outlier detection methods . . . 182
6.5 Discussion . . . 184
6.5.1 Limitations . . . 186
6.6 Conclusions . . . 186

7 A review of monitoring-related methodology literature 189
7.1 Introduction . . . 190
7.2 Aims and objectives . . . 191
7.3 Methods . . . 191
7.4 Results . . . 192
7.4.1 Development and evaluation of monitoring strategies . . . 193
7.4.2 Screening . . . 202
7.4.3 Time-dependent ROC curves . . . 203
7.4.4 Differentiating measurement change from measurement variability . . . 206
7.4.5 Health economic approaches . . . 209
7.5 Summary and conclusions . . . 210
7.5.1 Limitations . . . 213
7.5.2 Application to thesis: methods used . . . 213

8 Simulating monitoring data and evaluating monitoring strategies 215
8.1 Introduction . . . 216
8.2 Aims and objectives . . . 218
8.3 Methods . . . 218
8.3.1 Simulation of true disease progression . . . 221
8.3.2 Simulation of observed values . . . 225
8.3.3 Data sources . . . 225
8.3.4 Implementation of a monitoring strategy . . . 229
8.3.5 Evaluation of a monitoring strategy . . . 231
8.4 Results . . . 234
8.4.1 Reference monitoring strategy (strategy A) . . . 234

8.4.2 Comparing strategies with changes to individual components to the reference strategy . . . 236
8.4.3 Sensitivity analyses . . . 239
8.4.4 Comparison to ELUCIDATE data . . . 241
8.5 Discussion . . . 244
8.5.1 Reference strategy . . . 244
8.5.2 Comparing strategies with changes to individual components to the reference strategy . . . 245
8.5.3 Estimates of test performance and disease progression (sensitivity analyses) . . . 247
8.5.4 Comparison of modelled data to ELUCIDATE . . . 248
8.5.5 Limitations . . . 248
8.5.6 Further work . . . 250
8.6 Conclusions . . . 251

9 Discussion and Conclusions 253

9.1 Overview of thesis . . . 254
9.1.1 Research questions . . . 254
9.1.2 What are the current methods for assessment of biological variability? 256
9.1.3 How well are biological variability studies designed, analysed and reported? . . . 257
9.1.4 Can the design and analysis of biological variability studies be improved, specifically sample size planning and outlier detection methods? Are the current methods for analysis of biological variability studies valid, considering sample size planning and outlier detection methods? . . . 259
9.1.5 What are the current methods for the design and analysis of monitoring strategies? . . . 260
9.1.6 Can modelling methods be used to predict the performance of monitoring strategies, to identify optimal strategies to be evaluated in an RCT? . . . 261

9.2 Strengths and Limitations . . . 262
9.2.1 Strengths . . . 262
9.2.2 Limitations . . . 262
9.3 Implications for practice . . . 264
9.4 Future research . . . 265
9.5 Conclusions . . . 267

A Biological variability studies: review of design, analysis, and reporting 269
A.1 Studies identified for review of biological variability studies . . . 269

B Analysis of biological variability 289
B.1 Detailed results of analyses . . . 289

C Sample size guidance and justification for studies of biological variation 293
C.1 Results of the normally distributed data simulation; varying sample size . . . 293
C.2 Results for log-normal data simulation; varying sample size . . . 297
C.3 Results of the normally distributed data simulation; varying variability . . . 306
C.4 Results of the log-normal data simulation; varying variability . . . 310
C.5 Sensitivity analyses . . . 320

D The impact of outlier detection and removal on studies of biological variability 341
D.1 Outlier detection methods with outlier simulation–poor test performance . . . 341
D.2 Outlier detection methods with outlier simulation–increased n . . . 358
D.3 Outlier detection methods with outlier simulation . . . 363

E A review of monitoring-related methodology literature 383
E.1 Summary of identified studies . . . 383

F Simulating monitoring data and evaluating monitoring strategies 391
F.1 Detailed simulation results . . . 391
F.2 Detailed simulation results–sensitivity analyses . . . 396

References 417

List of Figures

1.1 The general monitoring process. 4

1.2 Visualisation of pre-analytical, analytical, within-individual and between-individual variability. 9

1.3 Method for selection of monitoring strategies. 13

1.4 Basic test accuracy measures. 15

1.5 Design of monitoring RCTs. 17

1.6 Impact of monitoring on patients. 20

2.1 Biological variability study design. 30

3.1 Flowchart studies included in biological variabilty review. 63

3.2 Histogram of study sample sizes in identified biological variability studies. . . 68

4.1 eGFR-C biological variability sub-study design. 92

4.2 Histogram of original and log transformed measures. 98

4.3 Beeswarm plot of data and removed data. 106

5.1 Illustration of biological variability data simulation process. 120

5.2 Histogram and plot of simulated biological variability data with normally distributed variability. 121

5.3 Histograms and plots of simulated biological variability data with log-normal distributed variability. 122
5.4 Coverage estimates from biological variabilty data simulations varying sample size. 128
5.5 SD estimates from biological variabilty data simulations varying sample size. 132
5.6 CV estimates from biological variability data simulations varying sample size. 133
5.7 II and RCV estimates from biological variability data simulations varying sample size. 134
5.8 Log-normal biological variability sample size simulation: CV estimates from biological variability data simulations varying sample size. 135
5.9 Coverage estimates from biological variabilty data simulations varying SD_A, SD_I and SD_G. 137
5.10 SD estimates from biological variability data simulations varying test variability. 141
5.11 CV estimates from biological variability data simulations varying test variability. 142
5.12 II and RCV estimates from biological variability data simulations varying test variability. 143

6.1 Histograms showing the number of measurements removed. 160
6.2 Estimates of CVs when using different outlier detection strategies. 165
6.3 Estimates of II and RCV when using different outlier detection strategies. . . 168
6.4 Estimates of asymmetric RCV when using different outlier detection strategies. 169
6.5 Outlier detection methods with outlier simulation (magnitude 2)–analytical CV. 176
6.6 Outlier detection methods with outlier simulation (magnitude 10)–analytical CV. 177
6.7 Outlier detection methods with outlier simulation (magnitude 2)–within-individual CV. 178
6.8 Outlier detection methods with outlier simulation (magnitude 10)–within-individual CV. 179
6.9 Outlier detection methods with outlier simulation (magnitude 2)–between-individual CV. 180

6.10 Outlier detection methods with outlier simulation (magnitude 10)–between-individual CV. 181

6.11 Comparison of average percentage bias for analytical, within-individual and between-individual standard deviations between outlier detection methods. . 183

7.1 Process of screening patients. 203

8.1 The design of the ELUCIDATE study. 219

8.2 Illustration of the monitoring data simulation process. 223

8.3 Observed ELF measures from monitoring data simulation. 227

8.4 Implementing a monitoring strategy using simulated data. 232

8.5 Performance of various monitoring strategies on simulated monitoring data with PPV of 25%. 236

9.1 Pathway of designing monitoring studies. 266

C.1 Log-normal biological variability sample size simulation: SD estimates from biological variability data simulations varying sample size. 297

C.2 Log-normal biological variability sample size simulation: II and RCV estimates from biological variability data simulations varying sample size. 298

C.3 Log-normal biological variability sample size simulation: asymmetric RCV estimates from biological variability data simulations varying sample size. . . 299

C.4 Log-normal biological variability sample size simulation: SD estimates from biological variability data simulations varying test variability. 310

C.5 Log-normal biological variability sample size simulation: CV estimates from biological variability data simulations varying test variability. 311

C.6 Log-normal biological variability sample size simulation: II and RCV estimates from biological variability data simulations varying test variability. 312

C.7 Log-normal biological variability sample size simulation: asymmetric RCV estimates from biological variability data simulations varying test variability. 313

F.1 Adjusted fibrosis progression rate–performance of various monitoring strategies on simulated monitoring data with PPV of 25%. A is the simple threshold strategy; B is the retest strategy; C is the decreased monitoring frequency strategy; D is the absolute increase from initial value strategy; E is the absolute increase from last value strategy; F is the relative increase from initial value strategy; G is the relative increase from last value strategy; H is the linear regression strategy. 396

List of Tables

1.1 Monitoring and variability terminology. 3

2.1 ANOVA table. 36
2.2 Commonly reported variability measures. 44
2.3 Examples of variability: measures for normally distributed data. 45
2.4 Examples of variability: measures for log-normally distributed data. 45
2.5 Biological variability study checklist. 51

3.1 Identified biological variability studies by search. 64
3.2 Aims of identified biological variability studies. 66
3.3 Populations studied in identified biological variability studies. 67
3.4 Sample size and study duration in identified biological variability studies. . . 70
3.5 Variability levels assessed in identified biological variability studies. 73
3.6 Analysis methods of studies in identified biological variability studies. 77
3.7 Reporting of identified biological variability studies. 79

4.1 Characteristics of patients recruited to eGFR-C biological variability sub-study. 95
4.2 Summaries and normality testing for non-transformed and natural log transformed Iohexol, Creatinine and Cystatin C data. 97

4.3 Analysis of the eGFR-C biological variability study–outlier detection using the Fraser-Harris method. 99
4.4 Analysis of the eGFR-C biological variability study–results using the Fraser-Harris method. 100
4.5 Analysis of the eGFR-C biological variability study–results of Iohexol analyses. 102
4.6 Analysis of the eGFR-C biological variability study–results of Creatinine analyses. 103
4.7 Analysis of the eGFR-C biological variability study–results of Cystatin C analyses. 104

5.1 Notation description for biological variability study sample size simulation method. 118
5.2 Performance measures to assess biological variability sample size simulation results. 124
5.3 Biological variability study sample size simulation results–bias performance measures varying number of participants, observations and assessments. . . . 129
5.4 Biological variability study sample size simulation results–accuracy and coverage performance measures varying number of participants, observations and assessments. 130
5.5 Biological variability study sample size simulation results–bias performance measures varying CV_A, CV_I and CV_G. 138
5.6 Biological variability study sample size simulation results–accuracy and coverage performance measures varying CV_A, CV_I and CV_G. 139

6.1 Dixon's Q tables. 156
6.2 Outliers removed by each detection method for the 5,000 simulations. 159
6.3 Outlier detection methods with no outlier simulation–bias performance measures. 162
6.4 Outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 162
6.5 Outlier detection methods with no outlier simulation–SDs. 164
6.6 Outlier detection methods with no outlier simulation–CVs. 164

6.7 Outlier detection methods with no outlier simulation–II, RCV and mean. . . 167
6.8 Outlier detection methods with no outlier simulation–asymmetric RCVs. . . . 167
6.9 Outlier detection methods with outlier simulation–outliers removed by each detection method. 172

7.1 Focus of monitoring papers. 193

8.1 Monitoring simulation model notation. 220
8.2 Trial consideration estimates used in monitoring simulation modelling. 224
8.3 Data used in monitoring simulation model. 226
8.4 Monitoring simulation results by observation point for the reference strategy (strategy A). 235
8.5 Monitoring simulation results of various monitoring strategies. 237
8.6 Monitoring simulation results of using the reference strategy when changing estimates required for data simulation. 240
8.7 Comparison of monitoring simulation to trial data–results of analysis of randomisation ELF and analysis of variance for ELF measurements at all time points. 242
8.8 Results of multilevel model of repeated ELF measures from ELUCIDATE trial and monitoring simulation. 243
8.9 Results of multilevel model (with random slope estimated) of repeated ELF measures from ELUCIDATE trial and monitoring simulation. 243
8.10 Comparison of outcomes for trial and simulated data. 244

A.1 Studies identified for review of biological variability studies. 270
A.2 Details of studies identified for review of biological variability studies. 279

B.1 Analysis of the eGFR-C biological variability study–results of iohexol outlier tests. 290
B.2 Analysis of the eGFR-C biological variability study–creatinine outlier tests. . 291
B.3 Analysis of the eGFR-C biological variability study–Cystatin C outlier tests. 292

C.1 SD estimates from biological variability data simulations varying number of participants, observations and assessments. 294

C.2 CV estimates from biological variability data simulations varying number of participants, observations and assessments. 295

C.3 II, RCV and mean estimates from biological variability data simulations varying number of participants, observations and assessments. 296

C.4 Log normal simulation: bias performance measures varying number of participants, observations and assessments. 300

C.5 Log normal simulation: accuracy and coverage performance measures varying number of participants, observations and assessments. 301

C.6 Log normal simulation: SD estimates from biological variability data simulations varying number of participants, observations and assessments. 302

C.7 Log normal simulation: CV estimates from biological variability data simulations varying number of participants, observations and assessments. 303

C.8 Log normal simulation: II, RCV and mean estimates from biological variability data simulations varying number of participants, observations and assessments. 304

C.9 Log normal simulation: asymmetric RCV estimates from biological variability data simulations varying number of participants, observations and assessments. 305

C.10 SD estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 307

C.11 CV estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 308

C.12 II, RCV and mean estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 309

C.13 Log normal simulation: bias performance measures varying CV_A, CV_I and CV_G. 314

C.14 Log normal simulation: accuracy and coverage performance measures varying CV_A, CV_I and CV_G. 315

C.15 Log normal simulation: SD estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 316

C.16 Log normal simulation: CV estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 317

C.17 Log normal simulation: II, RCV and mean estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 318

C.18 Log normal simulation: asymmetric RCV estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 319

C.19 Increased base n_1, n_2 and n_3: bias performance measures varying number of participants, observations and assessments. 321

C.20 Increased base n_1, n_2 and n_3: accuracy and coverage performance measures varying number of participants, observations and assessments. 322

C.21 Increased base n_1, n_2 and n_3: bias performance measures varying CV_A, CV_I and CV_G. 323

C.22 Increased base n_1, n_2 and n_3: accuracy and coverage performance measure varying CV_A, CV_I and CV_G. 324

C.23 Increased base n_1, n_2 and n_3: SD estimates from biological variability data simulations varying number of participants, observations and assessments. . . 325

C.24 Increased base n_1, n_2 and n_3: CV estimates from biological variability data simulations varying number of participants, observations and assessments. . . 326

C.25 Increased base n_1, n_2 and n_3: II, RCV and mean estimates from biological variability data simulations varying number of participants, observations and assessments. 327

C.26 Increased base n_1, n_2 and n_3: SD estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 328

C.27 Increased base n_1, n_2 and n_3: CV estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 329

C.28 Increased base n_1, n_2 and n_3: II, RCV and mean estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 330

C.29 Increased base σ_A, σ_I and σ_G: bias performance measures varying number of participants, observations and assessments. 331

C.30 Increased base σ_A, σ_I and σ_G: accuracy and coverage performance measures varying number of participants, observations and assessments. 332

C.31 Increased base σ_A, σ_I and σ_G: bias performance measures varying CV_A, CV_I and CV_G. 333

C.32 Increased base σ_A, σ_I and σ_G: accuracy and coverage performance measures varying CV_A, CV_I and CV_G. 334

C.33 Increased base σ_A, σ_I and σ_G: SD estimates from biological variability data simulations varying number of participants, observations and assessments. . . 335

C.34 Increased base σ_A, σ_I and σ_G: CV estimates from biological variability data simulations varying number of participants, observations and assessments. . . 336

C.35 Increased base σ_A, σ_I and σ_G: II, RCV and mean estimates from biological variability data simulations varying number of participants, observations and assessments. 337

C.36 Increased base σ_A, σ_I and σ_G: SD estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 338

C.37 Increased base σ_A, σ_I and σ_G: CV estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 339

C.38 Increased base σ_A, σ_I and σ_G: II, RCV and mean estimates from biological variability data simulations varying σ_A, σ_I and σ_G. 340

D.1 Increased CV_A: outlier detection methods with no outlier simulation–outliers removed by each detection method. 342

D.2 Increased CV_A: outlier detection methods with no outlier simulation–bias performance measures. 343

D.3 Increased CV_A: outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 343

D.4 Increased CV_A: outlier detection methods with no outlier simulation–SD. . . 344

D.5 Increased CV_A: outlier detection methods with no outlier simulation–CV. . . 344

D.6 Increased CV_A: outlier detection methods with no outlier simulation–II, RCV and mean. 345

D.7 Increased CV_A: outlier detection methods with no outlier simulation–asymmetric RCVs. 345
D.8 Increased CV_I: outlier detection methods with no outlier simulation–outliers removed by each detection method. 346
D.9 Increased CV_I: outlier detection methods with no outlier simulation–bias performance measures. 347
D.10 Increased CV_I: outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 347
D.11 Increased CV_I: outlier detection methods with no outlier simulation–SD. . . . 348
D.12 Increased CV_I: outlier detection methods with no outlier simulation–CV. . . 348
D.13 Increased CV_I: outlier detection methods with no outlier simulation–II, RCV and mean. 349
D.14 Increased CV_I: outlier detection methods with no outlier simulation–asymmetric RCVs. 349
D.15 Increased CV_G: outlier detection methods with no outlier simulation–outliers removed by each detection method. 350
D.16 Increased CV_G: outlier detection methods with no outlier simulation–bias performance measures. 351
D.17 Increased CV_G: outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 351
D.18 Increased CV_G: outlier detection methods with no outlier simulation–SD. . . 352
D.19 Increased CV_G: outlier detection methods with no outlier simulation–CV. . . 352
D.20 Increased CV_G: outlier detection methods with no outlier simulation–II, RCV and mean. 353
D.21 Increased CV_G: outlier detection methods with no outlier simulation–asymmetric RCVs. 353
D.22 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–outliers removed by each detection method. 354
D.23 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–bias performance measures. 355

D.24 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 355
D.25 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–SD. 356
D.26 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–CV. 356
D.27 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–II, RCV and mean. 357
D.28 Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–asymmetric RCVs. 357
D.29 Increased n: outlier detection methods with no outlier simulation–outliers removed by each detection method. 359
D.30 Increased n: outlier detection methods with no outlier simulation–bias performance measures. 360
D.31 Increased n: outlier detection methods with no outlier simulation–accuracy and coverage performance measures. 360
D.32 Increased n: outlier detection methods with no outlier simulation–SD. 361
D.33 Increased n: outlier detection methods with no outlier simulation–CV. 361
D.34 Increased n: outlier detection methods with no outlier simulation–II, RCV and mean. 362
D.35 Increased n: outlier detection methods without outlier simulation–asymmetric RCVs. 362
D.36 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–bias performance measures. 364
D.37 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–accuracy and coverage performance measures. 364
D.38 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–SD. 365
D.39 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–CV. 365
D.40 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–II, RCV and mean. 366

D.41 Outlier detection methods with outlier simulation (0.5% and magnitude 2)–asymmetric RCVs. 366
D.42 Outlier detection methods with outlier simulation (1% and magnitude 2)–bias performance measures. 367
D.43 Outlier detection methods with outlier simulation (1% and magnitude 2)–accuracy and coverage performance measures. 367
D.44 Outlier detection methods with outlier simulation (1% and magnitude 2)–SD. 368
D.45 Outlier detection methods with outlier simulation (1% and magnitude 2)–CV. 368
D.46 Outlier detection methods with outlier simulation (1% and magnitude 2)–II, RCV and mean. 369
D.47 Outlier detection methods with outlier simulation (1% and magnitude 2)–asymmetric RCVs. 369
D.48 Outlier detection methods with outlier simulation (2% and magnitude 2)–bias performance measures. 370
D.49 Outlier detection methods with outlier simulation (2% and magnitude 2)–accuracy and coverage performance measures. 370
D.50 Outlier detection methods with outlier simulation (2% and magnitude 2)–SD. 371
D.51 Outlier detection methods with outlier simulation (2% and magnitude 2)–CV. 371
D.52 Outlier detection methods with outlier simulation (2% and magnitude 2)–II, RCV and mean. 372
D.53 Outlier detection methods with outlier simulation (2% and magnitude 2)–asymmetric RCVs. 372
D.54 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–bias performance measures. 373
D.55 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–accuracy and coverage performance measures. 373
D.56 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–SD.374
D.57 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–CV.374
D.58 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–II, RCV and mean. 375

D.59 Outlier detection methods with outlier simulation (0.5% and magnitude 10)–asymmetric RCVs. 375
D.60 Outlier detection methods with outlier simulation (1% and magnitude 10)–bias performance measures. 376
D.61 Outlier detection methods with outlier simulation (1% and magnitude 10)–accuracy and coverage performance measures. 376
D.62 Outlier detection methods with outlier simulation (1% and magnitude 10)–SD. 377
D.63 Outlier detection methods with outlier simulation (1% and magnitude 10)–CV. 377
D.64 Outlier detection methods with outlier simulation (1% and magnitude 10)–II, RCV and mean 378
D.65 Outlier detection methods with outlier simulation (1% and magnitude 10)–asymmetric RCVs. 378
D.66 Outlier detection methods with outlier simulation (2% and magnitude 10)–bias performance measures. 379
D.67 Outlier detection methods with outlier simulation (2% and magnitude 10)–accuracy and coverage performance measures. 379
D.68 Outlier detection methods with outlier simulation (2% and magnitude 10)–SD. 380
D.69 Outlier detection methods with outlier simulation (2% and magnitude 2)–CV. 380
D.70 Outlier detection methods with outlier simulation (2% and magnitude 10)–II, RCV and mean. 381
D.71 Outlier detection methods with outlier simulation (2% and magnitude 10)–asymmetric RCVs. 381

E.1 Summary of reviewed monitoring and monitoring related methodology literature.384

F.1 Monitoring simulation–results using retest monitoring strategy (strategy B) by observation point. 392
F.2 Monitoring simulation–results using reduced frequency of monitoring strategy (strategy C) by observation point. 392
F.3 Monitoring simulation–results using absolute increase from start value monitoring strategy (strategy D) by observation point. 393

F.4 Monitoring simulation–results using absolute increase from last value monitoring strategy (strategy E) by observation point. 393
F.5 Monitoring simulation–results using relative increase from start value monitoring strategy (strategy F) by observation point. 394
F.6 Monitoring simulation–results using relative increase from last value monitoring strategy (strategy G) by observation point. 394
F.7 Monitoring simulation–results using linear regression monitoring strategy (strategy H) by observation point. 395
F.8 Monitoring simulation–results using reference strategy with decreased measurement error by observation point. 397
F.9 Monitoring simulation–results using reference strategy with decreased measurement error by observation point and PPV at 25%. 398
F.10 Monitoring simulation–results using reference strategy with increased measurement error by observation point. 398
F.11 Monitoring simulation–results using reference strategy with increased measurement error by observation point and PPV at 25%. 399
F.12 Monitoring simulation–results using reference strategy with decreased between-individual variability by observation point. 399
F.13 Monitoring simulation–results using reference strategy with decreased between-individual variability by observation point and PPV at 25%. 400
F.14 Monitoring simulation–results using reference strategy with increased between-individual variability by observation point. 400
F.15 Monitoring simulation–results using reference strategy with increased between-individual variability by observation point and PPV at 25%. 401
F.16 Monitoring simulation–results using reference strategy with decreased fibrosis progression rate by observation point. 401
F.17 Monitoring simulation–results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%. 402
F.18 Monitoring simulation–results using reference strategy with increased fibrosis progression rate by observation point. 402

F.19 Monitoring simulation–results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%. 403

F.20 Monitoring simulation–results of strategies A-H for adjusted fibrosis progression estimate data. 404

F.21 Monitoring simulation–results by observation point for the reference strategy (strategy A) for adjusted fibrosis progression estimate data. 405

F.22 Monitoring simulation–results using retest monitoring strategy (strategy B) by observation point for adjusted fibrosis progression estimate data. 405

F.23 Monitoring simulation–results using reduced frequency of monitoring strategy (strategy C) by observation point for adjusted fibrosis progression estimate data. 406

F.24 Monitoring simulation–results using absolute increase from start value monitoring strategy (strategy D) by observation point for adjusted fibrosis progression estimate data. 406

F.25 Monitoring simulation–results using absolute increase from last value monitoring strategy (strategy E) by observation point for adjusted fibrosis progression estimate data. 407

F.26 Monitoring simulation–results using relative increase from start value monitoring strategy (strategy F) by observation point for adjusted fibrosis progression estimate data. 407

F.27 Monitoring simulation–results using relative increase from last value monitoring strategy (strategy G) by observation point for adjusted fibrosis progression estimate data. 408

F.28 Monitoring simulation–results using linear regression monitoring strategy (strategy H) by observation point for adjusted fibrosis progression estimate data. . 408

F.29 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results of using the reference strategy when changing estimates required for data simulation. 409

F.30 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased measurement error by observation point. 410

F.31 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased measurement error by observation point and PPV at 25%. 410

F.32 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased measurement error by observation point.411

F.33 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased measurement error by observation point and PPV at 25%. 411

F.34 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased between-individual variability by observation point. 412

F.35 Monitoring simulation adjusted fibrosis progression sensitivity analysesresults using reference strategy with decreased between-individual variability by observation point and PPV at 25%. 412

F.36 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased between-individual variability by observation point. 413

F.37 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased between-individual variability by observation point and PPV at 25%. 413

F.38 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased fibrosis progression rate by observation point. 414

F.39 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%. 414

F.40 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased fibrosis progression rate by observation point. 415

F.41 Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%. 415

List of Boxes

2.1 The Shapiro-Wilk test. 35

3.1 Multiple testing situations in identified biological variability studies. 65

3.2 Biological variability studies with large sample sizes–utilising existing data collection. 69

3.3 Biological variability studies: Analysis where CV_A is estimated separately to main variability study. 72

3.4 Biological variability studies: symmetric and non-symmetric RCVs. 80

3.5 Biological variability studies: reporting exemplars. 80

4.1 Fraser Harris framework. 93

7.1 Signal and noise monitoring examples. 195

7.2 Joint latent class models example. 196

7.3 Bayesian hierarchical change-point models examples. 197

7.4 Non-linear models examples. 198

7.5 Alternative modelling approaches examples. 199

7.6 Simulation studies examples. 200

7.7 Pooled analyses examples. 201

7.8 Estimating the duration of the pre-clinical stage of disease. 204
7.9 Optimal screening frequency. 205
7.10 Optimal decision rules for novel tests. 206
7.11 Sensitivity and specificity. 207
7.12 Modelling to produce time-dependent ROC curves. 208
7.13 Methods accounting for test variability. 209
7.14 Statistical process control and statistical rules for interpretation of sequential tests. 210
7.15 Issues for methods using test variability. 211
7.16 Decision analytic models examples. 211
7.17 Real options approaches examples. 212

Dissemination of research

The work presented in this thesis has, in part, been disseminated by publication and presentations at conferences.

Some outputs from this thesis are direct outputs with the author conducting all research, and writing with guidance from supervisors and other collaborators. In addition, there are related outputs from this thesis where the author was part of a collaborative effort in delivering.

Direct outputs

Publications

Selby PJ, Banks RE, Gregory W, Hewison J, Rosenberg W, Altman DG, Deeks JJ, McCabe C, Parkes J, Sturgeon C, Thompson D, Twiddy M, Bestall J, Bedlington J, Hale T, Dinnes J, Jones M, Lewington A, Messenger MP, Napp V, Sitch A, Tanwar S, Vasudev NS, Baxter P, Bell S, Cairns DA, Calder N, Corrigan N, Del Galdo F, Heudtlass P, Hornigold N, Hulme C, Hutchinson M, Lippiatt C, Livingstone T, Longo R, Potton M, Roberts S, Sim S, Trainor S, Welberry Smith M, Neuberger J, Thorburn D, Richardson P, Christie J, Sheerin N, McKane W, Gibbs P, Edwards A, Soomro N, Adeyoju A, Stewart GD, Hrouda D. Methods

for the evaluation of biomarkers in patients with kidney and liver diseases: multicentre research programme including ELUCIDATE RCT. Chapter 5: A review of monitoring-related methodology literature. Programme Grants for Applied Research 2018.

Selby PJ, Banks RE, Gregory W, Hewison J, Rosenberg W, Altman DG, Deeks JJ, McCabe C, Parkes J, Sturgeon C, Thompson D, Twiddy M, Bestall J, Bedlington J, Hale T, Dinnes J, Jones M, Lewington A, Messenger MP, Napp V, Sitch A, Tanwar S, Vasudev NS, Baxter P, Bell S, Cairns DA, Calder N, Corrigan N, Del Galdo F, Heudtlass P, Hornigold N, Hulme C, Hutchinson M, Lippiatt C, Livingstone T, Longo R, Potton M, Roberts S, Sim S, Trainor S, Welberry Smith M, Neuberger J, Thorburn D, Richardson P, Christie J, Sheerin N, McKane W, Gibbs P, Edwards A, Soomro N, Adeyoju A, Stewart GD, Hrouda D. Methods for the evaluation of biomarkers in patients with kidney and liver diseases: multicentre research programme including ELUCIDATE RCT. Chapter 7: Simulating monitoring data and evaluating monitoring strategies. Programme Grants for Applied Research 2018.

Rowe C, Sitch A, Barratt J, Brettell E, Cockwell P, Dalton N, Deeks J, Eaglestone G, Pellatt-Higgins T, Kalra P, Khunti K, Loud F, Morris F, Ottridge R, Stevens P, Sharpe C, Sutton A, Taal M, Lamb E. Biological variation of measured and estimated glomerular filtration rate (GFR) in patients with chronic kidney disease: the eGFR-C Study. Kidney International (in press).

Oral conference presentations

Sitch A, Dinnes J, Parkes J, Gregory W, Hewison J, Altman D, Deeks J. Simulation modelling to identify optimal monitoring strategies: the use of the ELF biomarker in liver disease monitoring. 2nd Clinical Trials Methodology Conference: Methodology Matters, Edinburgh, UK. 18-19 November 2013.

Sitch A, Mallett S, Deeks J. Biological variability studies: design, analysis and reporting. 4th Methods for Evaluating Medical Tests and Biomarkers (MEMTAB) Symposium, Birmingham, UK. 19-20 July 2016.

Sitch A, Mallett S, Deeks J. Sample size guidance and justification for studies of biological variation. EuroMedLab–22nd IFCC-EFLM European Congress of Clinical Chemistry and Laboratory Medicine, Athens, Greece. 11-15 June 2017.

Sitch A, Mallett S, Deeks J. The impact of outlier detection and removal on studies of biological variability (BV). Methods for Evaluation of medical prediction Models, Tests And Biomarkers (MEMTAB), Utrecht, Netherlands. 2-3 July 2018.

Related outputs

Publications

Selby PJ, Banks RE, Gregory W, Hewison J, Rosenberg W, Altman DG, Deeks JJ, McCabe C, Parkes J, Sturgeon C, Thompson D, Twiddy M, Bestall J, Bedlington J, Hale T, Dinnes J, Jones M, Lewington A, Messenger MP, Napp V, Sitch A, Tanwar S, Vasudev NS, Baxter P, Bell S, Cairns DA, Calder N, Corrigan N, Del Galdo F, Heudtlass P, Hornigold N, Hulme C, Hutchinson M, Lippiatt C, Livingstone T, Longo R, Potton M, Roberts S, Sim S, Trainor S, Welberry Smith M, Neuberger J, Thorburn D, Richardson P, Christie J, Sheerin N, McKane W, Gibbs P, Edwards A, Soomro N, Adeyoju A, Stewart GD, Hrouda D. Methods for the evaluation of biomarkers in patients with kidney and liver diseases: multicentre research programme including ELUCIDATE RCT. Chapter 4: Has the randomised controlled trial design been successfully used to evaluate strategies for monitoring disease progression or recurrence? An assessment of experience to date. Programme Grants for Applied Research 2018.

Lamb EJ, Brettell EA, Cockwell P, Dalton N, Deeks JJ, Harris K, Higgins T, Kalra PA, Khunti K, Loud F, Ottridge RS, Sharpe CC, Sitch AJ, Stevens PE, Sutton AJ, Taal MW on belhalf of the eGFR-C study group. The eGFR-C study: accuracy of glomerular filtration rate (GFR) estimation using creatinine and cystatin C and albuminuria for monitoring disease progression in patients with stage 3 chronic kidney disease–prospective longitudinal study in a multiethnic population. BMC Nephrol 2014.

Chapter 1

Introduction

1.1 Evidence based monitoring

Does monitoring of patients improve patient care, and subsequently patient outcomes? What evidence do we have to suggest that routine monitoring is beneficial to patients? With increased knowledge, could we improve how we monitor patients?

Patient monitoring is often performed, at great cost, when the benefit of monitoring has not been evaluated or evidence for monitoring is weak.[1,2] When monitoring is formally assessed, strategies should be evidence-based;[3] however, strategies selected for assessment are not always developed using existing evidence.[4] Monitoring strategies are complex interventions and should be developed and evaluated accordingly.[5]

The aim of this thesis is to investigate optimal monitoring of progressive and recurrent disease. This thesis focusses on two main areas: the design, analysis and reporting of biological variability studies, which provide estimates of measurement error, and modelling to develop

optimal monitoring strategies for further investigation, combing available evidence.

1.1.1 Definition of monitoring

Monitoring of the health of patients is defined here as: *'scheduled repeated testing, where pre-defined test results prompt a change in patient management'.*[6]

When monitoring patients to identify disease progression or recurrence, to begin monitoring patients must have previous disease that could potentially recur or early stage disease that may progress. Selected patients are monitored using a strategy (test or tests used at a series of monitoring points with a decision rule to declare a test result as positive or negative). A decision rule is used at each monitoring point to identify if the result is positive or negative (this may be assessing an image for indication of disease, comparing a test value to a defined threshold or previous measures etc.). If the result is considered negative, monitoring will continue at given time intervals until a positive result is achieved or a given amount of time has passed. When a patient is considered positive, they are no longer monitored in the same way, meaning a change in patient management (new treatment, more intense monitoring, further testing etc.)[7] see Figure 1.1. For a guide to the terminology used in this thesis see Table 1.1.

1.1.2 Monitoring of disease progression and recurrence

Here, the focus is monitoring to identify progression or recurrence of disease, also referred to as surveillance and watchful waiting. Patients would be managed using a monitoring strategy after treatment for a recurrent condition or in the early stages of progressive disease. A monitoring strategy involves repeated use of a test (or multiple tests) in these patients with a specified rule for the test results that would be considered positive and that would lead to a change in patient management. Examples of monitoring routinely performed in the UK are cystoscopy for detecting recurrence of bladder cancer in patients who have previously had tumours removed, with identification of further tumours resulting in subsequent surgery;[8]

Table 1.1: Monitoring and variability terminology.

Term	Explanation
Monitoring	
Monitoring strategy	A monitoring strategy specifies the monitoring test(s), frequency of monitoring, total duration of monitoring and the decision rule used to identify a test result as positive or negative.
Monitoring test	The test used in a monitoring strategy.
Monitoring frequency	The frequency of repeat testing in a monitoring strategy (not necessarily at regular intervals).
Monitoring duration	The total duration of use of a monitoring strategy, may potentially be until disease progression or recurrence, or death.
Decision rule/monitoring rule	The decision rule specifies which results from the monitoring test would be a positive and a negative result. The decision rule may use only the last value from the monitoring test or may rely on previous values also.
Threshold	The threshold is a set level to measure test results against to identify positive results.
'track-shot' rule	A 'track-shot' rule uses multiple results for individuals and assesses these results together to identify if a result is positive or negative.
'snap-shot' rule	A 'snap-shot' rule uses a generic threshold for all individuals using the last obtained monitoring test result to identify if a result is positive or negative.
Absolute change	Changes in test values for an individual on the absolute scale, for example values of 1 and 2 units show an absolute increase of 1 unit.
Relative change	Changes in test values for an individual on the relative scale, for example values of 1 and 2 units show a relative increase of 100%.
Patient management	The management of a patient, for example monitoring, intensive monitoring, treatment, invasive test to identify need for treatment etc.
Variability	
Biological variability	Biological variability is the variability in test measures between and within individuals whilst in a stable disease state.
Pre-analytical variability	Pre-analytical variability is the variability of a test measure due to differences in the how samples have been obtained, stored and transported (prior to evaluation of the sample).
Analytical variability (imprecision)	Analytical variability is the variability of multiple assessments of a single sample (same participant and observation point).
Within-individual variability	Within-individual variability is the variability between test measures for a single individual over time in a stable disease state.
Homeostatic setting point	True value of the test for an individual.
Measurement error	Measurement error is the variability in a test measure around the true value for an individual, this is analytical and within-individual variability combined.
Between-individual variability	Between-individual variability is the the variability in a test measure between individuals.
Assessor variability	Assessor variability is the variability in assessing the result of a test, this is most often seen in imaging studies.
Inter reader variability	Inter reader variability is the variability between readers in assessing a test result, this is most often seen in imaging studies.
Intra reader variability	Intra reader variability is the variability within readers (repeated reads by the same reader) in assessing a test result, this is most often seen in imaging studies.
Evaluation	
True positive	Positive test result at a point when a patient is diseased.
False positive	Positive test result at a point when a patient is not diseased.
True negative	Negative test result at a point when a patient is not diseased.
False negative	Negative test result at a point when a patient is diseased.
Sensitivity	Proportion of diseased participants that have a positive test result at a test point.
Specificity	Proportion of non diseased participants that have a negative test result at a test point.
Positive predictive value (PPV)	Proportion of test positive participants that have a positive test result.
Negative predictive value (NPV)	Proportion of test negative participants that have a negative test result.
Coefficient of variation (CV)	Standard deviation of variability (at the analytical, within-individual and between-individual levels) expressed as a ratio to the mean value.
Index of individuality (II)	The ratio of analytical and within-individual variability compared to between-individual variability. Indicates how much variability is within measures for an individual compared to a group of people.
Reference change value (RCV)	Change in a measure suggesting true change based on estimates of variability.

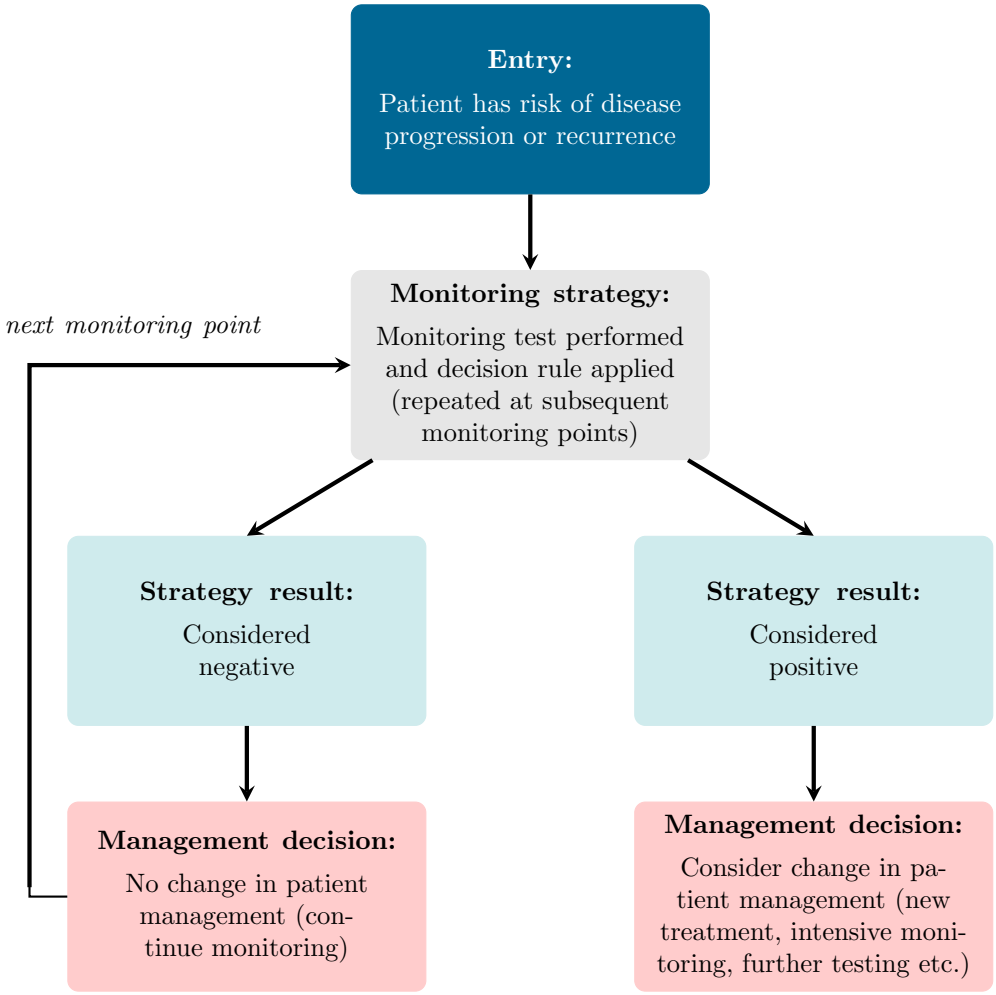


Figure 1.1: The general monitoring process.

and, repeated blood tests to measure CD4+ cell counts in patients with HIV, to indicate progression of disease and allow antiretroviral treatment to be appropriately considered.[9]

There are many disease areas where monitoring is for treatment titration, with the purpose of modifying doses. Whilst methods and theories relating to treatment titration are considered, where appropriate, research in this area of monitoring is not directly applicable to the work presented in this thesis. Examples of diseases where monitoring is for treatment titration are hypertension[10] and diabetes.[11]

Test and biomarkers with continuous measurements rather than binary or categorical are considered in this thesis.

1.1.3 Monitoring data

Stevens and colleagues developed a general model for monitoring data. This model defines Y_{it} as the observed monitoring values and U_{it} as the 'true' underlying ('latent') value, which can never be directly observed.

$$U_{it} = \alpha_i + \beta_{it} \text{ and } Y_{it} = U_{it} + \omega_{it},$$

where α_i is the latent value at time 0, β_{it} is the change in the latent value over time and ω_{it} is measurement error.[12]

1.1.4 Measurement error

The estimation of measurement error is a major theme of this thesis, as measurement error was identified as a key component of monitoring data and vital when considering optimal monitoring strategies. The design of studies and methods used to estimate test variability, particularly biological variability studies, are investigated in this thesis.

1.1.5 Biological variability

Studies of biological variability often estimate the partitioned variances of a test (at the analytical, within-individual and between-individual levels) allowing calculation of measurement error.

Estimates of biological variability provide vital information when using tests for diagnosis and monitoring.[13] Essentially, biological variability estimates quantify the natural fluctuation in test results which is useful when developing a monitoring strategy as true change ('signal') can be detected over random and expected fluctuation ('noise').[14] Biological variability is defined as: *'The natural variability in a laboratory parameter due to physiologic differences among subjects and within the same subject over time.'*[15]

The biological variability of a test can indicate the appropriate and optimal use for monitoring purposes; estimates suggest whether a test is best used with a threshold value for all participants ('snap-shot' rule) or whether differences from previous values for each individual should be considered ('track-shot' rule).[9,16] Results of variability studies will also guide the threshold values used (for the entire population or considering changes from previous values) to define a positive test result, as these results suggest the magnitude of change implying a real change in condition.[16]

1.1.6 Issues in the field

Issues when developing monitoring strategies are:

- Monitoring is performed as standard with many monitoring strategies not subjected to formal evaluation.[2]
- Monitoring strategies (test frequencies, decision rules, thresholds and duration) used to monitor patients with progressive and recurrent disease are rarely evidence based.[2]
- With limited information on variability, as is often the case, test frequencies are commonly based on routine care schedules and test thresholds are chosen arbitrarily.[1]

- When monitoring strategies are evaluated using randomised controlled trials there is limited evidence supporting the components of the strategies.[4]

Issues in studies of biological variability are:

- Estimates of biological variability are necessary for planning monitoring strategies;[3] this includes the design (specifically sample size), conduct, methods for analysis (including outlier detection methods) and validity of analysis for studies estimating variability.

- Sources of estimates of biological variability may be important; current estimates from test manufacturers are designed for the purpose of proving tests meet minimal quality assurance standards and such studies often use spiked or calibrated samples (for example Bargnoux et al[17]) this may not be appropriate when estimating variability to inform monitoring of patients with potential disease progression or recurrence.

- Studies of biological variability may recruit healthy participants[13,18] rather than those with disease, the population of relevance for monitoring.

- For many tests and patient conditions there may be insufficient evidence for estimates of biological variability used to plan monitoring strategies.

This lack of evidence for estimating test variability and designing monitoring strategies is concerning not only due to the high cost of monitoring and studies evaluating monitoring but the multiple opportunities monitoring has to benefit and harm patients. Strategies should be optimised to ensure the greatest possible benefit to patients and evaluated.

1.2 Design and evaluation of biological variability studies

1.2.1 Aims of biological variability studies

The aim of biological variability studies is to quantify the inherent variability of test results.[16]

1.2.2 Variability

When considering a standard laboratory based test there are four types of variability to estimate; these are: pre-analytical, analytical, within-individual and between-individual variability.[16] Studies are often designed to assess analytical, within-individual and between-individual level variability for a biomarker, for example estimated glomerular filtration rate (eGFR).[19] Figure 1.2 shows these measures of variability around the mean value.

1.2.2.1 Pre-analytical variability

'The pre-analytical phase entails all those actions that are necessary in order to obtain diagnostic specimens.'[20] Pre-analytical variability is due to the differences in how subjects have prepared for a sample to be taken and the process of taking the sample or performing the test. The leading causes of pre-analytical variability (as reported in a review by Lippi et al)[20] are: patient preparation (fasting status, exercise and posture), blood drawing (misidentification, insufficient volume, spurious haemolysis, contamination, venous stasis and blood collection devices), sample handling (mixing), sample transportation (time, temperature and integrity) and sample preparation (centrifugation and automation). Often measures are taken to keep pre-analytical variability at a minimum,[16] by keeping testing conditions consistent. The factors influencing pre-analytical variability vary from test to test; with different tests requiring stability of different factors. Pre-analytical variability is not a focus of this thesis, and it is assumed to be minimised.

1.2.2.2 Analytical variability

'The analytical component of variation is derived from replicate analysis of subject samples.'[16] Analytical variability is the variation of results from a single sample and is often assessed by taking a sample and replicating the analysis of this sample.[16] Analytical variability is also known as imprecision and is: *'the closeness of agreement between independent results of measurements obtained under stipulated conditions.'*[21] To assess analytical variability the same

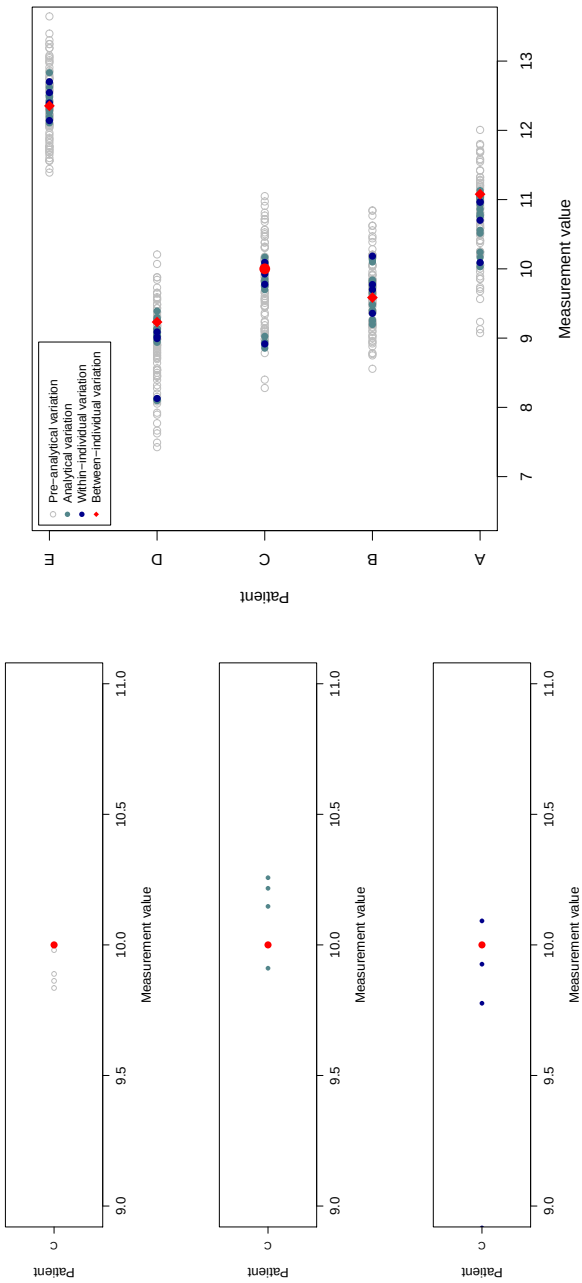


Figure 1.2: Visualisation of pre-analytical, analytical, within-individual and between-individual variability. Left: results for a single participant, separately showing pre-analytical (top), analytical (middle) and within-individual (bottom) variability. Right: results for five participants, showing pre-analytical, analytical, within-individual and between-individual variability. Results show a hypothetical testing situation where the true mean value is 10 units, values at each level of variability follow a normal distribution and the standard deviations at each level are: 0.5 units for pre-analytical, 0.1 units for analytical, 0.5 units for within-individual, and 1 unit for between-individual variability. Image shows values for 5 patients with 4 measurements each, assessed 4 times and each from 4 different testing conditions.

sample is tested using the same testing procedure at the same time (analysis in duplicate) and the difference in results at this level is assessed.[16]

1.2.2.3 Within-individual variability

Within-individual variability (also known as within-subject, within-person and intra-individual variability) is the variability within test measures for an individual occurring due to normal (and expected) variation in participants, for example variability may be detected in repeated measures for an individual if measures are taken at different times of the day and a daily rhythm exists; at different times of the month when a monthly rhythm exists; and, at different seasons when season is a factor.[16] The test values for participants fluctuate for many tests used in patient care without signifying a change in condition, and this is assessed by taking multiple samples from participants over a short time period (when the disease state for participants is expected to stay the same). Assessment of multiple measures for participants allow this level of variability to be quantified.[16]

1.2.2.4 Between-individual variability

Between-individual variability (also known as between-subject, between-person and inter-individual variation)[22] is the variation between the central values for individuals. For some tests individuals will have very similar results and for other tests very different results. Estimates of between-individual and within-individual variability allow us to identify if a test is best used with a threshold to identify a positive or negative result for the population or if the results for each individual should be considered separately with their own individual threshold value.

1.2.2.5 Assessor variability

To evaluate the impact of different assessors on test results intra-inter reader studies are used, often when assessing imaging tests. Individual readers will evaluate a result multiple

times (intra-reader variability) and multiple readers will assess the same result (inter-reader variability).[23] For example, a study used readers to interpret the results of ultrasound scans to measure bladder wall thickness, looking at the variability between and within readers.[24]

Assessor variability is not present when using most laboratory based tests as analytical procedures will provide the healthcare professional with a value of the test result. In the case of assessing an image the process of obtaining a value (for example, measuring tumour size) is variable, with variability occurring between and within assessors. The judgement required in these tests will introduce additional variability.

1.2.2.6 Other sources of variability

Variability in test results can also be due to other factors. Other variability sources can be: intra-inter assay variability,[25] within and between batch variability[26] or intra-inter laboratory variability.[27]

1.3 Designing and evaluating monitoring strategies

1.3.1 Strategy design

Monitoring strategies are complex interventions. Each strategy has several components:

- monitoring test(s);
- monitoring rule–decision rule (for defining test positives);
- monitoring frequency;
- and, monitoring duration.

The review by Selby and colleagues[4] identified less than 50% of monitoring trials assessed gave evidence to support the frequency of monitoring and the intervention patients received

after a positive result used in monitoring strategies; and less than 60% of trials gave evidence to support the threshold used as part of the monitoring strategy.

1.3.1.1 Test(s)

The monitoring test or tests are chosen for their ability to identify the target condition in patients with potential progressive or recurrent disease. The tests used are often chosen based on accuracy estimates, usually for the purpose of diagnosis rather than monitoring or with unclear evidence for the selection of the test.[1,4]

1.3.1.2 Monitoring rule

The monitoring rule (decision rule) is the most complex component of a monitoring strategy. A simple decision rule would use a single threshold for all patients; those with a test result exceeding the threshold would be positive and otherwise negative ('snap-shot' rule).[9] A more complex decision rule may have the threshold based on change from previous results for each patient ('track-shot' rule), meaning an individual threshold for each participant.[9] Decision rules can also be designed to use not only test information but other factors (such as previous results and medical history) with an algorithm to provide a positive or negative result.[28]

1.3.1.3 Monitoring frequency

The frequency of monitoring is how often monitoring tests are performed within a period, for example two tests per year. The monitoring interval is the length of time between each monitoring test. The frequency of monitoring can be the same for each patient for the duration of monitoring or can be more complex. Monitoring may be performed more or less frequently in the early or latter stages of monitoring or depending on previous results or patient history, making the monitoring frequency specific to the individual.[29]

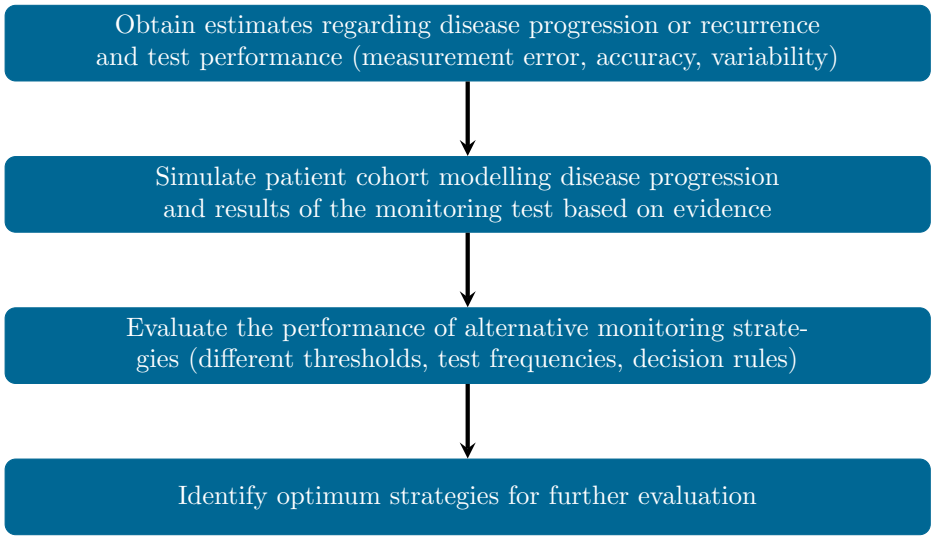


Figure 1.3: Method for selection of monitoring strategies.

1.3.1.4 Monitoring duration

The duration of monitoring is dependent on the clinical situation. Monitoring may continue indefinitely and usually for the duration of a trial, if a trial is assessing a monitoring strategy. However, for some conditions it may be monitoring ceases after a certain period with no positive results, or that if patients appear to be 'low risk' from previous results monitoring can be stopped.[8]

1.3.2 Method for identifying monitoring strategies

The purpose of the work presented in this thesis is to provide guidance, considering the evidence available, for selecting monitoring strategies for further evaluation, see Figure 1.3.

The approach used in this thesis follows from the method introduced by Stevens et al,[12] see §1.1.3. Stevens and colleagues reviewed statistical models used for the control phase of monitoring, including literature based methods, where parameter estimates are obtained from reviewing the literature.

The model is used to simulate monitoring data for a cohort of patients and assess the performance of monitoring strategies in this cohort by comparing the observed data (Y_{it}) to the

latent test value (U_{it}). The model uses the initial value for each individual (α_i), the change over time (β_{it}), and the measurement error (ω_{it}).

To use this approach, available information is collected regarding the progression or recurrence of the disease monitored, and also information on test performance, relating to accuracy and variability. Progression from an initial study starting point is estimated, the modelled latent disease state of patients through time. Using information regarding test performance and variability the observed results seen at monitoring points are estimated.

As both the latent underlying disease status of individuals and the monitoring test results that would have been observed are estimated, the performance of various monitoring strategies can be compared, with those appearing optimal highlighted for full evaluation in a trial, in line with best practice.[5]

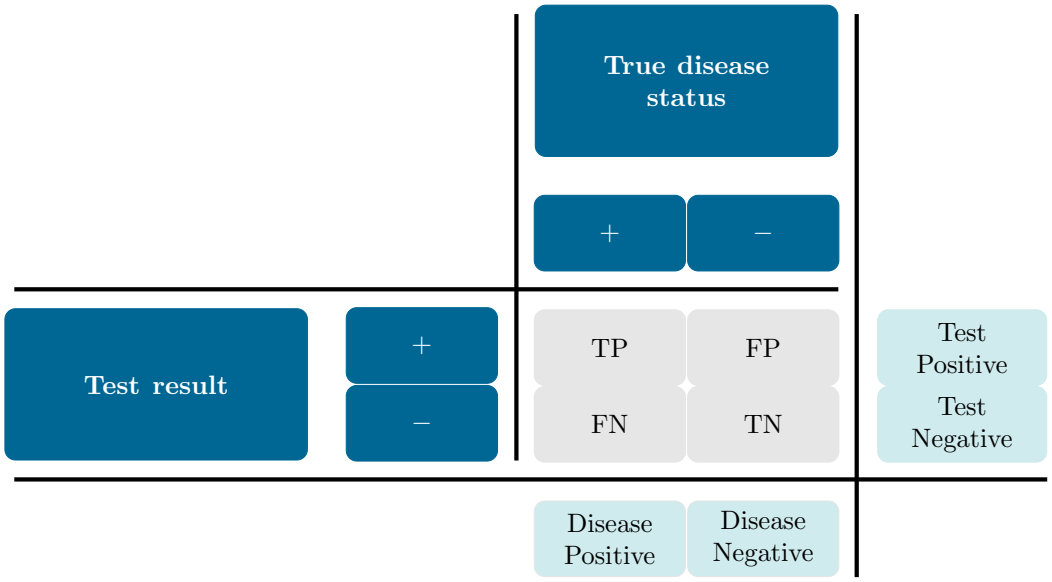
1.3.3 Assessing monitoring strategies

When assessing the ability of monitoring strategies, it may be desirable to assess patient outcomes. However, the purpose of the work in this thesis is to present a method allowing identification of optimal strategies for further evaluation in RCTs of monitoring.

1.3.3.1 Assessing test performance

In studies of diagnostic test accuracy, the test result is compared to the true disease status (measured by a gold standard or reference test, or obtained by follow up of patients). If the test result correctly detects the true disease status of a patient, this result is a true positive (TP) if the test result is positive, or a true negative (TN) if the result is negative. If the patient has a positive test result when the true disease status of the patient is no disease, this is a false positive (FP) result. If a patient received a negative test result when the true disease status of that patient is positive this is a false negative result (see Figure 1.4).

Basic measures of test performance are calculated using the number of patients with each combination of test results and true disease status. Prevalence is the proportion of the



$$\begin{aligned}\text{Prevalence} &= (TP + FN)/(TP + FN + FP + TN)\\ \text{Sensitivity} &= TP/(TP + FN)\\ \text{Specificity} &= TN/(TN + FP)\\ \text{PPV} &= TP/(TP + FP)\\ \text{NPV} &= TN/(TN + FN)\end{aligned}$$

Figure 1.4: Basic test accuracy measures.

population with the disease. Sensitivity is the proportion of patients with disease correctly identified as positive by the test (TP) and specificity is the proportion of patients without disease correctly identified as negative by the test (TN). The positive predictive value (PPV) is the proportion of test positive patients that have the disease; and, the negative predictive value (NPV) is the proportion of test negative patients that do not have the disease.

1.3.3.2 Assessing monitoring test performance

Usual test accuracy measures are more complex in monitoring studies as participants receive multiple negative results and the disease status of participants is not constant (participants develop disease). There are methods for calculating time-dependent sensitivity, specificity and ROC curves, with these methods reviewed by Pepe and colleagues[30] (see Chapter 7).

Adapted from the guidance Li and Gatsonis[31] provide (see Chapter 7 for full details) the outcomes chosen to assess monitoring strategies are: number of tests, positive predictive value

and delay to diagnosis. The number of tests reflects the resource need of a strategy; the positive predictive value shows the percentage of truly positive patients from those identified as test positive; and, the delay to diagnosis reflects the harm to patients by the monitoring strategy as it is the time between developing progressive or recurrent disease and the monitoring strategy detecting disease. To allow these measures to be assessed objectively the positive predictive value (PPV) was fixed at an acceptable level for the clinical question and the number of tests and delay to diagnosis were compared. The threshold used in the decision rule was adjusted to allow the PPV to be at the chosen level. The delay to diagnosis estimate indicates the impact of false negatives and the false positives are controlled by fixing the PPV.

1.3.4 Monitoring studies

Arguably the most robust study design to use when comparing patient health outcomes between monitoring strategies (or comparing a monitoring strategy to routine care) is a randomised controlled trial (RCT).[4] In a typical RCT comparing a monitoring strategy to standard care, patients are recruited and randomised to receive either the monitoring strategy (in addition to routine care) or routine care only. Those receiving monitoring have the monitoring test at specified time intervals and if they have a positive result the management of their condition will change. If monitored patients have a negative test result, they continue monitoring and are tested again at the next testing point.[4] Patients in both randomisation arms receive standard care and may have a change in management, see Figure 1.5.

RCTs evaluating monitoring may compare standard care with monitoring to standard care alone; alternatively studies can compare two (or more) monitoring strategies (differing by test used, test threshold or decision rule employed, frequency of monitoring etc.) or comparing monitoring to immediate treatment.[4]

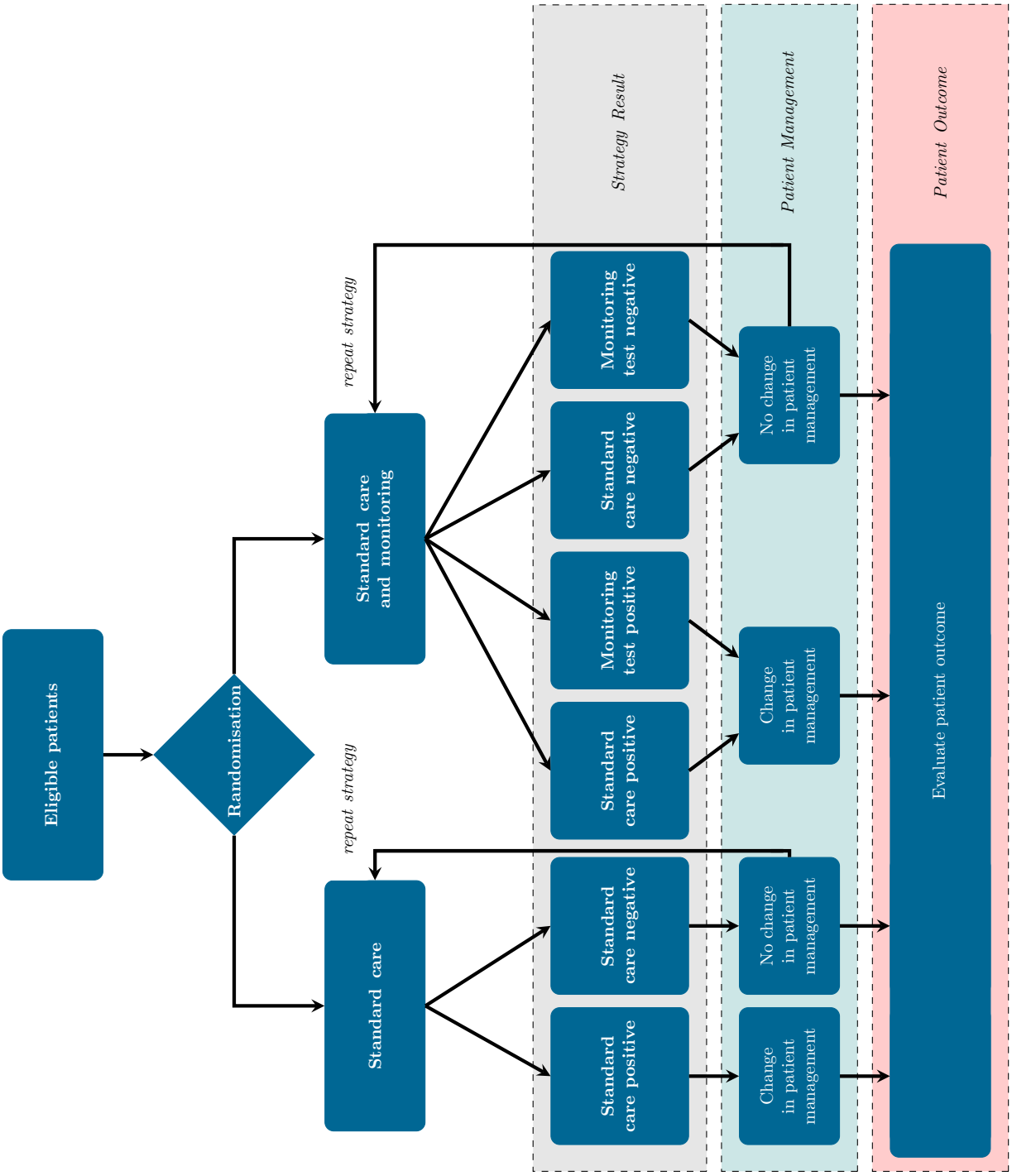


Figure 1.5: Design of monitoring RCTs.

1.3.5 Outcomes

As with trials of treatments, ideally trials of monitoring tests would evaluate differences in patient outcomes across the arms. The review by Selby et al,[4] showed for 49 trials reported in 58 publications, patient outcomes were used in approximately half the trials assessed and were generally patient mortality or the event of new or recurrent disease. Monitoring RCTs may instead evaluate the difference in a process type outcome, such as the percentage of patients having the condition detected. Process outcomes may be used rather than patient outcomes as patient outcomes occur after a lengthy period of follow up, rarity of events or other reasons. Trialists should ensure outcome assessment is not biased across randomisation arms, such as evaluating positives detected using the monitoring test in one arm but an alternative method in the other.[4] The review by Selby and colleagues[4] showed 17% of trials had biased outcome assessment.

1.3.6 Impact on patients

Monitoring strategies are complex interventions, with the monitoring decision informing patient management. Whilst strategies can be evaluated by process measures, such as the number of patients with identified disease, patient benefit is the true outcome of interest which can only be measured with patient outcomes. From a recent review of monitoring RCTs,[4] in most trials the addition of monitoring was thought to lead to earlier treatment or better selection of patients for treatment.

There are many ways tests can change the management of patients providing benefit and/or harm. Ferrante di Ruffano et al[32] identified ways the care pathway could be changed when introducing diagnostic tests, and this has been modified for monitoring tests by Selby et al.[4] The items reported were: test delivery (test feasibility, procedure and frequency), test result (interpretability, clinical validity, timing of test result, detection of long-term change), management decision (added clinical value, timeframe of management decision, clinical confidence) and treatment implementation (timing of treatment, efficacy and adherence). Due to the complex relationship between monitoring and patient impact, full scale investigation

in correctly powered trials with patient outcomes is optimal; and it is essential trials use the evidence available to ensure the monitoring strategies are most beneficial.

Work by Selby and colleagues[7] (adapted from Adriaensen et al[33]) shows the positive and negative consequences of each type of test result on patients when monitoring a progressive or recurrent condition. The positive consequences of monitoring are experienced by those patients with true positive or true negative results; patients with true positive results benefit from earlier treatment and patients with true negative results correctly avoiding treatment or further testing. The patients with false positive and false negative results have undesirable outcomes; patients with false negative results do not receive necessary treatment and have the false confidence of a negative result, and patients with false positive results face the harm of unnecessary treatment, further testing and overdiagnosis. All participants may benefit from monitoring as their exposure to more invasive testing may be reduced; however, on the contrary, monitoring tests may cause harm. There may also be psychological benefits and harm from monitoring, see Figure 1.6.[7,33]

1.3.7 Relationship to screening

Monitoring of patients with progressive or recurrent disease is different to screening of the asymptomatic population, as the disease prevalence in populations receiving testing in the monitoring and screening situations is very different, and the participants without disease will be different to healthy participants taking part in a screening programme. When using tests for monitoring or screening purposes the specific use of such tests should be tailored with the test variability information obtained using the appropriate population, and tests, test frequencies, and decision rules specifically chosen.

There are similarities between monitoring and screening processes. With the methods for screening being further developed, there are potential methods that can be adapted for use in monitoring as appropriate.

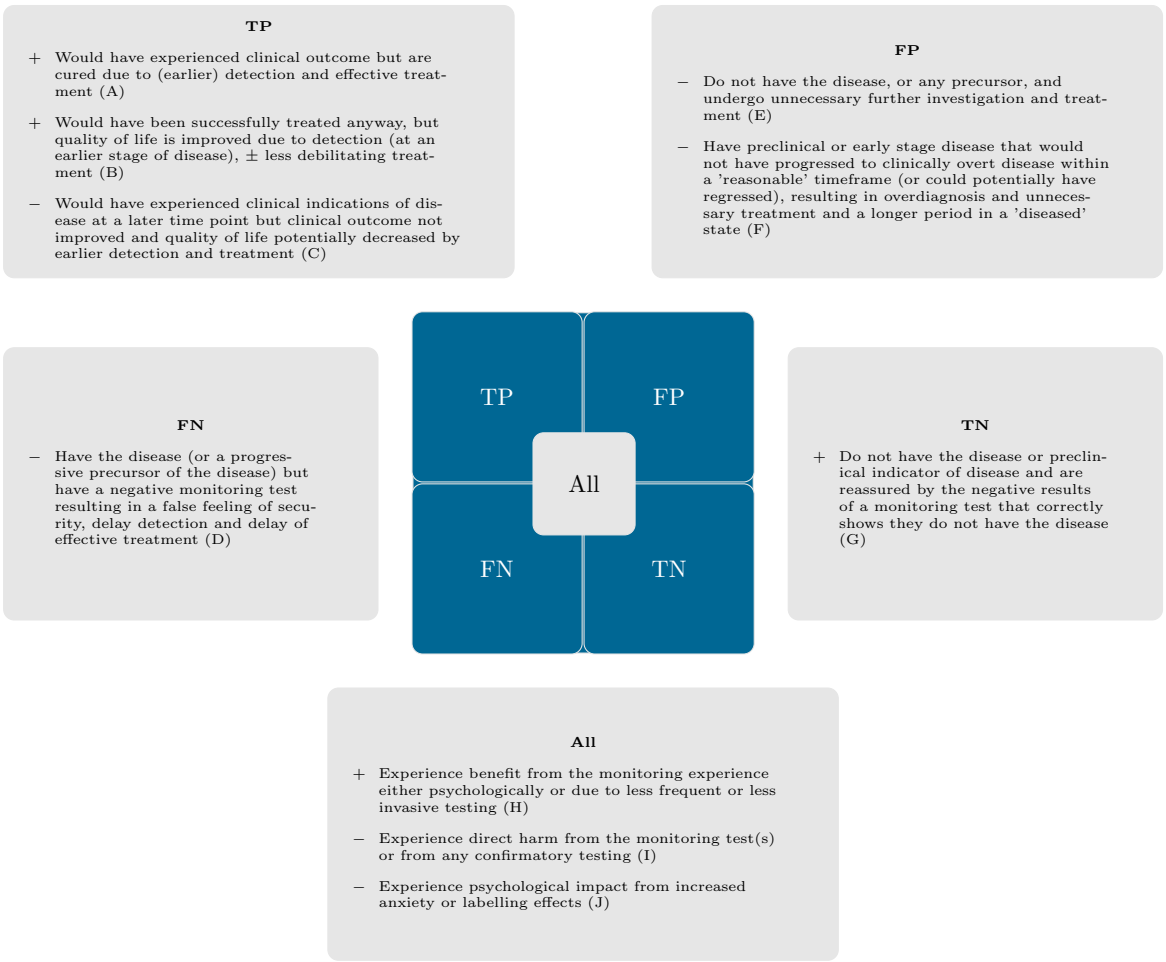


Figure 1.6: Impact of monitoring on patients (from Selby et al,[7] adapted from Adriaensen and colleagues)[33].

1.4 Research questions and thesis outline

1.4.1 Research questions

How can optimal monitoring strategies be designed? What are the appropriate study designs and methods for estimating variability of tests? In order to deliver on these main questions, this thesis aims to answer the following questions:

- What are the current methods for assessing biological variability?
- How well are biological variability studies designed, analysed and reported?
- Can the design and analysis of biological variability studies be improved, specifically sample size planning and outlier detection methods? Are the current methods for analysis of biological variability studies valid, considering sample size and outlier detection?
- What are the current methods for the design and analysis of monitoring strategies?
- Can modelling methods be used to predict the performance of monitoring strategies, to identify optimal strategies to be evaluated in an RCT?

1.4.2 Thesis outline

The aim of this thesis is to investigate optimal monitoring of progressive and recurrent disease. To achieve this aim the thesis looks broadly at two areas: the design, analysis and reporting of biological variability studies, which provide estimates of measurement error, and the use of modelling techniques to combine evidence and allow comparison of monitoring strategies, so that optimal strategies can be used in further investigations.

1.4.2.1 Biological variability studies

Chapters 2 to 6 investigate biological variability studies by: reviewing the current methods (Chapter 2), reviewing the current state of the field in terms of design, analysis and reporting

(Chapter 3), providing critical evaluation and recommendations for these studies through analysis of a case study (Chapter 4), providing guidance for sample size justification (Chapter 5), and evaluating the impact of outlier detection methods and subsequent removal of data (Chapter 6). The validity of current methods is also evaluated considering different sample sizes (Chapter 5) and outlier detection methods (Chapter 6).

Chapter 2 provides detailed explanation of the methods reviewed and used in Chapters 3 to 6. The purpose of this Chapter is to introduce concepts and terminology regarding the design, analysis and reporting of biological variability studies.

Chapter 3 investigates biological variability studies by reviewing the literature across a range of tests and test situations (laboratory, physiological and imaging tests). The purpose of this review was to assess the design, analysis and reporting of biological variability studies. Evaluation of studies identified in this review identified common weaknesses of studies, and recommendations for reporting are proposed at the end of this Chapter.

Chapter 4 provides a detailed analysis of a biological variability study using a case study. The aim of this chapter is to identify how the standard methods work using an example and investigate the use of log-transformation and outlier detection methods.

Chapter 5 uses simulation to validate the confidence intervals for coefficients of variation (see Chapter 2) provided by Burdick and Graybill[34] and the variability of estimates from standard analysis methods using different sample sizes. A tool was developed allowing researchers to plan the sample size (number of participants, observations and assessments) based on the precision of key estimates from biological variability studies. This Chapter provides guidance for sample size when planning a biological variability study.

Chapter 6 uses simulation to investigate the results from biological variability studies when (commonly used) outlier detection methods are used, with simulated data both including and excluding outliers. The purpose of this Chapter is to understand how methods for outlier detection impact the estimates of variability and the validity of these estimates. This Chapter ends with recommendations for the use of outlier detection methods and interpretation of studies using these methods.

1.4.2.2 Designing monitoring strategies

Chapters 7 and 8 identify optimal evidence-based monitoring strategies for further assessment explaining the methods and appropriate data required. Methods for monitoring strategy design and related areas were reviewed (Chapter 7) and using the knowledge of test variability, along with disease progression and test performance a model was developed allowing the comparison of monitoring strategies (Chapter 8) in terms of test performance.

Chapter 7 provides an overview of the literature, with searches performed to identify the current methods for the design of monitoring strategies and also methods from similar fields (for example screening and biomarker development). The purpose of this review was to identify appropriate methods for modelling monitoring strategies for progressive and recurrent disease discussed.

Chapter 8 uses the methods identified in the review of monitoring and monitoring related areas, to develop a model enabling the simulation of monitoring data. The purpose of this model is to estimate the performance of strategies in terms of test performance (allowing candidate strategies to be identified). The aim of this Chapter is to identify the data required to develop a model to evaluate monitoring strategies, and assess and validate the model itself.

Chapter 9 summarises the findings and concludes the thesis. A pathway of studies and evidence required to develop a monitoring strategy is introduced and the implications for practice discussed. This Chapter also suggests future work and discusses the limitations and strengths of the thesis.

Chapter 2

Introduction to assessment of biological variability: design, analysis and reporting

Summary

Biological variability studies assess the between and within-individual variability of test data. A review of biological variability studies was conducted and key methods were reported.

There are many alternative measures of variability with different terms favoured in each area of research. Biological variability studies in laboratory medicine have a framework for design and evaluation proposed by Fraser and Harris in 1989.[35] Key papers focus on the analysis of biological variability studies (usually ANOVA or equivalent methods) with little discussion of the design of studies.

2.1 Introduction

To assess biological variation of a test the variability within and between individuals needs to be estimated after accounting for analytical variability,[16] see Chapter 1. Estimates of variability help place tests efficiently and optimally in the care pathway and indicate the appropriate difference in test results required to change patient management.[13] Knowledge of variability allows the test results triggering a change in patient management to be reflective of a change in the disease state of a patient rather than merely reflecting the usual variability in multiple test measures.[14]

The levels of variability considered are: pre-analytical variability, analytical variability, within-individual variability and between individual variability.

Pre-analytical variability is variability in a test due to the differences in how subjects have prepared for a sample to be taken (for example, diet prior to sample being taken and time of day) and the process of taking the sample or performing the test (for example, the equipment used to perform a test and the person performing the test) and how this sample has been treated prior to assessment (for example, in the case of laboratory based tests this may be procedures for storing and transporting samples). Pre-analytical variability is usually minimised where possible.[16]

Analytical variability is the variability of results from a single sample and is often assessed by taking a sample and replicating the analysis of this sample.[16] Analytical variability for laboratory based tests is generally expressed by how well a laboratory process can replicate results when using identical samples, but for other testing settings, such as imaging testing, this process is different with variability in the assessment of an image being due to the clinician tasked with reviewing the image. The analytical variation in imaging tests is often assessed by the use of inter and intra reader studies. For physiological tests it is often not possible to assess analytical variability.

Within-individual variability is the variability due to fluctuation in test results for an individual with the underlying disease status of the individual remaining consistent for the time

period assessed. Between-individual variability is the variability in test measures between individuals in a population.[16]

Biological variation is a complex component of test evaluation and it is especially vital this is accurately estimated when considering the repeated use of a test in a population to detect disease progression or recurrence.[13] Assessment of biological variability, prior to devising a strategy of monitoring by repeated testing of individuals over time, is crucial to identify optimal use of the test in the strategy (knowing if the test can be used with a constant threshold for the whole population or if changes for each individual are more meaningful) and to fully understand the impact of changes in test results for the multiple tests evaluated for each participant by knowing the likely variability in results.[13]

To understand the scope of studies evaluating biological variability, the design, methods for analysis, reporting and overall quality a review of studies of biological variation was conducted (see Chapter 3), with the key methodological papers influencing this work identified and summarised in this Chapter.

In many situations test data are not normally distributed and are log-transformed, the methods and results for this special case are considered in this Chapter.

2.2 Methods

A review of the design, analysis and reporting of biological variability studies was conducted. To ensure the review covered a wide range of test areas there were many elements to the search: searches for key terms were performed, hand searching through relevant journals, identification of papers from a database of biological variability studies and searches specific to three selected clinical areas. For full details of the search see Chapter 3. The searches were developed to ensure that laboratory, physiological and imaging tests were identified in the review.

Papers were included in the review if they reported a study where the purpose (primary or secondary) was to assess the variability of measuring or evaluating test results in participants

thought to be in a stable health state. Studies included were required to have multiple test assessments for participants under the same conditions (testing and patient care). There was no restriction placed on study participants for inclusion, with studies of healthy participants and participants with disease included.

Multiple test assessment included repetition of part or all of the testing procedure; participants could be tested multiple times or a test could be assessed multiple times (for example, clinicians assessing imaging results or retesting of stored samples). To obtain an estimate of biological variability (*'The natural variability in a lab parameter due to physiologic differences among subjects and within the same subject over time.'*[15]), analytical variability needs to be assessed and allowed for when measuring within-individual and between-individual variability, so studies assessing analytical variability only were included. It was also required that assessment was of participants or participant samples rather than control samples.

Whilst conducting the review, key methodological papers influencing the design, analysis and reporting of the selected studies were identified and further texts were found from 'snowball' searches.[36] The main findings from the methodological papers are reported here.

2.3 Results

The literature for designing, analysing and reporting variability studies is better defined in laboratory science than for other areas of health care. This guidance is generally applicable to any test providing a continuous value, although some tests do not lend themselves to the assessment of analytical variability.

Studies of imaging tests have different methods for design, analysis and reporting as these studies include the variability from having a reader interpret the image. The image may provide a continuous value that can be used to guide patient care or it may provide a binary result (presence or absence of disease).

2.3.1 Design of biological variability studies

2.3.1.1 Studies of laboratory tests

In the area of clinical chemistry there are developed frameworks that are adhered to when studying the variability of tests. These frameworks are often used with laboratory tests but would be generally applicable to any test with a continuous outcome.

Fraser and Harris

The work of Fraser and Harris,[35] published in 1989, provides a framework for the design of biological variability studies conducted in the clinical chemistry laboratory setting. Fraser and Harris[35] state that studies with the aim of estimating biological variability should reduce pre-analytical variation where possible with this including a strict testing protocol; this may include tests being performed at the same time of the day, in the same place, by the same person and at regular time periods in addition to the participant having prepared for the test in the same way (diet and exercise) and consistent handling, transport and storage of samples.

With regard to analysing the samples in the laboratory the favoured approach (introduced by Cotlove and colleagues in 1971)[37] is to allow all samples to be collected for the duration of the study and then analyse all samples at the same time in duplicate.[35] The benefit of this approach is there is no variation due to the samples being analysed in different runs and the analytical variation can be estimated as the samples are analysed in duplicate. If the samples are not analysed in duplicate the analytical variance needs to be estimated using quality control materials, stored samples or existing literature. Fraser and Harris warn that this is not ideal, as the estimate of analytical variability: *'may not accurately reflect the true analytical variation achieved with the specimens from the subjects'*.[35] It is, however, acknowledged that in some cases (such as when samples are unstable) analytical variability will have to be assessed externally.[35] Fraser and Harris[35] also criticise the approach where samples are analysed as they are collected, as this will incur between-run variation.

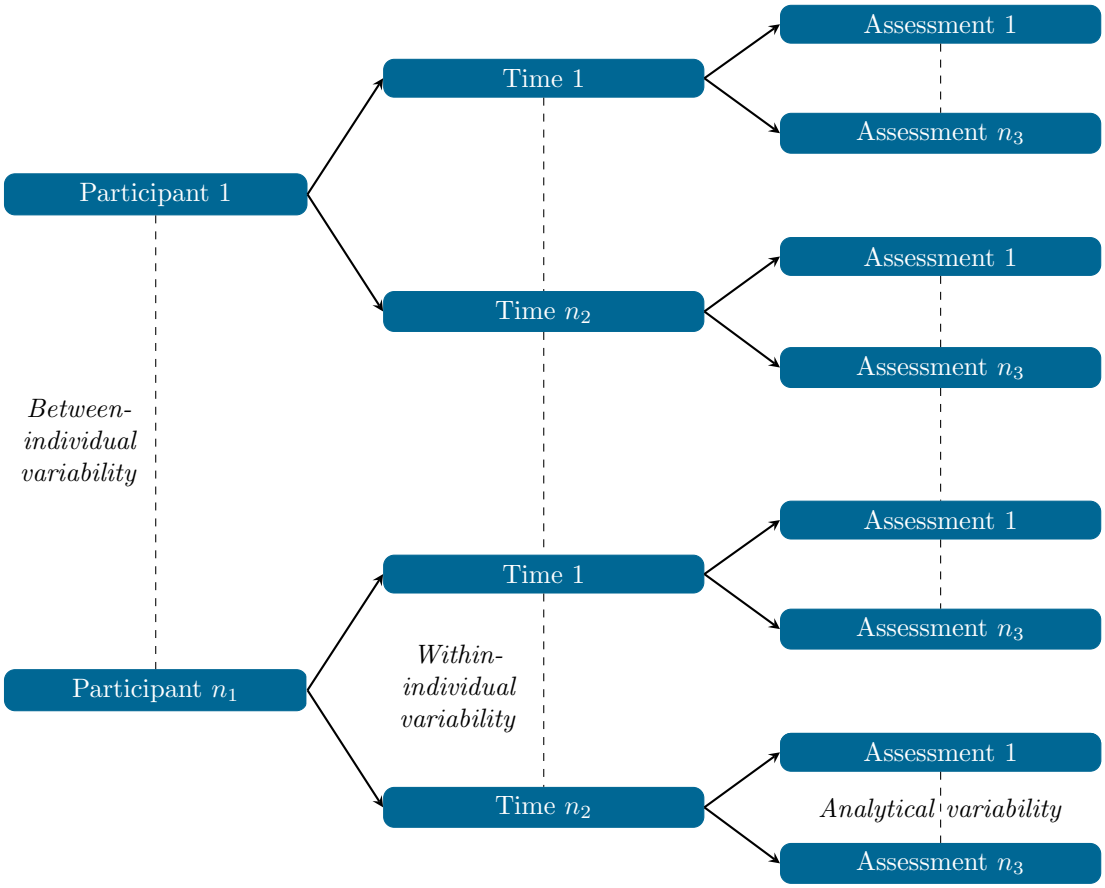


Figure 2.1: Biological variability study design.

In order to assess analytical, within-individual and between-individual variability, the design often used in laboratory test experiments is to recruit n_1 participants and take observations from these participants at n_2 time points, with each observation assessed n_3 times,[35] see Figure 2.1.

European Federation of Clinical Chemistry and Laboratory Medicine

The EFLM (European Federation of Clinical Chemistry and Laboratory Medicine) released a consensus statement in 2015[38] following the conference held in Milan updating the guidelines created in Stockholm 15 years previously.[39] This consensus statement offers three broad approaches; the first concerns analytical performance and the impact of this on patient outcomes; the second is based on estimating biological variation; and the final approach looks at comparing analytical variability to performance goals known as 'the-state-of-the-art'. The authors advocate the use of the first and second model (individually and in combination)

over the third as this does not consider impact on patients. There is also a move to make the catalogued biological variability information more reliable.[13]

2.3.1.2 Studies of physiological tests

Non-laboratory based studies may be designed to recruit a group of patients and take repeated measures for these patients without formal assessment of analytical variability. Indeed, for some physiological measures it may not be possible to directly assess analytical variability, for example spirometry and blood pressure testing. Use of these designs mean measurement error is estimated when calculating the variability within participants, combining analytical and within-individual variability. Laboratory studies may employ this design and calculate within-individual variability by obtaining an estimate of analytical variability from an external source. Simundic et al[22] stress the importance of clarification when reported results showing the combined analytical and within-individual variability.

2.3.1.3 Studies of imaging tests

Imaging tests can incur variability at the analytical, within and between-individual level, but often the focus of the studies is the variability within and between readers, which is not generally a concern for laboratory and physiological tests. For studies of intra and inter reader variability of imaging or patient charts, a typical study design would be to recruit and test each participant. Each result would then be independently assessed by multiple readers, with a reader(s) assessing each result multiple times.[23,40] Intra and inter reader studies often have few readers.[41] Repeat reading of images by readers may only be performed by a subset of readers or even a single reader.[23] In some studies stock images/records are repeatedly read (for example To et al[42]). Readers may be chosen purposively to represent experienced and inexperienced readers.[40]

2.3.1.4 Populations investigated

The recommendation of Fraser and Harris,[35] when assessing the variability of laboratory based tests, is to only include apparently healthy participants in studies assessing biological variability. The Milan consensus issued by the EFLM discussed how results can vary due to the population and health care setting and participants being in a 'steady disease state'.[38] This issue is also considered by Aarsand et al[13] when discussing the reliability of estimates from biological variability studies.

For imaging and physiological tests no guidelines were identified suggesting the populations to be tested but studies identified in the review (see Chapter 3) indicate participants with disease were tested more often.

2.3.1.5 Sample size

There is limited guidance to inform the sample size, number of repeat measures, and timing of measures when designing studies of test variability.

Studies of laboratory tests

Fraser and Harris[35] use previous studies to support the view: *'valid estimates of the components of variation can be obtained from relatively small numbers of specimens collected from a small group of subjects over a reasonably short period of time'.*[35] Subsequent work by Fraser[16] states: *'the number of subjects is a compromise between the large number that is the ideal and the smaller number that can be handled in any good experimental design'.*

Another approach for laboratory tests has been to evaluate the number of repeat samples required (n) for an estimate to be within x% of the homeostatic setting point, which is $n = \left(1.96\left(\frac{\sqrt{CV_A^2+CV_I^2}}{x}\right)\right)^2$.[16] The homeostatic setting point is the 'true' value for an individual. This approach considers only the estimate of within-individual variability and focusses on the number of repeat measures that would be required each time the test is performed. See §2.3.2.4 for explanation of notation.

More recently Røraas and colleagues[43] assessed the power of biological variability studies using simulation based on analysis using ANOVA, for varying numbers of individuals, samples, replicates and levels of analytical variability. The tables detailing the power of studies are available to guide investigators to design biological variability studies with adequate power.

Studies of physiological tests

In the medical testing setting studies of variability are often called reliability studies which may be referred to as generalisability studies (G) and studies where the most reliable strategy involving a decision making process is assessed, which are decision (D) studies.[44] de Vet et al[44] suggest that a sample size of 50 is appropriate for the measurement of variability reasoning that this sample size will be adequate to fill a 2×2 table and will allow a Bland-Altman plot to be reasonably populated.

de Vet and colleagues[44] discuss how statistical significance is not a component of devising the sample size required to produce a reliability estimate, as it is the value of the estimate which provides information about the ability of the measuring system rather than the difference from zero. The importance of an adequate sample size is stressed as this will allow an adequate 95% confidence interval for the reliability parameter to be estimated with Streiner et al[23] also commenting on how sample size influences the accuracy of the reliability coefficient.

de Vet et al[44] acknowledge that sample size guidance is difficult to locate and provide sample size formula for intraclass correlation coefficients (ICCs) allowing the sample size (n) to be calculated for a pre-specified CI. The formula is taken from Giraudeau and Mary:[45]

$$n = \frac{8z^2_{1-\alpha/2}(1 - ICC)^2[1 + (n_2 - 1)ICC]^2}{n_2(n_2 - 1)w^2},$$

where n_2 is the number of measures of each participant and w is the width of the $100(1-\alpha)\%$ confidence interval for the ICC. See §2.3.2.4 for explanation of notation.

Studies of imaging tests

When imaging studies produce a binary outcome, the measure commonly used is the Kappa statistic (see §2.3.2.3). It is acknowledged that sample sizes for Kappa statistics need to be larger as the data is categorical; however, sample size calculations for a Kappa statistic are more difficult to perform as an estimate of Kappa and other distributional knowledge is required.[44]

2.3.2 Analysis of biological variability studies

2.3.2.1 Preparing data for analysis

In studies of laboratory tests data is tested for normality and subjected to outlier detection methods prior to analysis. For imaging and physiological studies no such practices were identified.

Normality of data

For laboratory studies of biological variation, Fraser and Harris[35] advocate the use of the Shapiro-Wilk test for normality. This test is applied to results for each individual separately and the data is transformed (log transformation) if the results for many individuals are not normally distributed.[35] Braga et al[46] have offered further guidance introducing Kolmogorov-Smirnov tests in addition to Shapiro-Wilk (see Box 2.1).

When using a model to evaluate three levels of variability, the model assumes normality of the variability parameters at the analytical, within-individual and between-individual levels. If this assumption is not held results from the model may not be valid. Simply assessing the normality of the test measures may not be sufficient to investigate if the data meets the assumptions of the model. Many of the outlier detection methods (see Chapter 6) rely on the data to be normally distributed, hence the further requirement for normality prior to assessing outliers.

It may also be desirable to convert to the log scale as when biological variability data are log-

normally distributed, coefficients of variation (on the original scale) can easily be estimated using the standard deviation of the log transformed data.[47]

For studies of physiological and imaging tests no guidance on normality checking and transformation of data was identified.

Box 2.1: The Shapiro-Wilk test.

The Shapiro-Wilk test is the primary test used for assessing normality of data in studies of laboratory test variability. The Shapiro-Wilk test uses analysis of variance to test for normality, with the null hypothesis that data are normally distributed.[48] Given a variable y is ordered, such that $y_1 < y_2 < \ldots < y_n$ the test-statistic for the Shapiro-Wilk test is:

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean value of the variable and $a_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$. $m = (m_1, \ldots, m_n)^T$ are the expected values of ordered statistics and V is the corresponding covariance matrix.[49]

Outliers

In the laboratory setting, it is recommended data are assessed for outlying results and these should be removed prior to analysis. Most commonly, Cochran C test and Reed's criterion are used to identify data to be removed, as recommended in the Fraser-Harris Framework.[16,35] The Cochran C test is used to assess variances within the duplicated results and if an outlier is detected both results are removed; this test is used again to assess the variances of results for each individual. Reed's criterion is used to identify if the mean value for any individual is an outlier.[16] Other methods of outlier detection are used in the laboratory setting and are further discussed (See Chapter 6).

In other areas of test evaluation an alternative view of outliers is taken. According to de Vet et al outliers should not be deleted as errors do occur and may indicate difficulties with measurement read outs or interpretation of scales.[44]

2.3.2.2 General method for analysis of test variability studies

Across variability studies for laboratory, physiological and imaging tests the analysis allowing the estimates of variability to be quantified is the same and comes from a model allowing the variability at each level to be estimated, or equivalently analysis of variance, ANOVA.

The differences between studies of test types occur after this analysis has been performed. Fraser and Harris[35] provide guidance for the analysis of biological variability studies conducted in the clinical chemistry laboratory setting. ANOVA is suggested as the method for analysis providing estimates for each component of variance, which is equivalent to fitting a linear regression model with only random effects when the number of observations within participants is consistent for all participants and the number of assessments at each observation point is consistent for all observation points (a balanced design). ANOVA is performed by the calculations shown in Table 2.1.

Table 2.1: ANOVA (adapted from Burdick and Graybill[34]).

	DF	SS	MS	EMS
Between participants	$G = n_1 - 1$	$n_2 n_3 \sum_i (\bar{y}_i - \bar{y}_i)^2$	S_G^2	$\theta_G = \sigma_A^2 + n_3\sigma_I^2 + n_2 n_3 \sigma_G^2$
Within participants	$I = (n_2 - 1)n_1$	$n_3 \sum_i \sum_j (\bar{y}_{ij} - \bar{y}_i)^2$	S_I^2	$\theta_I = \sigma_A^2 + n_3\sigma_I^2$
Within assessments	$A = (n_3 - 1)n_1 n_2$	$\sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{ij})^2$	S_A^2	$\theta_A = \sigma_A^2$
Total	$T = n_1 n_2 n_3 - 1$	$\sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y})^2$		

DF is degrees of freedom; SS is sum of squares; MS is mean squares; and EMS is expected mean squares

Where S^2, the mean squares, is equal to the sum of squares divided by the degrees of freedom, as is estimated by θ. In model notation this can be expressed as $y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, where $i = 1, .., n_1$, $j = 1, .., n_2$ and $k = 1, .., n_3$, y_{ijk} is the test measure for the ith participant at the jth time point and for the kth assessment, μ is the mean value of the measure, $\alpha_i \sim N(0, \sigma_G^2)$, $\beta_{ij} \sim N(0, \sigma_I^2)$ and $\epsilon_{ijk} \sim N(0, \sigma_A^2)$.

When using a linear regression model to estimate the variance parameters, the use of restricted maximum likelihood estimation (which estimates the fixed and random effects independently) is preferred as the estimates are less biased compared to maximum likelihood estimation methods, especially with small sample sizes.[50] McNeish and Stapleton[50] recommend a minimum of ten observations per individual to estimate a variance parameter.

Specialist ANOVA methods

Specifically referring to studies of laboratory tests, further work by Røraas et al[51] has investigated the performance of the standard ANOVA method alongside ln-ANOVA and CV-ANOVA methods. The ln-ANOVA method uses log transformed values and the method CV-ANOVA normalises data for each individual (by dividing by the mean) prior to ANOVA being performed. The ln-ANOVA method is often advocated as it easily allows estimation of coefficients of variation (CV), see §2.3.2.4. The CV-ANOVA method is an alternative approach to obtaining CVs; however, it is not possible to estimate between-individual variability with this method.[51]

2.3.2.3 Alternative methods for analysis of test variability studies

Bland-Altman method

The Bland-Altman method[52] is primarily used when comparing two methods of measuring the same outcome in the same individuals with the Bland-Altman method providing a measure of agreement, and is often used for physiological measures. The method comprises the calculation of the mean difference between the measures and 95% limits of agreement around this estimate ($\bar{d} \pm 1.96\sigma_d$), where σ_d is the standard deviation of the differences between measures for each individual, this is often displayed graphically also. When describing this method, Bland and Altman provide detail regarding the estimation of repeatability of a measure rather than comparison between two measures, suggesting ANOVA is used to provide an estimate of within-individual standard deviation which can then be used to compare repeatability between measurement methods.

Bland and Altman advise that repeatability is evaluated and this is taken into account when comparing methods of measurement. Bland and Altman also provide the repeatability coefficient (CR), which is calculated as $1.96\sqrt{2}\sigma_{A+I}$, where σ_{A+I} is the calculated standard deviation relating to measurement error. The repeatability coefficient is how close two readings made using the same method will be for 95% of subjects (an absolute measure), with this measure akin to the reference change value (RCV) but with variability expressed as standard deviations and the within-individual variation estimate including analytical variation rather

than estimated separately as measures are not performed in duplicate. This method is used when assessing clinical measures.[53]

Kappa

Agreement within and between observers (or raters) is known as intra and inter rater variability. A basic way of expressing this level of agreement is percentage agreement, which is the percentage of measurements showing agreement of the total number assessed. Cohen's Kappa extends this idea by allowing for agreement occurring by chance. Kappa is calculated as:

$$\kappa = \frac{p_o - p_e}{1.0 - p_e},$$

where p_o is the proportion of cases where there is agreement and p_e is the proportion of cases where agreement would be expected by chance. The corresponding standard error for Kappa is:

$$SE(\kappa) = \sqrt{\frac{p_o(1-p_o)}{n_1(1.0-p_e)^2}},$$

where n_1 is the number of participants in the study. Cohen's Kappa measures agreement by providing a value between -1 and 1, with a value of 0 indicating chance agreement.[54]

Kappa is used in situations where the result of a test is categorical rather than a continuous measurement as it is difficult to achieve complete agreement when a continuous measure is used.

Capability measures

Outside of medicine, in the area of industrial management, there are measures of variability for monitoring ongoing processes. Processes are monitored by taking repeated measures, and controlled by comparing these to 'capability measures' calculated using variability estimates. Precision to tolerance ratio (PTR) and signal to noise ratio (SNR) are capability measures. PTR is calculated as $\frac{k\sqrt{\sigma_A^2}}{USL-LSL}$, where USL and LSL are specification limits and k is equal to

5.15 or 6 (this is the number of standard deviations between the 'natural' tolerance limits and correspond to the central 99% or 99.73% of the 'process' respectively) and σ_A^2 is the variance of the measurement system. SNR is $\sqrt{2\frac{\sigma_G^2}{\sigma_{A+I}^2}}$, using the ratio of the process variance or between individual variance (σ_G^2) to the measurement variance (σ_{A+I}^2). The SNR value indicates how many categories can be reliably identified; SNR values of five and above mean monitoring is recommended and SNR values of less than two indicate little benefit of monitoring. SNR and PTR are plotted against each other, with 95% confidence intervals to account for uncertainty of the estimates and these plots are used to assess the measurement system. Also used are: intraclass correlation (ICC), a measure of the proportion of the total variation that is attributed to the process, $\rho = \frac{\sigma_G^2}{\sigma_G^2+\sigma_{A+I}^2}$; and, process capability, $C_p = \frac{USL-LSL}{6\sqrt{\frac{\sigma_G^2}{\sigma_{A+I}^2}}}$, this is to understand the ability of the process rather than the measurements.[55]

2.3.2.4 Reported measures from test variability studies

There are many different terms used to describe measures of test variability. These terms refer to related concepts, and although some terms have precise definitions, in practice, they are often used interchangeably. Streiner et al[23] list some of the definitions used for variability as: reliability, objectivity, reproducibility, stability, agreement, association, sensitivity, precision, accuracy, dependability, repeatability and consistency. The exact definitions of these concepts are not defined here, but the statistical measures presented to describe variability are discussed. Some measures of test variability are akin to those in other research areas but may be specified differently.

Reported measures for studies of laboratory tests

Estimated standard deviations are used to set goals for analytical performance. Measures include: coefficients of variation, reference change values and index of individuality.[56] See Table 2.2 for details of all measures reported in this section.

Coefficient of variation, or CV, is defined as σ/μ where σ is the standard deviation and μ is the mean.[16] Coefficients of variation are often expressed as percentages and allow measures

to be interpreted in relation to the value of the mean. When calculating the coefficient of variation at the analytical, within-individual and between-individual levels the corresponding standard deviations are used and are divided by the overall (grand) mean to estimate the CV.

The reference change value (RCV), also known as critical difference (CD), estimates the difference between two results for an individual that would indicate a real change above the expected random fluctuation in test measures.[16] RCV is calculated using $\sqrt{2}Z\sqrt{(CV_A^2 + CV_I^2)}$, where Z is selected from the normal distribution (usually 1.96) and CV_A and CV_I are the coefficients of variation at the analytical/reassessment and individual level respectively.[16]

The index of individuality (II), provides information regarding the best use of a test by quantifying the ratio of variability at the analytical and within-individual levels to the between-individual level, and is calculated using $\sqrt{(CV_A^2 + CV_I^2)}/CV_G$ where CV_A, CV_I and CV_G are coefficients of variation analytical, within and between individuals respectively. II is often simplified to CV_I/CV_G, with the assumption that $CV_A << CV_I$.[16,22] Higher values indicate a general threshold would be reasonable and lower values suggest changes from previous results for an individual will more be meaningful.[16,57]

Also used is the index of analytical error, CV_A/CV_I. This measure is the ratio of analytical variation to within-individual variation and is often used for setting analytical variability goals.[16]

Total variance is the total variability in the data and is calculated by combining variances at the analytical, within-individual and between-individual levels, $\sigma_A^2 + \sigma_I^2 + \sigma_G^2$.[46] This combined variability can also be expressed as a standard deviation (σ_{A+I+G} or σ_{TOT}) or as a coefficient of variation (CV_{A+I+G} or CV_{TOT}). Another combined variance is total error, or measurement error, combing the analytical and within-individual variability. Total error variance is calculated as: $\sigma_A^2 + \sigma_I^2$. Again, total error can also be expressed as a standard deviation (σ_{A+I} or σ_{TE}) or coefficient of variation (CV_{A+I} or CV_{TE}).

The index of heterogeneity (IH) is a measure of the heterogeneity of within-individual observations. IH measures the ratio of CV_{A+I} to the theoretical CV and is calculated as:

$IH = \frac{CV_{A+I}}{\sqrt{\frac{2}{n_2-1}}}$, where n_2 is the number of observations for each individual.[46] Under the assumption of no heterogeneity, the expected value is $1/\sqrt{2n_2}$. If the calculated IH is more than twice the expected value, the IH indicates heterogeneity.[46]

The Validity Coefficient (VC) represents the difference between the measured value and the true value, due to variability. $VC = \sqrt{\left(1 + \frac{\sigma_I^2}{n_2\sigma_G^2}\right)^{-1}}$, where σ_I and σ_G are the standard deviations at the within-individual and between-individual level respectively and n_2 is the number of observations from each individual.[58]

To demonstrate how the measures of variability differ in different situations estimates are produced for four test scenarios: scenario A is a test with good analytical variability and within-individual variability is much lower than between-individual variability; scenario B is a test with the within-individual variability larger in comparison to between-individual variability; scenario C is a test with poorer analytical variability; and scenario D is a test with decreased between-individual variability, see Table 2.3. For further details of how these measures can be employed for use in monitoring individuals see Chapter 7.

Studies of laboratory tests–log-normal data

Often in laboratory based studies data are considered to be log-normally distributed. Different measures are reported when data follow a log-normal distribution.

When data requires log-transformation prior to analysis using ANOVA or modelling, the estimated standard deviations are geometric CVs and these can be multiplied by 100 to be expressed as a percentage.[59] The exact geometric CV of values on the original scale (assuming a log-normal distribution) can be estimated by $\sqrt{exp(\sigma^2) - 1}$, where σ is the standard deviation of the log transformed values, with $\sqrt{exp(\sigma^2) - 1} \times 100$ to express CV as a percentage.[59,60] This calculation of the coefficient of variation arises as standard deviations of (natural) log transformed data represent a fraction standard deviation, the CV, and with the equation expressed giving the exact relationship.[59] An alternative way of using log transformed data to provide an estimate of geometric CV is given by Kirkwood,[61] as $exp(\sigma) - 1$ and with the CV expressed as a percentage using $(exp(\sigma) - 1) \times 100$, where σ is again the standard deviation of the log transformed values.

When using log-normal data RCV limits are asymmetrical, meaning the positive and negative difference are calculated separately. The percentage RCV limits when using log-normal data are calculated using: $RCV_{pos} = [exp(Z \times \sqrt{2}\tau) - 1] \times 100$, and $RCV_{neg} = [exp(-Z \times \sqrt{2}\tau) - 1] \times 100$ (where $\tau = \sqrt{ln(CV_{A+I}^2 + 1)}$, Z is selected from the normal distribution (usually 1.96) and CV_{A+I} is the coefficient of variation for the total imprecision, $CV_{A+I} = \sqrt{CV_A^2 + CV_I^2}$, with CV_A and CV_I).[62,63]

The measures for log-normal data for the four test scenarios are shown in Table 2.4.

Reported measures for studies of physiological tests

In general, reliability parameters are used to assess the ability of participants to be distinguished from each other and are forms of ICC. The reliability parameter is

$$R = \frac{\sigma_G^2}{\sigma_{TOT}^2} = \frac{\sigma_G^2}{\sigma_{A+I}^2 + \sigma_G^2},$$

where σ_{TOT}^2 is the total variance of measures and σ_G^2 is the 'true' variability between participants. Here, σ_{A+I}^2 is 'measurement error' which includes both within individual variability and analytical variability, this would be obtained if the study design included only repeat measures of a group of individuals and no assessment of analytical variability by repeated assessment of measures. The reliability parameter takes values between 0 and 1, with a value of 1 meaning the system is perfect at identifying individuals, and 0 meaning it is unreliable.

Reported measures for studies of imaging tests

ICC measures are further defined for studies considering the variance due to raters (identified in inter-intra reader studies of imaging tests), these measures are further defined as ICCs for agreement and consistency.

$$ICC_{agreement} = \frac{\sigma_G^2}{\sigma_A^2 + \sigma_I^2 + \sigma_G^2},$$

where σ_A^2 is residual variance, σ_I^2 is the rater variance and σ_G^2 is the between-individual variance, and

$$ICC_{consistency} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_G^2},$$

where σ_A^2 is the residual variance and σ_G^2 is the variance of the 'true' scores of the participants.

For the consistency measure the rater variance is not considered; however, if interest is in agreement between measures the variance due to raters is included.[44]

The reliability parameter can also be used when there are multiple (k) measures used to generate an average,

$$R = \frac{k^2\sigma_G^2}{k\sigma_{A+I}^2 + k^2\sigma_G^2}.$$

A more reliable measure will always be achieved when based on the average of measurements as the measurement error becomes smaller.[44] Streiner et al advise that averaging of measures is only used if this reflects practice.[23]

The relationship between the reliability parameter, R, and the standard error of measurement (SEM) is: $SEM = \sigma\sqrt{1-R}$, where σ is the standard deviation of observed values.[23] When concerned with the variability of a measure itself it is of interest to assess the magnitude of measurement error. For continuous outcomes the standard error of measurement error and coefficients of variation are also often reported, and when comparing agreement between tests the limits of agreement from the Bland-Altman method are used. When the outcome is categorical (say a diagnosis) Cohen's Kappa and weighted Kappa are often used to judge agreement. There are no measurement error estimates for categorical outcomes; instead the percentage of outcomes in agreement is usually used in some way.[44] The use of ICCs and ANOVA is advocated by Streiner et al rather than Bland-Altman and Kappa.[23]

Table 2.2: Commonly reported variability measures.

Measure	**Formula**
Commonly used in studies of laboratory tests	
CV	σ/μ
Reference change value (RCV)	$\sqrt{2}Z\sqrt{(CV_A^2+CV_I^2)}$, where Z is selected from the normal distribution
Index of individuality (II)	$\sqrt{(CV_A^2+CV_I^2)}/CV_G$
Index of analytical error	CV_A/CV_I
Total variance	$\sigma_A^2+\sigma_I^2+\sigma_G^2$
Index of heterogeneity (IH)	$\frac{CV_{A+I}}{\sqrt{\frac{2}{n_2-1}}}$
Validity coefficient (VC)	$\sqrt{\left(1+\frac{\sigma_I^2}{n_2\sigma_G^2}\right)^{-1}}$
After log transformation	
Exact geometric CV	$\sqrt{exp(\sigma^2)-1}$
Alternative exact geometric CV	$exp(\sigma)-1$
Asymmetric RCV bounds	$exp(\pm Z\times\sqrt{2}\tau)-1$, where $\tau=\sqrt{ln(CV_{A+I}^2+1)}$
Commonly used in studies of physiological and imaging tests	
Reliability parameter (R)	$\frac{\sigma_G^2}{\sigma_{A+I}^2+\sigma_G^2}$
Reliability parameter with k measurements	$\frac{k^2\sigma_G^2}{k\sigma_{A+I}^2+k^2\sigma_G^2}$
ICC agreement	$\frac{\sigma_G^2}{\sigma_A^2+\sigma_I^2+\sigma_G^2}$
ICC consistency	$\frac{\sigma_A^2}{\sigma_A^2+\sigma_G^2}$

Table 2.3: Examples of variability: measures for normally distributed data.

	σ_A	σ_I	σ_G	CV_A	CV_I	CV_{A+I}	CV_G	CV_{TOT}	RCV	II	IA	IH	Bound[a]	VC	ICC_A	ICC_I	ICC_G	CR
A	0.2	0.6	2	2	6	6.32	20	20.98	17.53	0.32	0.33	0.08	0.35	6.74	0.01	0.08	0.91	1.75
B	0.2	1.6	2	2	16	16.12	20	25.69	44.69	0.81	0.13	0.20	0.35	2.69	0.01	0.39	0.61	4.47
C	0.4	0.6	2	4	6	7.21	20	21.26	19.99	0.36	0.67	0.09	0.35	6.74	0.04	0.08	0.88	2.00
D	0.4	0.6	1	4	6	7.21	10	12.33	19.99	0.72	0.67	0.09	0.35	3.48	0.11	0.24	0.66	2.00

[a]IH bound assuming 4 observations per individual.
Mean value is 10 for all scenarios. All CVs and RCVs are expressed as percentages.
Scenario A is a test with good analytical variability and within-individual variability is much lower than between-individual variability; scenario B is a test with the within-individual variability larger in comparison to between-individual variability; scenario C is a test with poorer analytical variability; and scenario D is a test with decreased between-individual variability.

Table 2.4: Examples of variability: measures for log-normally distributed data.

	σ_A	σ_I	σ_G	CV_A	CV_I	CV_{A+I}	CV_G	CV_{TOT}	RCV	RCV -	RCV +	II	IA	IH	Bound[a]	VC	ICC_A	ICC_I	ICC_G
A	0.02	0.06	0.2	2.00	6.01	6.33	20.20	21.21	17.55	-16.08	19.16	0.31	0.33	0.08	0.35	6.74	0.01	0.08	0.91
B	0.02	0.16	0.2	2.00	16.10	16.23	20.20	26.12	44.98	-36.04	56.35	0.80	0.12	0.20	0.35	2.69	0.01	0.39	0.61
C	0.04	0.06	0.2	4.00	6.01	7.22	20.20	21.50	20.00	-18.12	22.13	0.36	0.67	0.09	0.35	6.74	0.04	0.08	0.88
D	0.04	0.06	0.1	4.00	6.01	7.22	10.03	12.38	20.00	-18.12	22.13	0.72	0.67	0.09	0.35	3.48	0.11	0.24	0.66

[a]IH bound assuming 4 observations per individual. All CVs are geometric exact. All CVs and RCVs are expressed as percentages.
Scenario A is a test with good analytical variability and within-individual variability is much lower than between-individual variability; scenario B is a test with the within-individual variability larger in comparison to between-individual variability; scenario C is a test with poorer analytical variability; and scenario D is a test with decreased between-individual variability.

2.3.2.5 Use of confidence intervals

In the laboratory test setting, Fraser and Harris comment on the use of confidence intervals to express uncertainty of estimates produced but caution these are often not considered due to distributional assumptions.[35] Røraas et al[43] state that biological variability estimates should be presented with confidence intervals to allow the uncertainty around the estimate to be understood. It is acknowledged that confidence intervals are rarely seen in biological variability studies and as a consequence it is difficult to compare results across studies. This issue was raised by Henderson[64] with debate from Harris[65] in 1993 in the Journal of Clinical Chemistry. Henderson called for the consistent use of confidence intervals in the journal, arguing use was commonplace in other fields of research, and requested this be made a requirement; however, Harris was reluctant to employ this. At present the author guidelines for the Journal of Clinical Chemistry[66] state confidence intervals should be used 'when appropriate'.

Røraas et al[43] considered the standard biological variability study design with individuals providing multiple samples and replicated analyses of these samples using ANOVA. Using the formula introduced by Burdick and Graybill,[34] Røraas et al provided confidence intervals calculated for a varying numbers of individuals, samples, replicates and levels of analytical variability. Researchers are encouraged to use the tables provided, demonstrating the width of confidence intervals under certain designs, to estimate the width of the confidence interval around an estimate. Røraas et al[67] have subsequently investigated the ability of methods to generate confidence intervals for estimates of within-individual biological variability. Burdick and Graybill[34] proposed methods to calculate approximate intervals for variance components (see §2.3.2.5).

Burdick et al[55] also comment on the use of confidence intervals, as for gauge reliability and reproducibility studies, it is often not possible to calculate exact confidence intervals for estimates of variability. Two alternatives are discussed, firstly, the modified large sample (MLS) approach[55] and, secondly, the computer intensive approach using generalised confidence intervals. The authors warn intervals obtained from likelihood based methods (for example REML models) are only valid for large samples and may not be appropriate.

For other test types, the use of confidence intervals is commonplace but the issues with obtaining exact confidence intervals for measures of variability remain. The default of many computer packages is to display confidence intervals using an approximation via the delta method.

Confidence intervals for variance components

These formulas are appropriate for any test type where estimates of variability at the analytical, within-individual and between-individual levels have been estimated.

Using the assumptions of the general model for data from biological variability studies: $y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, where μ is the mean value of the measure, $\alpha_i \sim N(0, \sigma_G^2)$, $\beta_{ij} \sim N(0, \sigma_I^2)$, $\epsilon_{ijk} \sim N(0, \sigma_A^2)$ and $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$, $n_1 S_1^2/\theta_1$, $n_2 S_2^2/\theta_1$ and $n_3 S_3^2/\theta_1$ are jointly independent chi-squared random variables, see ANOVA notation in Table 2.1. Using the formulas of Burdick and Graybill[34] it is possible to calculate confidence intervals calculated using the expected mean squares (as shown in Table 2.1). These equations have also been coded in a shiny application allowing visualisation of the confidence intervals for different estimates and sample sizes, `https://alicesitch.shinyapps.io/bvs_cis/`.

An exact two-sided confidence interval for analytical variation (σ_A^2), estimated by θ_A is given by: $\left[\frac{S_A^2}{F_{\alpha:A,\infty}}; \frac{S_A^2}{F_{1-\alpha:A,\infty}}\right]$, with this converted to give the confidence interval for σ_A by taking the square root of both the lower and upper bound.

To calculate a confidence interval for within-individual variation (σ_I^2) the formula

$$\left[\frac{S_I^2 - S_A^2 - \sqrt{V_{IL}}}{n_3}, \frac{S_I^2 - S_A^2 + \sqrt{V_{IU}}}{n_3}\right]$$

is used and for between-individual variation (σ_G^2) the formula

$$\left[\frac{S_G^2 - S_I^2 - \sqrt{V_{GL}}}{n_2 n_3}, \frac{S_G^2 - S_I^2 + \sqrt{V_{GU}}}{n_2 n_3}\right]$$

is used where:

$$V_{GL} = G_G S_G^4 + H_I^2 S_I^4 + G_{GI} S_G^2 S_I^2$$

$$V_{GU} = H_G S_G^4 + G_I^2 S_I^4 + H_{GI} S_G^2 S_I^2$$

$$V_{IL} = G_I S_I^4 + H_A^2 S_A^4 + G_{IA} S_I^2 S_A^2$$

$$V_{IU} = H_I S_I^4 + G_A^2 S_A^4 + H_{IA} S_I^2 S_A^2$$

$$G_G = 1 - \frac{1}{F_{\alpha:G,\infty}}$$

$$G_I = 1 - \frac{1}{F_{\alpha:I,\infty}}$$

$$G_A = 1 - \frac{1}{F_{\alpha:A,\infty}}$$

$$H_G = \frac{1}{F_{1-\alpha:G,\infty}} - 1$$

$$H_I = \frac{1}{F_{1-\alpha:I,\infty}} - 1$$

$$H_A = \frac{1}{F_{1-\alpha:A,\infty}} - 1$$

$$G_{GI} = \frac{(F_{\alpha:G,I-1})^2 - G_G^2 F_{\alpha:G,I} - H_I^2}{F_{\alpha:G,I}}$$

$$G_{IA} = \frac{(F_{\alpha:I,A-1})^2 - G_I^2 F_{\alpha:I,A} - H_A^2}{F_{\alpha:I,A}}$$

$$H_{GI} = \frac{(1 - F_{1-\alpha:G,I})^2 - H_G^2 F_{1-\alpha:G,I}^2 - G_I^2}{F_{1-\alpha:G,I}}$$

$$H_{IA} = \frac{(1 - F_{1-\alpha:I,A})^2 - H_I^2 F_{1-\alpha:I,A}^2 - G_A^2}{F_{1-\alpha:I,A}}.$$

Confidence intervals for coefficients of variation

Obtaining exact confidence intervals for coefficients of variation involves solving non-linear equations. Assuming the coefficient of variation is positive, the lower bound of a confidence interval can be obtained solving the following for β:

$$\alpha/2 = F_{NCT}(n-1, \sqrt{n}/\beta)(\bar{X}/(\sigma/\sqrt{n})),$$

where $1-\alpha$ represents the confidence level required, $F_{NCT}(n-1, \sqrt{n}/\beta)$ is a non-central t-distribution with $n-1$ degrees of freedom and $\sqrt{n}/\beta$ non-centrality parameter, $\bar{X}$ is the sample mean, σ is the sample standard deviation and n is the sample size. To obtain the upper bound, the following equation must be solved for β:

$$1-\alpha/2 = F_{NCT}(n-1, \sqrt{n}/\beta)(\bar{X}/(\sigma/\sqrt{n})).^{60,68}$$

In addition to the computational complexities in obtaining exact confidence intervals for coefficients of variation, this method can give confidence bounds of infinity. There are also approximate methods for calculating confidence intervals for coefficients of variation, but the issue of infinite upper bounds is more apparent for approximate methods.[68] There is no closed form solution for deriving confidence intervals for reference change values.[51]

Confidence intervals for coefficients of variation–log-normal data

When data are log-normal, the distribution of log transformed data (Y_i) is normally distributed ($Y_i \sim N(\mu, \sigma^2)$) and a $1-\alpha$ confidence interval for σ^2 is $[a_L, a_U]$.

$$a_L = (n-1)\sigma^2/(F_{\chi^2}(n-1)^{-1}(1-\alpha/2)), \text{ and}$$

$$a_U = (n-1)\sigma^2/(F_{\chi^2}(n-1)^{-1}(\alpha/2)),$$

where σ is the sample standard deviation of the log transformed data (Y_i) and $F_{\chi^2}(n-1)(x)$ is a cumulative distribution function of a central chi-squared distribution with $n-1$ degrees of freedom. This formula is equivalent to the formula for a confidence interval for σ_A provided by Burdick and Graybill due to the χ^2 distribution and F distribution being equivalent under certain conditions.

Hence, to obtain a $1-\alpha$ confidence interval for the coefficient of variation for log-normal data, the following is used:

$$\left[\sqrt{exp(a_L)-1}, \sqrt{exp(a_U)-1}\right].^{68}$$

2.3.2.6 Reporting of biological variability studies

Studies of laboratory tests

There are no guidelines for the reporting of laboratory biological variability data, unlike the reporting of reference range estimates. Bartlett and colleagues,[69] on behalf of the Biological Variation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine (EFCCLM) developed a checklist for the appraisal of existing and future publications of biological variability data.[69] This checklist is developed from the idea of a minimum dataset that must be provided to allow readers to accurately interpret results of biological variability studies.[70] The minimum dataset to be reported includes information on the: test, population, study, analysed data and also suggests linking to a publication with further details and a rating of the study (in development).[69,70] The checklist proposed has six domains: title/abstract/keywords, introduction, data analysis, results and discussion, see Table 2.5. Simundic et al suggest that consistent notation is used to avoid confusion when reporting results from biological variability studies.[22]

Studies of physiological and imaging tests

For variability studies in other test areas there are reporting guidelines (GRRAS).[40] These guidelines include clear identification of the study type in the title; describing the testing measurement or device, the population, rater population and study rationale; explanation of the chosen sample size (number of subjects, raters and replicates), the sampling method, the measurement and rating process; acknowledgement of independence; description of statistical analyses; the results must state the number of raters, subjects and duplicates, describe the raters and subjects and give 'reliability and agreement' measures with statistical uncertainty. Further to these points, there should be discussion of the relevance of results and further detailed results.

Table 2.5: Biological variability study checklist. After Bartlett.[69]

Section and topic	Item	Evidenced
Title/abstract/keywords	1	The title should indicate that the content relates to a study of biological variation, the subject of the study, the sample matrix, and the population studied. Analyte (component being measured), the measurand/s (the quantity or quantities to be measured, see Section 1.1), and state of well-being of the subjects under study should be clearly and unambiguously identified. Relevant coding systems might be employed, (e.g., LOINC,[71] SNOMED,[72] C-NPU.[73])
Abstract	1.1	As a minimum it should contain the headline biological variation data, the major characteristics of the population studied (numbers of subjects with demographics), clearly identify the analyte and measurand/s studied [the analyte quantities studied in a particular sample matrix, (e.g., concentration of glucose in plasma)], the statistical approach taken, the duration of the study and the geographical location of the study.
Introduction	2	Introduction should clearly identify the context and aims of the study and cite any previous relevant studies of biological variability of the target analyte. Recommended terminology to be adopted re description of variability.[22]
Methods	3	Described in enough detail to facilitate transportability of the derived data across populations and health care systems. The biological variation data produced are effectively reference data and their applicability requires delivery of appropriately described metadata to enable their use as such.
Analyte/measurand	3.1	The described study should clearly identify the target analyte and measurand/s. Where available internationally agreed terminology and codings should be utilised.
Subjects	3.2	The description of the subjects and population studied should be detailed enough to enable transportability of the biological variation data. Minimum data set should be present.[70,74,75] This should include number of subjects studied, age, gender, and state of well-being.
Measurement procedure	3.3	A clear description of the analytical methodology used should form part of the metadata. This may be made available via an appropriate reference or be presented within the publication. Deviation from standard operating procedures, use of adaptations of published methods, and deviation from manufacturers recommended methods in the case of commercially available systems should be documented. Standardisation and traceability should be clearly identified.
Length of study	3.4	Length of the study periods should be clearly identified.

[a] Tests to determine the power of a study to identify heteroscedasticity need to be developed. If variances are not homogenous derived estimates of biological variation cannot be trusted, and are not representative for the population in which it is examined.

Section and topic	Item	Evidenced
Sampling	3.5	Sampling protocols (e.g., subject preparation, sampling conditions) that minimise pre-analytical variation should be adequately described to enable transportability of the data.[35] Numbers of samples taken should be sufficient to deliver the required power to the study.[35,43].
Samples	3.6	Recorded details should include the beginning and end date of the study and timings of sampling. Sampling conditions and sample type should be described in detail. Pre-analytical storage conditions of samples should be described.
Conditions for analysis of samples	3.7	A description of conditions under which the samples were analysed. Analytical protocols should be designed to minimise sources of analytical variation (Optimal Conditions Precision).[76]
Data analysis	4	Data analysis techniques should be described. The power of the study to identify indices of biological variation should be calculated and presented[a].[43]
Outlier analysis	4.1	Outliers should be excluded from the final analysis of the data. Test for outliers should be applied to all levels of data (between replicate analysis, between samples within subject, between subjects).[35] The numbers of outliers and reasons for their exclusion must be given.
Heterogeneity of variance	4.2	Subjects with outlying within subject variance should be rejected from calculations used to determine an estimate of common true variance. The numbers of outliers and reasons for their exclusion must be given[a].
Statistical methods described and appropriate	4.3	Statistical methods used should be appropriately identified, fit for purpose and referenced. Data that do not conform to a normal distribution should be appropriately transformed.[35]
Results	5	Unified terminology[22] should be used and appropriately defined metadata clearly presented to enable understanding and transportation of the data through time and across health care systems.
Terminology	5.1	Terms and symbols should be used to describe biological variation should conform standards identified by Simundic et al.[22]
Results clearly presented and managed	5.2	Biological variation data, with derived indices, should be tabulated in a format that enables extraction of the key data unambiguously associated with a minimum data set to enable transportability of the data. Power of the study and confidence limits around estimates of biological variation should be presented.[43] The results section should clearly identify the results of outlier analysis undertaken and confirm homogeneity of the data sets. If data are stratified the variables used to enable this should be clearly characterised.

[a]Tests to determine the power of a study to identify heteroscedasticity need to be developed. If variances are not homogenous derived estimates of biological variation cannot be trusted, and are not representative for the population in which it is examined.

Section and topic	Item	Evidenced
Discussion	6	The discussion of the data should clearly include a focus on factors that impact on the transportability of the data to other settings. Limitations and strengths of the study should be addressed. If the data are used to set analytical performance specifications, derive reference change values and study individuality, the recommendations of Simundic et al. should be followed.[22]

[a]Tests to determine the power of a study to identify heteroscedasticity need to be developed. If variances are not homogenous derived estimates of biological variation cannot be trusted, and are not representative for the population in which it is examined.

2.4 Summary and conclusions

There are many summary measures of variability across various fields, some are analogous and others expressed in different ways in different areas (such as coefficient of variation, where the standard deviation is divided by the mean value and expressed as a percentage), but are fundamentally expressing the same information to explain variability of tests. Simundic et al[22] called for standardisation of notation for variability measures reported from laboratory studies but this could be extended to all studies of variability as the array of definitions used is a source of confusion, making the results of studies inaccessible to those outside of the field.

The design of studies of test variability is well documented for clinical laboratory based tests with the Fraser and Harris framework.[35] For other areas a paucity of literature concerning the design of variability studies was identified. The framework of Fraser and Harris is lacking with no advice given regarding sample size of variability studies. Røraas et al[43] have developed resources to help plan study sizes.

The methods for evaluating variability studies are again well defined for clinical laboratory tests due to the framework of Fraser and Harris.[35] However, some of the methods advocated require evaluation, specifically the use of outlier detection methods and deletion of values prior

to analysing data as this may impact variability estimates. In other areas the methods of analysis appear to be more ad hoc with estimates of variability, such as ICCs, expressed.

The use of confidence intervals to express uncertainty of an estimate is an issue for biological variability studies due to difficulties with calculation, although it is generally accepted confidence intervals would be useful. Again, the work of Røraas et al[43] provided investigators with help generating confidence intervals. Confidence intervals are commonly used in the medical field and would be reported as standard from ANOVA and multi-level modelling analyses.

In order to understand the quality of biological variability studies, the review of studies will address the areas of design, analysis and reporting with focus on specific elements that have been highlighted as causing concern. These areas are sample size, data transformations, outlier detection and the use of confidence intervals. The review will reveal the depth and the scope of these issues in the current research (see Chapter 3).

To fully investigate the impact of the methods used to investigate biological variability, empirical analyses will be performed to investigate these issues (see Chapter 4).

The issue of sample size specification is also considered further in Chapter 5 and the use of outlier detection methods in Chapter 6.

Chapter 3

A review of biological variability studies: design, analysis, and reporting

This work has been partly presented in the following form:
Sitch, A, Mallett, S, Deeks, J. Biological variability studies: design, analysis and reporting. 4th Methods for Evaluating Medical Tests and Biomarkers (MEMTAB) Symposium, Birmingham, UK. 19-20 July 2016.

Summary

Biological variability studies provide key information for using tests in patient care. These studies are required to provide accurate information.[13] A purposive search was conducted to identify and review the design, methods and reporting of biological variability studies, to identify key issues.

Studies were difficult to find and the majority of those located were for laboratory tests, with few studies of variability for imaging and physiological tests identified. The studies of laboratory tests often used the same method for analysis, ANOVA or random effects linear regression modelling, and outlier detection and deletion, as specified by Fraser and Harris.[35] Sample size was not planned and/or reported in nearly all identified studies and the use of confidence intervals to express uncertainty was rare.

This review identified the need for guidance when planning sample sizes for biological variability studies and calculating confidence intervals, and further investigation into the methods used for outlier detection.

3.1 Introduction

As outlined in Chapter 2, for studies of laboratory tests a general framework for the design and analysis of studies is advocated by Fraser and Harris.[35] Røraas and colleagues have recently provided guidance on sample size and expressing uncertainty of estimates from biological variability studies.[43] Also, Simundic et al[22] have provided a guide for standardising notation to aid understanding and Bartlett et al[69] have developed a checklist for these studies (for further information see Chapter 2). The Working Group on Biological Variation (WG-BV) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) have been active in developing this field further.[13,77]

For studies of physiological and imaging tests, guidance on design was not identified and methods for analyses varied as the types of variability estimated are often different (analytical variability cannot be assessed using some tests and inter-intra reader variability is of interest when studying an imaging test).

The purpose of this review is to understand and evaluate the state of the field by investigating the design, analysis and reporting. This review will allow investigation into how well studies adhere to the recommendations,[69] where there is variation in methods used, and where there is clear need for improvement.

3.2 Aims and objectives

The aim of this review was to evaluate the design, analysis and reporting of biological variability studies. The study objectives were to identify:

- the current state of the field (aims of studies, test types and disease areas);
- how studies are designed;
- the methods used to analyse these studies;
- the quality of reporting;
- and, differences between studies assessing laboratory, imaging and physiological tests.

3.3 Methods

Several searches were used to identify published articles reporting test variability studies (between November 2014 and May 2015). These searches were developed purposely to ensure different test types (laboratory based tests, imaging tests and physiological tests) were included and the studies covered a range of clinical areas. The searches were not intended to be fully comprehensive, but were designed to provide a representative sample of studies. The identified studies were assessed for suitability in the review and key information was extracted.

3.3.1 Searches used to identify studies reporting biological variability studies

The search strategy was developed adapting to the identified studies, to ensure a sufficient number of studies were identified and these were representative. The first search was a broad search; hand searching of specific journals was then used to ensure full coverage of these

sources; the Westgard data base was utilised; and, specific searches in targeted clinical areas were used.

3.3.1.1 Search A: Broad search–November 2014

- Key word search (bio* AND vari*) in title and/or abstract for the period 1st November 2013 to 31st October 2014.

The initial search was used to identify a broad range of studies. Using the results of this search, searches were refined to identify appropriate studies of biological variability. Searches were performed in PubMed in November 2014.

3.3.1.2 Search B: Hand search–January 2015

- All articles published in the journals Clinical Biochemistry (search B1), Radiology (search B2) and Clinical Chemistry (search B3) during the period 1st January 2014 to 31st December 2014. Searches were performed in PubMed in January 2015.

This search looked at all articles for specific journals for a one year period. Journals were targeted, informed by results from search A, and were chosen to include laboratory and imaging studies.

3.3.1.3 Search C: Test specific searches–May 2015

Detailed searches for three different test types (imaging, laboratory and physiological) in specific clinical areas:

- Search C1: Ultrasound imaging to assess bladder wall thickness in patients with incontinence (exp Urinary Incontinence/ AND exp Ultrasonography).

- Search C2: Creatinine and Cystatin C measurements to estimate glomerular filtration rate (GFR) in patients with chronic kidney disease (CKD) (exp Renal Insufficiency,

Chronic/ AND exp Glomerular Filtration Rate/ AND (exp Cystatin C/ OR exp Creatinine/).

- Search C3: Spirometry to measure forced expiratory volume (FEV) in patients with chronic obstructive pulmonary disorder (COPD) (exp Pulmonary Disease, Chronic Obstructive/ AND exp Spirometry/ AND exp Forced Expiratory Volume/).

These particular tests and populations were considered due to ongoing work in these areas, enabling knowledge of the literature and access to experts to develop the searches. Searches were performed in PubMed in May 2015 with no restriction on the date of publication.

3.3.1.4 Search D: Westgard QC database–May 2015

Studies recorded in the Westgard QC database[78] (an online resource giving biological variability information for laboratory based tests collated from published studies[18,79,80]) published from 1st January 2000 onwards. Accessed in 2015.

As studies have already been assessed prior to inclusion in the database and this is specifically for laboratory tests, identifying studies only from this source would be limited.

3.3.1.5 Search E: Expert search

In addition to the studies identified by these searches, published articles identified by previous and concurrent work meeting the criteria were included to enrich the sample. The test specific searches (C1, C2 and C3) described did not manage to identify all known biological variability studies in the specific clinical areas. Clinical experts provided further studies.

3.3.2 Eligibility criteria

Studies were included in the review if they met the following criteria:

- Study aim: the purpose (primary or secondary) was to assess the variability of measuring or evaluating test results in participants, in a stable health state over the period of testing.

- Study test(s): include assessments of the test(s) for participants under the same conditions.

- Study participants: there was no restriction placed on study participants, with studies of healthy participants and participants with disease included. However, it was required that assessment was of participants or participant samples rather than control/spiked samples.

- Language: only studies reported in English were included in the review.

Studies were required to have multiple test assessment which included repetition of part or all of the testing procedure; this could mean participants were tested multiple times or test output/samples were assessed multiple times (for example, clinicians assessing imaging results or retesting of stored samples). To obtain an estimate of biological variability (*"The natural variability in a lab parameter due to physiologic differences among subjects and within the same subject over time."*[15]), analytical variability needs to be assessed and allowed for when measuring within-individual and between-individual variability so studies assessing analytical variability only were also included.

3.3.3 Review of selected studies and data extraction

Titles and abstracts were assessed for inclusion, full text was obtained for potentially eligible articles and reviewed against the eligibility criteria. Details of the study design, methods of analysis, and reporting were extracted and assessed by a single reviewer. The data extraction form was adapted from Bartlett's checklist,[69] including fields appropriate for studies of physiological and imaging tests. Additional items were included for aspects of study design, analysis and reporting.

The following information was extracted from the included studies:

1. What are the aims of these studies and in which tests and test areas are they seen?
    (a) Test type
    (b) Study aim
    (c) Situations assessed
2. How are studies assessing biological variability of tests designed?
    (a) Participants included
    (b) Sample size (any justification of the number of participants, observations per participant and assessment of observations)
    (c) Duration of study and time between repeated assessments
    (d) Levels of variability assessed
    (e) Reduction of pre-analytical variability
    (f) Blinding
3. How are studies assessing biological variability of tests analysed?
    (a) Methods for analyses (reported explicitly or judgement from results)
    (b) Data transformation
    (c) Outlier identification and exclusion
4. How are studies assessing biological variability of tests reported?
    (a) Title identifies study as biological variability
    (b) Clarity of reporting for study design and methods for analysis
    (c) Biological variability estimates and the corresponding uncertainty

3.4 Results

3.4.1 What is the current state of the field? What are the aims of these studies and in which tests and test areas are they seen?

3.4.1.1 Studies identified

One-hundred-and-one studies were included in the review, see Figure 3.1. The Westgard QC database contributed more studies to the review (n=57, 56%) than any other search, see Table 3.1. Of the 101 studies, 75 (74%) were studies of clinical chemistry laboratory tests, with 20 (20%) studies of imaging tests and 6 (6%) studies of physiological tests. The tests evaluated in these studies varied, (including imaging to measure bladder wall thickness and tumour size; spirometry to measure forced expiratory volume (FEV); and laboratory tests measuring glomerular filtration rate (GFR), vitamin uptake, HbA1c, hepatic enzymes, hormone levels and cardiac troponin), see Appendix A.

3.4.1.2 Study aims

Extracting the study aim gives insight into the main purpose of each study. Some studies evaluated the variability of just one test (n=37, 37%) whereas others looked at multiple tests, or multiple measurement types from a single test (n=64, 63%). Some studies (n=27, 27%) also evaluated tests in multiple populations (or subpopulations, defined by gender, age, medication status etc.) or over different time ranges (n=11, 11%) and reported the results separately for these situations. The aims of these studies varied also with some studies looking at the variability of a test measure secondarily to evaluating the test for a different property (n=25, 25%). The median number of testing situations (different measurement type, population or time point) was 4 (Q1, Q3: 2, 7). One study estimated variability in 53 different test situations,[82] see Box 3.1 for examples and Table 3.2.

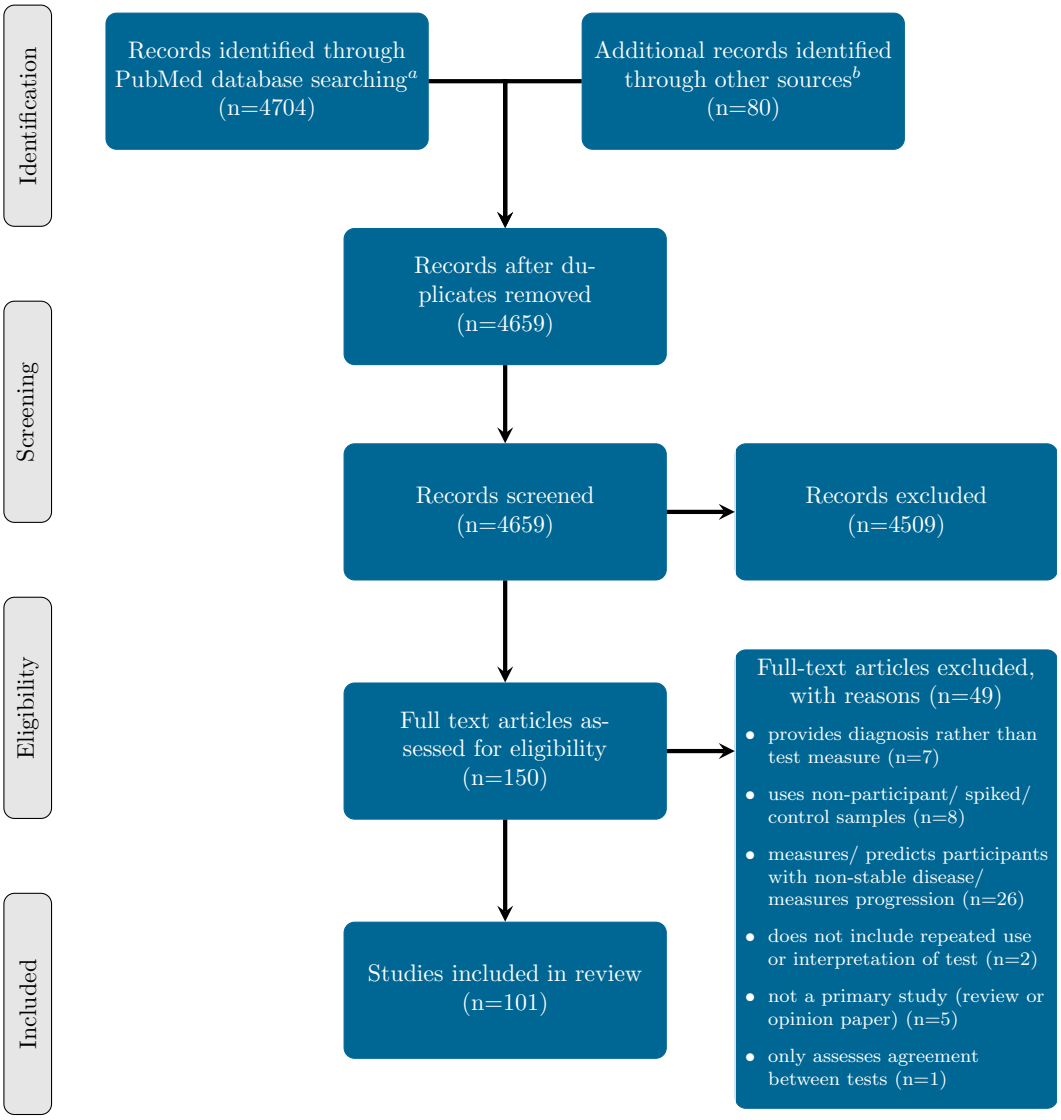


Figure 3.1: Flowchart studies included in biological variability review, from Moher[81]. [a]search A 1024 studies; B1 318 studies; B2 500 studies; B3 313 studies; C1 219 studies; C2 1578 studies; C3 752 studies. [b]search D 65 studies; search E 15 studies (see §3.3.1).

Table 3.1: Identified biological variability studies by search.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Search*; n(%)				
A	1 (5)	4 (5)	0 (0)	5 (5)
B1	0 (0)	2 (3)	0 (0)	2 (2)
B2	8 (40)	0 (0)	0 (0)	8 (8)
B3	0 (0)	5 (7)	0 (0)	5 (5)
C1	6 (30)	0 (0)	0 (0)	6 (6)
C2	0 (0)	5 (7)	0 (0)	5 (5)
C3	0 (0)	0 (0)	6 (100)	6 (6)
D	0 (0)	57 (76)	0 (0)	57 (56)
E	5 (25)	6 (8)	0 (0)	11 (11)
Year				
Median	2012	2008	2006	2009
Q1, Q3	(2008, 2014)	(2003, 2012)	(2003, 2007)	(2003, 2013)
Range	[1994-2014]	[1988-2014]	[1996-2013]	[1988-2014]
Journal; n(%)				
Annals of Clinical Biochemistry	0 (0)	6 (8)	0 (0)	6 (6)
Chest	0 (0)	0 (0)	2 (33)	2 (2)
Clinical Biochemistry	0 (0)	3 (4)	0 (0)	3 (3)
Clinical Chemistry	0 (0)	24 (32)	0 (0)	24 (24)
Clinical Chemistry and Laboratory Medicine	0 (0)	21 (28)	0 (0)	21 (21)
Clinica Chimica Acta	0 (0)	2 (3)	0 (0)	2 (2)
European Journal of Obstetrics, Gynaecology, & Reproductive Biology	2 (10)	0 (0)	0 (0)	2 (2)
Kidney International	0 (0)	2 (3)	0 (0)	2 (2)
Radiology	8 (40)	0 (0)	0 (0)	8 (8)
Ultrasound Obstetrics Gynaecology	2 (10)	0 (0)	0 (0)	2 (2)
World Journal of Urology	2 (10)	0 (0)	0 (0)	2 (2)
Other	6 (30)	17 (23)	4 (67)	27 (27)

*Four studies were identified by more than 1 search. See §3.3.1 for detail of searches.

Differences between test types

Multiple tests were commonly evaluated across studies of each of the test types (imaging n= 14/20 (70%); laboratory n= 45/70 (60%); and physiological n=4/6 (67%)). More studies of laboratory tests evaluated multiple populations (n=25/75, 33%) than for imaging (n=1/20, 5%) and physiological tests (n=1/6, 17%). Across all test types, studies showed analysis of tests separate to estimating variability (imaging n= 9/20 (45%); laboratory n= 15/70 (20%); and physiological n=1/6 (17%)).

The purpose of 16 (16%) studies was to assess inter and/or intra reader variability, these studies were all studies of imaging tests, see Table 3.2. Some of these studies investigated reassessment of images (for example a study of pelvic floor muscle contraction by ultrasonography where 'frozen images' were reassessed)[83] but in other imaging studies multiple images were taken and assessed (for example a study investigating ultrasonography and Doppler velocimetric assessment of the levator ani muscle used two investigators each performing the imaging procedures)[84].

Box 3.1: Multiple testing situations in identified biological variability studies.

Multiple tests and populations
Pediatric within-day biological variation and quality specifications for 38 biochemical markers in the CALIPER Cohort, by Bailey et al[82] looked at 53 different testing situations. This included 38 biochemical markers, for example Albumin, Glucose, and Magnesium. Some of the biochemical markers were assessed separately for children in different age bands, for example AST was assessed separately for children up to seven years old, from seven years up to 12 years old and then from 12 years to 19 years old.

Multiple tests and time periods
Weekly and 90-minute biological variations in cardiac troponin T and cardiac troponin I in hemodialysis patients and healthy controls, by Aakre et al[85] evaluated Cardiac troponin T and Cardiac Troponin I for two distinct time ranges. The variability of the tests was assessed by evaluating patients at 90 minute intervals over a 6 hour testing period and weekly for a period of 10 weeks.

Table 3.2: Aims of identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Single test evaluated; n(%)	6 (30)	30 (40)	2 (33)	37 (37)
Multiple tests evaluated; n(%)	14 (70)	45 (60)	4 (67)	64 (63)
Number of tests evaluated				
Median (Q1, Q3)	4 (1, 6)	2 (1, 5)	5 (1, 11)	2 (1, 6)
Range	[1-19]	[1-38]	[1-14]	[1-38]
Multiple populations evaluated; n(%)	1 (5)	25 (33)	1 (17)	27 (27)
Multiple time periods evaluated; n(%)	1 (5)	8 (11)	2 (33)	11 (11)
Number of situations evaluated				
Median (Q1, Q3)	5 (1, 8)	4 (2, 7)	8 (2, 14)	4 (2, 7)
Range	[1-19]	[1-53]	[1-30]	[1-53]
Inter/intra rater variability; n(%)	16 (80)	0 (0)	0 (0)	16 (16)
Test(s) evaluated separately to variability; n(%)	9 (45)	15 (20)	1 (17)	25 (25)

3.4.2 How are studies assessing biological variability of tests designed?

3.4.2.1 Participants included

Biological variability estimates are required for the population that will receive the test. Variability estimates in healthy participants may be useful if using a test to screen or developing test reference ranges; however estimates are required for diseased participants if a test is used for monitoring progressive or recurrent disease.

The participants included in many of the studies were healthy (n=48, 48%). A further 21 (21%) studies assessed mixed populations (with some participants healthy and some diseased), and 22 (22%) studies tested only diseased participants. The healthy status of these populations was rarely confirmed (n=3, 3%), for example in a study assessing high-sensitivity troponin T[62] healthy individuals were verified through physical examination, MRI analysis including adenosine perfusion or dobutamine stress, lung function testing, and blood sample testing, see Table 3.3.

Differences between test types

Assessment of healthy participants was more common in studies of laboratory tests (n=43/75, 57%) than imaging (n=4/20, 20%) and physiological tests (n=1/6 (17%)). Use of a mixed population of healthy and diseased participants was seen across studies of all test types (imaging n= 6/20 (30%); laboratory n= 14/70 (19%); and physiological n=2/6 (33%)), see Table 3.3.

Table 3.3: Populations studied in identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Population; n(%)				
Healthy population	4 (20)	43 (57)	1 (17)	48 (48)
Confirmed healthy population	0 (0)	3 (4)	0 (0)	3 (3)
Diseased population	6 (30)	12 (16)	3 (50)	21 (21)
Mixed healthy/diseased population	6 (30)	14 (19)	2 (33)	22 (22)
Unknown population	4 (20)	6 (8)	0 (0)	10 (10)

3.4.2.2 Sample size

Study sample sizes require adequate consideration to ensure estimates produced are meaningful. Sample size calculations were rarely performed in the identified biological variability studies (n=1, 1%), with studies routinely omitting justification of the number of participants included. The single study reporting a sample size justification used a previous study estimate of variability and calculated the number of repeated measures required for each subject to allow for small differences to be detected.[86] The smallest studies had four participants[87,88] and the largest had 7,101 participants.[89] Many of the studies identified had few participants (see Figure 3.2); the median number of participants was 24 (Q1, Q3: 15, 40). Of the 99 studies reporting a sample size, eight (8%) had more than 100 participants and two (2%) had more than 1,000. Fourteen (14%) studies had 10 participants or fewer; 27 (27%) studies had between 11 and 20 participants; and, 24 (24%) studies had between 21 and 30 participants. Studies with larger sample sizes utilised routinely collected data rather than following a prospective plan for data collection, see Box 3.2.

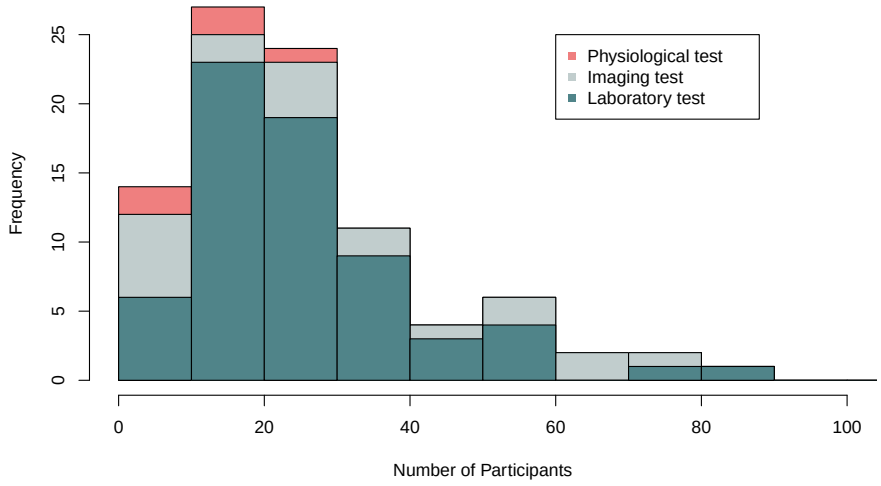


Figure 3.2: Histogram of study sample sizes in identified biological variability studies (excluding the 8 studies with sample size exceeding 100).

Many studies also looked at subgroups of participants or of the tests performed, and the sample sizes for these analyses were smaller with some studies using data from only two participants to provide estimates of variability (median=11; Q1, Q3: 7, 15). The median number of observations (calculated for studies where there are multiple measures taken for each person and the number of measures was reported) obtained from each participant to allow an estimate of within-individual variability to be made was 5 (Q1, Q3: 3, 10) and ranged from 2 to 40. The median number of assessments was 2 (Q1, Q3: 2, 3). For inter-intra reader studies the median number of readers was 2 (Q1, Q3: 2, 2) and median number of duplicate reads was 2 (Q1, Q3: 1, 2), with the common design using two readers for each image and one reader duplicating all readings, see Table 3.4.

Differences between test types

The median total sample size was similar across studies of all test types (imaging n=26; laboratory n=24; and physiological n=16), and for the number of repeats for each participant (imaging n=2; laboratory n=5; and physiological n=5). The number of analyses for each measure is only given for laboratory tests, as for physiological and imaging tests it is not possible to assess variability at this level. All inter-intra reader studies were of imaging tests,

see Table 3.4.

Box 3.2: Biological variability studies with large sample sizes–utilising existing data collection.

Variability of spirometry in chronic obstructive pulmonary disease: results from two clinical trials, by Herpel et al[89] used data collected from two clinical trials (the National Emphysema Treatment Trial, NETT, and the Lung Health study, LHS). This study had a large sample size of 7,101. Trial participants had two spirometry measurements taken during baseline study investigations and the data from this was analysed to assess variability of spirometry in measuring FEV (forced expiratory volume) and FVC (forced vital capacity) in this study.

Use of observed within-person variation of cardiac troponin in emergency department patients for determination of biological variation and percentage and absolute reference change values, by Simpson et al[90] used repeat blood sample assessments of Cardiac Troponin (hs-cTnI) taken routinely for patients presenting in the emergency department evaluated for acute coronary syndrome. The analysis of these data to estimate variability was for the 283 individuals with two test results that did not have acute cardiac disease at the time of testing in the emergency department.

3.4.2.3 Study duration and time between repeats

The duration of studies and the timing of repeats indicate the resources required for these studies and the typical design. The median total study duration was 5 weeks (Q1, Q3: 2, 14), ranging from just a single day to five years. The median time between test repeats for the studies was 1 week (Q1, Q3: 1, 4), with the most frequent repeats within a single day and the least frequent six monthly, see Table 3.4.

Differences between test types

Studies of imaging tests were carried out over a shorter testing period (median of 0.5 weeks) compared to laboratory and physiological tests (median of 7 weeks and 5 weeks respectively). For the eight imaging studies where the time between repeats was reported the interval (median of 1.5 weeks) was similar to using a laboratory or physiological test (median of 1 week for both test types), see Table 3.4.

Table 3.4: Sample size and study duration in identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Sample size				
Total sample size calculation/justification provided; n(%)	1 (5)	0 (0)	0 (0)	1 (1)
Total study sample size	(n=20)	(n=73)	(n=6)	(n=99)
Median (Q1, Q3)	26 (10, 52)	24 (4, 39)	16 (10, 30)	24 (15, 40)
Range	[7-77]	[4-1103]	[9-7101]	[4-7101]
Multiple measures per person	(n=7)	(n=66)	(n=5)	(n=77)
Median (Q1, Q3)	2 (2, 2)	5 (4, 10)	5 (3, 21)	5 (3, 10)
Range	[2-6]	[2-23]	[2-40]	[2-40]
Total number of analyses for each measure	-	(n=31)	-	(n=31)
Median (Q1, Q3)		2 (2, 3)		2 (2, 3)
Range		[2-4]		[2-4]
Subgroup study sample size	(n=2)	(n=31)	(n=2)	(n=35)
Median (Q1, Q3)	13 (7, 19)	12 (6, 15)	10 (10, 10)	11 (7, 15)
Range	[7-19]	[2-118]	[10-10]	[2-118]
Subgroup multiple measures per person	(n=1)	(n=29)	(n=1)	(n=31)
Median (Q1, Q3)	2 (2, 2)	5 (4, 6)	3 (3, 3)	5 (4, 6)
Range	[2-2]	[2-11]	[3-3]	[2-11]
Subgroup number of analyses for each measure	-	(n=14)	-	(n=14)
Median (Q1, Q3)		2 (2, 2)		2 (2, 2)
Range		[2-3]		[2-3]
Number of readers	(n=16)	-	-	(n=16)
Median (Q1, Q3)	2 (2, 2)			2 (2, 2)
Range	[1-3]			[1-3]
Number of duplicate reads	(n=15)	-	-	(n=15)
Median (Q1, Q3)	2 (1, 2)			2 (1, 2)
Range	[1-3]			[1-3]
Study duration				
Total duration (weeks)	(n=12)	(n=69)	(n=3)	(n=84)
Median (Q1, Q3)	0.5 (0.1, 2)	7 (3, 20)	5 (1, 8)	5 (2, 14)
Range	[0.1-4]	[0.1-260]	[1-8]	[0.1-260]
Overall time between repeats (weeks)	(n=8)	(n=64)	(n=2)	(n=74)
Median (Q1, Q3)	1.5 (0.5, 3)	1 (1, 4)	1 (0.1, 2)	1 (1, 4)
Range	[0.1-4]	[0.1-26]	[0-2]	[0.1-26]

3.4.2.4 Variability assessed

The design of studies allows assessment of different types of variability, this information suggests the estimates researchers focus on. Most studies provided estimates of between-individual variation (n=60, 59%) and within-individual variation (n=72, 71%). For laboratory tests the estimated variability between and within individuals was also accompanied by the assessment of analytical/reassessment variability in 35 (35%) studies. Sixteen (16%) studies (all of imaging tests) explored within and between reader variability.

For laboratory based studies an estimate of analytical variability is required to enable the calculation of between and within-individual variation. For studies not directly estimating the analytical variation, an estimate of analytical/reassessment variation obtained from a source external to the study (usually another study or published work) or from the analysis of control samples, is used to calculate the other levels of variability, this was reported in 24 (24%) studies and a further 14 (14%) studies did not explicitly report this but it is suspected, see Box 3.3. In addition to analytical, within-individual and between-individual variability, few studies also attempted to assess the variability at other levels, such as between centre variability, see Table 3.5.

Differences between test types

For studies of laboratory tests the variability at the analytical, within-individual and between-individual levels were commonly considered. Of the studies evaluating physiological tests, one (17%) evaluated variability at the within-individual level. Studies of imaging tests were mainly inter-intra reader studies and investigated the variability between and/or within readers, see Table 3.5.

3.4.2.5 Pre-analytical variability reduced prior to assessment

Pre-analytical variability was not a focus of this review but there is an assumption this is minimal; many studies (n=85, 84%) reported having undertaken measures to reduce pre-analytical variability. Most laboratory based studies reported using the same staff, instruments and con-

Box 3.3: Biological variability studies: Analysis where CV_A is estimated separately to main variability study.

Controls
Biological variation of seminal parameters in healthy subjects, by Alvarez et al,[91] looked at the variability of seminal parameters in 20 healthy donors. However, to estimate analytical variation quality control materials were assessed.

Sub-samples
Weekly and 90-minute biological variations in cardiac troponin T and cardiac troponin I in hemodialysis patients and healthy controls, by Aakre et al,[85] assessed the variability of cardiac troponin T and cardiac troponin I for 19 heamodialysis patients and 20 healthy controls. When assessing analytical variation, however, only half of the samples were selected to be analysed in duplicate in order for this to be estimated, with random selection stratified by sex.

External estimate
Intra-individual variation in creatinine and cystatin C, by Bandaranayake et al,[92] assessed the variability of creatinine and cystatin C in 10 healthy participants. Estimates of CV_A obtained from external sources were stated and used to calculate other variability estimates.

Analytical CV measured in study but external estimate used
Biological variation of myeloperoxidase, by Dednam et al,[93] investigated 12 healthy individuals to understand the biological variability of myeloperoxidase, a marker of coronary artery disease. Using samples for the 12 participants an estimate of CV_A was produced of 4%. The authors state the estimate was 'unrealistically low' and an estimate of 8.4% obtained from an external source was used in all calculations.

ditions for sample collection in addition to storage of samples and transport. For other types of studies, there was also reason to believe pre-analytical variation had been minimised (at various points) by ensuring participants were prepared for the test in a consistent way and using the same staff to carry out the test. Studies where measurements were not taken in controlled circumstances were unable to demonstrate minimising pre-analytical variation (for example, one study estimated biological variability using test measures obtained in intensive care unit patients)[94].

Differences between test types

Pre-analytical variability was often minimised across studies of all test types (imaging n=15/20 (75%); laboratory n=65/75 (87%); and physiological n=5/6 (83%)), see Table 3.5.

Table 3.5: Variability levels assessed in identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Variability level assessed; n(%)				
Between individuals	0 (0)	60 (80)	0 (0)	60 (59)
Within individuals	1 (5)	70 (93)	1 (17)	72 (71)
Analytical/reassessment	1 (5)	34 (45)	0 (0)	35 (35)
Analytical/reassessment evaluation external to study	0 (0)	24 (32)	0 (0)	24 (24)
Suspected Analytical/reassessment evaluation external to study	0 (0)	14 (19)	0 (0)	14 (14)
Within/between readers	16 (80)	0 (0)	0 (0)	16 (80)
Pre-analytical variation minimised	15 (75)	65 (87)	5 (83)	85 (84)

3.4.2.6 Blinding

Blinding is important when measuring quantities multiple times, especially for inter-intra reader studies. Only seven (7%) studies reported blinding of some type. Studies achieved blinding by keeping assessors and participants from knowing previously observed measures. Although blinding is not explicitly reported in the majority of studies, it may be in many studies, the design (with all samples being tested in a single batch at a later date) means the analysis of samples was blinded.

Differences between test types

The majority of the studies (6/7, 86%) reporting blinding were inter-intra variability studies using imaging tests. Blinding is integral to inter-intra variability studies as these studies assess variability between and within readers and so information on the interpretation of other readers and previous reads by the same reader would be detrimental to the study. For studies of laboratory (n=0/75, 0%) and physiological tests (n=1/6, 17%) blinding was often not reported and may not be as critical.

3.4.3 How are studies assessing biological variability of tests analysed?

3.4.3.1 Methods for analysis

The primary method for analysis was extracted to understand the approaches used. Eighty-eight (87%) studies appeared to use ANOVA or random effects modelling for the primary analyses; for 40 studies (40%) the method was not explicitly expressed but the results suggested ANOVA or random effects modelling. The framework introduced by Fraser and Harris[35] was referred to in the methods section by 32 (32%) studies and referenced by a total of 52 (51%) studies. Few studies used alternative methods, such as assessing change in a test value over time or comparing methods of measurement, with the alternative methods Bland-Altman (n=7, 7%), Kappa (n=1, 1%), other modelling (n=1, 1%) and other methods (n=10, 10%). Other methods were Chi-squared tests, t-tests, correlation coefficients, Kruskal-Wallis tests and Mann-Whitney U tests, see Table 3.6.

Differences between test types

Across the test types, the use of ANOVA or random effects modelling was common; the issue of assumed ANOVA or random effects modelling due to lack of clarity was common across the test types also. All studies identified as following and referencing the Fraser-Harris framework were evaluating laboratory tests. The use of Bland-Altman methods was more common in studies of imaging (n=4/20, 20%) and physiological tests (n=2/6, 33%), rather than laboratory tests (n=1/75, 1%). Kappa was used as the primary analysis for one (5%) study of an imaging test, see Table 3.6.

3.4.3.2 Normality checking and data transformation

As data transformations may impact the results obtained, the methods for normality checking were noted to understand how frequently this is performed, the methods used and the transformation approach taken. The normality of obtained data was tested in 38 (38%) studies, the methods used were: the Shapiro-Wilk test (n=12, 12%), Kolmogorov-Smirnov

test (n=10, 10%) or visual inspection (n=4, 4%), others stated assessing normality but did not specify a method (n=18, 18%). Twenty-two (22%) studies reported log transforming the data, with no alternative transformations reported. Of the 22 studies that log transformed the data, 17 (77%) reported testing for normality, see Table 3.6.

Differences between test types

Testing for normality was often seen in the laboratory based studies and was not as common in the non-laboratory test studies (laboratory n=34/75 (45%); imaging n=3/20 (15%); and physiological n=1/6 (17%). For the laboratory based studies, 21 studies (28%) reported log-transforming the data, whereas only one (5%) imaging study and none of the identified (0%) physiological test studies reported log-transformation of data, see Table 3.6.

3.4.3.3 Outlier detection and removal

As outlier detection and removal may impact variability estimates, details of outlier detection and removal processes were identified to understand the frequency of use and the procedures. Outliers were tested for in many of the studies identified (n=25, 25%), all were laboratory based studies. The methods for detecting outliers were mainly Cochran's C test and Reed's criterion (n=9, 9% and n=5, 5%), although in some studies the methods used were not specified (n=6, 6%). Other outlier detection methods reported were ±3SD, Dixon's Q test, Grubbs' test, Tukey's IQR rule and visual inspection. Outliers were reported to have been excluded in 27 (27%) studies, see Table 3.6. Of the 27 studies with outliers excluded, 4 studies did not report testing for outliers this could be due to 'error' data that was removed without formal testing or these studies may have neglected to report the method used for testing only reporting the consequential changes to the data.

Differences between test types

Testing for normality was often seen in the laboratory based studies; however, this process was not reported in the studies of imaging and physiological tests (laboratory n=25/75 (33%); imaging n=0/20 (0%); and physiological n=0/6 (0%). For the studies of laboratory tests,

27 (36%) reported removing outliers, whereas no studies of imaging and physiological tests reported outlier removal, see Table 3.6.

3.4.4 How are studies assessing biological variability of tests reported?

3.4.4.1 Title identifies study as biological variability

Following from the items in the Bartlett checklist,[69] the titles of studies were assessed to identify if they clearly labelled studies as biological variability studies. Of the studies identified by the search, 67 (66%) studies were clearly studies of biological variation from the title of the article, see Table 3.7.

Differences between test types

Studies of laboratory tests were more likely to clearly identify as studies of biological variability from the title (n=61/75, 81%) compared to studies of imaging (n=5/20, 25%) and physiological tests (n=1/6, 17%). This is likely due to the uniform terminology used in studies of laboratory tests which is not present for studies of imaging and physiological tests, see Table 3.7.

3.4.4.2 Clarity of design and methods

The clarity of reporting of the design and methods in the identified studies was variable. Studies did not adequately describe the: population (n=10, 10%), duration of study (n=17, 17%), method of measuring assessment variability (analytical variability) externally to the study or from control/spiked samples (n=14, 14%), method for analyses (n=40, 40%), normality testing procedure (n=31, 31%), outlier detection procedure (n=10, 10%), justification/rationale for the sample size (n=100, 99%), number of repeats (n=39, 39%), timing of repeats (n=42, 42%), and number of assessments duplicated (n=4, 4%), see Table 3.7.

Table 3.6: Analysis methods of studies in identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Primary analyses methods; n(%)				
ANOVA reported	2 (10)	33 (44)	2 (33)	37 (37)
RE reported	5 (25)	7 (9)	1 (17)	13 (13)
ANOVA/ RE assumed	6 (30)	33 (44)	1 (17)	40 (40)
Total ANOVA/RE	12 (60)	73 (97)	3 (50)	88 (87)
Bland-Altman	4 (20)	1 (1)	2 (33)	7 (7)
Kappa	1 (5)	0 (0)	0 (0)	1 (1)
ROC analysis	0 (0)	0 (0)	0 (0)	0 (0)
Other modelling	0 (0)	1 (1)	0 (0)	1 (1)
Other methods	6 (30)	2 (3)	2 (33)	10 (10)
Secondary analyses methods; n(%)				
Bland-Altman	1 (5)	5 (7)	0 (0)	6 (6)
Kappa	0 (0)	0 (0)	0 (0)	0 (0)
ROC analysis	3 (15)	2 (3)	0 (0)	5 (5)
Other modelling	4 (20)	15 (20)	1 (17)	20 (20)
Other methods	6 (30)	41 (55)	0 (0)	47 (47)
Fraser framework for analyses; n(%)				
Methods of analyses followed	0 (0)	52 (69)	0 (0)	52 (52)
Reference to framework	0 (0)	32 (43)	0 (0)	32 (32)
Transformation of data; n(%)				
Assessment of normality	3 (15)	34 (45)	1 (17)	38 (38)
Shapiro-Wilks test	2 (10)	9 (12)	1 (17)	12 (12)
Kolmogorov-Smirnov test	2 (10)	8 (11)	0 (0)	10 (10)
Visual/plot of data assessment	0 (5)	3 (4)	0 (0)	4 (4)
Unclear method	0 (0)	18 (24)	0 (0)	18 (18)
Data transformed	1 (5)	21 (28)	0 (0)	22 (22)
Log transformation	1 (5)	21 (28)	0 (0)	22 (22)
Other transformation	0 (0)	0 (0)	0 (0)	0 (0)
Unclear transformation	0 (0)	0 (0)	0 (0)	0 (0)
Outlier detection; n(%)				
Outliers tested	0 (0)	25 (33)	0 (0)	25 (25)
Cochran C test	0 (0)	9 (12)	0 (0)	9 (9)
Reed's test	0 (0)	5 (7)	0 (0)	5 (5)
Other outlier test	0 (0)	13 (17)	0 (0)	13 (13)
Unclear method	0 (0)	6 (8)	0 (0)	6 (6)
Outliers removed	0 (0)	27 (36)	0 (0)	27 (27)

Differences between test types

The population was not specified in more studies of imaging tests (n=4/20, 20%) compared to laboratory and physiological test studies. The issue of not reporting sample size justification was seen across all test types with only one imaging study reporting details of sample size. Details of study design (number of repeats, timing of repeats, duration of study and methods of analysis) were insufficiently reported in similar percentages of studies across the test types, see Table 3.7. The issue of measuring analytical variability outside of the study was unique to studies of laboratory tests, as was the lack of clarity when explaining the methods for testing for normality and outlier detection.

3.4.4.3 Biological variability estimates and uncertainty

The most common estimates reported were coefficients of variation (CV): estimates of assessment (analytical) variability (n=35, 35%), within-individual variability (n=72, 71%) and between-individual variability (n=60, 60%). Some studies reported analytical (2, 2%), within-individual (6, 6%) and between-individual variability (5, 5%) as standard deviations. Total variability was not often reported in these studies, along with total error and total imprecision. Four (4%) studies reported an exact CV after using log transformed data (methods for exact geometric CV after log transformation,[59,60] assuming the distribution was log-normal, were used). The RCV (or repeatability coefficient) was reported for 48 (48%) studies, 42 (42%) studies reported a symmetric RCV and 7 (7%) reported a non-symmetric RCV interval (one study reported both,[62] see Box 3.4). Forty-four (44%) studies reported the index of individuality (II) and 18 (18%) reported an ICC/reliability parameter. Estimates of percentage agreement, Kappa, AUROC, Bland-Altman limits, regression coefficients and other general estimates were seen in some studies, and were mainly produced for secondary aims or aims unrelated to assessing test variability, see Table 3.7.

The uncertainty around these estimates provided was rarely described with few studies (n=18, 18%) providing confidence intervals for any of the biological variability measures or describing the uncertainty in any other way (n=2, 2%), comparison to the reference range and caution when interpreting the results. Some of the studies reviewed were considered to have reported

Table 3.7: Reporting of identified biological variability studies.

	Test type			
	Imaging	**Laboratory**	**Physiological**	**All**
N	20	75	6	101
Identification; n(%)				
Biological variation study clear from title	5 (25)	61 (81)	1 (17)	67 (66)
Poor clarity (unclear or insufficient detail); n(%)				
Population	4 (20)	6 (8)	0 (0)	10 (10)
Sample size	19 (95)	75 (100)	6 (100)	100 (99)
Number of participants	0 (0)	2 (3)	0 (0)	2 (2)
Number of repeats	5 (25)	30 (40)	4 (67)	39 (39)
Timing of repeats	14 (70)	25 (33)	3 (50)	42 (42)
Number of assessments duplicated	1 (5)	3 (4)	0 (0)	4 (4)
Duration of study	8 (40)	6 (8)	3 (50)	17 (17)
Variability of measure of assessment external to study	0 (0)	14 (19)	0 (0)	14 (14)
Methods for analyses	6 (30)	33 (44)	1 (17)	40 (40)
Normality procedure	0 (0)	31 (41)	0 (0)	31 (31)
Outlier procedure	0 (0)	10 (13)	0 (0)	10 (10)
Estimates; n(%)				
Assessment (analytical) variability CV	1 (5)	34 (45)	0 (0)	35 (35)
Within-individual variability CV	1 (5)	70 (93)	1 (17)	72 (71)
Between-individual variability CV	0 (0)	60 (80)	0 (0)	60 (59)
Assessment (analytical) variability SD	0 (0)	2 (3)	0 (0)	2 (2)
Within-individual variability SD	0 (0)	5 (7)	1 (17)	6 (6)
Between-individual variability SD	0 (0)	5 (7)	0 (0)	5 (5)
Exact CV	0 (0)	4 (5)	0 (0)	4 (4)
RCV/repeatability coefficient	0 (0)	48 (64)	0 (0)	48 (48)
Symmetric RCV	0 (0)	42 (56)	0 (0)	42 (42)
Non-symmetric RCV	0 (0)	7 (9)	0 (0)	7 (7)
II	0 (0)	44 (59)	0 (0)	44 (44)
ICC/reliability parameter	10 (50)	6 (8)	2 (33)	1 (18)
Percentage agreement	0 (0)	0 (0)	1 (17)	1 (1)
Kappa	1 (5)	0 (0)	0 (0)	1 (1)
AUROC	1 (5)	1 (1)	0 (0)	2 (2)
Bland-Altman limits	3 (15)	2 (3)	2 (17)	7 (7)
Uncertainty; n(%)				
Confidence intervals	12 (60)	4 (5)	2 (33)	18 (18)
Other measure of uncertainty	0 (0)	2 (3)	0 (0)	2 (2)

Box 3.4: Biological variability studies: symmetric and non-symmetric RCVs.

Biological variation and reference change value of high-sensitivity troponin T in Healthy Individuals during short and intermediate follow up periods, by Frankenstein et al,[62] presented both symmetric and non-symmetric reference change values, for normal and log-normal data respectively. For the estimate of hourly hsTnT using the E 170 assay the RCV% for normal data was ±47 and for log-normal data was 64, -39.

the results well and are shown as exemplars, see Box 3.5.

Differences between test types

Estimates of CV were almost exclusively reported in studies of laboratory tests, with one (5%) study of an imaging test and one (17%) study of a physiological test reporting the CVs. RCVs and IIs were only reported in studies of laboratory tests. ICC measures were used in 10 (50%) studies of imaging tests, 2 (33%) studies of physiological tests and 6 studies (8%) of laboratory tests. The use of confidence intervals was seen more frequently in studies of imaging tests (n=12, 60%) and physiological tests (n=2, 33%) compared with laboratory tests (n=4, 5%), see Table 3.7.

Box 3.5: Biological variability studies: reporting exemplars.

Clarity of external analytical CV estimate
Intra-individual variation in creatinine and cystatin C, by Bandaranayake et al,[92] assessed the variability of creatinine and cystatin C in 10 healthy participants, with variability estimates calculated using CV_A obtained from external sources. This study clearly expressed external estimates of CV_A would be used in the materials and methods section of the article and stated the estimates of CV_A for both creatinine and cystatin C. Many other studies do not explicitly inform readers that external estimates of CV_A are used and do not give the value of the CV_A estimate.

Use of confidence intervals
Within-subject biological variation of glucose and HbA1c in healthy persons and in type 1 diabetes patients, by Carlsen et al,[95] was one of the few studies reviewed that provided any indication of the precision of the estimates generated by the study. This study reported 95% confidence intervals, generated using the methods of Burdick and Graybill.[34] Demonstrating precision of estimates is vital to allow the estimates presented to be appropriately interpreted and used further.

It is also noted, many of the studies published in laboratory clinical medicine journals were very short articles, and the low word count available may contribute to lack of clarity.

3.4.5 What are the differences between studies assessing biological variability of laboratory, imaging and physiological tests?

There were few studies assessing physiological (n=6, 6%) and imaging test (n=20, 20%) identified by the review, with the majority of studies (n=75, 74%) assessing laboratory tests. Biological variability studies appear to be defined for laboratory based studies (n=61/75, 81%) but not for imaging and physiological test studies, (n=5/20 (25%) and n=1/6 (17%) refer to biological variation in the study title respectively). In the previous sections detailed comparisons between test types were made; key issues from these analyses are presented here.

Many of the studies evaluating laboratory based tests assessed only healthy participants (n=43/75, 57%), and this was less common for physiological and imaging tests (n=5/26, 19%).

With 52 of the 74 (69%) studies assessing laboratory tests referencing or following the framework of design and analysis introduced by Fraser and Harris, the majority of laboratory test studies used similar study designs and methods. Many of the studies assessing laboratory tests, tested to identify non-normality of data (n=34/75, 45%) and outliers (n=25/75, 33%), whereas this was not as common in the non-laboratory test studies (n=4/26 (15%) and 0/26, (0%) for normality checking and outlier detection respectively). Sample sizes were generally small, with only a few exceptions.

The quality of reporting appeared similar across the studies of different test types. Biological variability studies of laboratory tests reported CVs (n=73/75, 93%), RCVs (and n=48, 64%) and II (n=44, 59%). The studies assessing non-laboratory tests mostly reported ICCs (n=12/26, 46%) and other measures (percentage agreement, Kappa, AUROC and Bland-Altman limits), only one study (n=1/26, 4%) reported CV_A and two studies (n=2/26, 8%) reported CV_I. Giving an estimate of uncertainty by producing confidence intervals was very rare in the laboratory test studies (n=4/75, 5%) but more common in imaging and physiological test studies (n=14/26, 54%).

3.5 Discussion

3.5.1 What is the current state of the field? What are the aims of these studies and in which tests and test areas are they seen?

It was not easy to identify studies of biological variability, suggesting these studies are not often performed and/or published. A paucity of biological variability studies would be concerning given the importance of these studies, not just for planning monitoring strategies for patients with potential disease progression and recurrence but also for selecting tests to be evaluated for accuracy, or conversely identifying tests that are not fit for purpose and should not be further investigated.

The review mainly identified studies investigating the variability of clinical laboratory tests. This may suggest studies of biological variability are more common in this area or it may be the searches used identified more studies of clinical laboratory tests than imaging and physiological tests.

Most studies did not aim to assess the variability of a single test situation but multiple tests, testing populations, measurements from tests, or time points.

3.5.2 How are studies assessing biological variability of tests designed?

The populations assessed in most of the identified studies were partly or fully formed of healthy individuals. This practice is problematic as test performance and variability may be different in non-healthy/diseased populations, in whom the tests will be used for diagnostic and monitoring purposes, compared to healthy populations. However, knowledge of variability of tests in healthy populations may be beneficial for screening and developing reference ranges.

Sample size justification was found to be absent in almost all studies identified; this is likely due to limited funding and the burden of repeated testing on participants, meaning only a small sample can be achieved. The recent work of Røraas and colleagues[43] may help improve

the planning of sample sizes for studies in the future and additional guidance may also be required to enable researchers to appropriately plan studies.

The relatively short duration of studies and fast rate of retesting is to be expected as measures should be taken over a stable period of disease. The small number of repeats generally seen in these biological variability studies is also likely linked to the burden on patients and the need to test within a stable period of disease. Guidance on the number of repeats necessary is required to help researchers plan these studies.

The levels of variability assessed were mainly within-individual variability and between-individual variability. The studies where analytical variability was estimated using control samples or an estimate from a separate source was used are concerning as this estimate of variability may not be appropriate. Encouragingly, reduction of pre-analytical variability prior to test assessment was seen in many of the studies.

Blinding is not explicitly reported in the majority of laboratory based studies, this is potentially due to blinding of assessors (laboratory workers) being assumed as the analysis of samples is performed separately without the input of clinicians or participants. The issue of blinding is much more important in the imaging test studies, especially those where inter and intra rater reliability is assessed as ideally observers would be blinded to the true condition of the participant and also to the measurement(s) obtained from the other observer(s). Blinding has perhaps been used in more of these studies but poor reporting means this is not clear.

3.5.3 How are studies assessing biological variability of tests analysed?

Methods for analyses were mainly ANOVA and random effects models. However, for some studies the methods were not clear and it was assumed ANOVA or random effects modelling methods had been used, based on the results reported. For the biological variability studies of laboratory tests the ANOVA and random effects modelling methods are likely used as the majority of studies followed the methods of Fraser and Harris;[35] studies of other test types more often used alternative methods.

The practice of transforming data and, identifying and removing outliers was used in many of the studies identified; again this is likely due to the framework of Fraser and Harris[35] and was mainly seen in studies of laboratory tests.

Due to the framework, it is assumed some studies of laboratory tests may have transformed the data, or at least tested the normality of the data, but not reported this explicitly; whereas, this is less likely for imaging and physiological test studies. Some studies may also be log transforming data as this simplifies calculations (see Chapter 2), rather than for distributional benefits. It is also anticipated (due to the framework) that although some studies do not directly report the identification of outliers and consequent removal this has been considered; however, the non-laboratory test studies may not consider outliers when analysing the data.

3.5.4 How are studies assessing biological variability of tests reported?

Studies of laboratory tests may have been identified more easily as the clinical laboratory community have defined terminology for these studies. Identification of biological variability studies was difficult, and the established database for laboratory tests provided most of the studies in the review. Correct terminology and labelling of these studies would make it easier for biological variability studies to be identified.

The clarity of design and methods for analysis was variable for the identified studies with studies lacking detail of: justification of sample size, number of measurements and timing of repeats and the methods for analysis. In some cases the detail regarding timing and frequency of measures is missing due to the ad hoc nature of taking measurements, for example the studies taking advantage of routinely collected data meaning measurements are made as and when standard practice dictates. In general, all aspects of reporting for these studies could be improved. The Bartlett checklist[69] and the exemplars given here will hopefully provide authors with guidance to improve the reporting of these studies. It is also noted many of the studies published in laboratory clinical medicine journals were very short articles, and the low word count available may contribute to lack of clarity.

Often biological variability estimates are not given with corresponding uncertainty estimates. Again the work of Røraas and colleagues[43] will hopefully give researchers guidance and lead to improvements.

3.5.5 What are the differences between studies assessing biological variability of laboratory, imaging and physiological tests?

There are clear differences between the studies of biological variability for laboratory tests compared with imaging and physiological tests. The laboratory tests have a set framework, whereas the other test types vary more. The imaging studies are also different as the purpose is often to explore inter and intra reader variability. Studies of different test types lend themselves to assessment of different levels of variability; some test types do not allow for assessment of analytical variability.

Issues with not justifying sample size and lack of clarity for reporting were seen in studies of all test types; however, studies of physiological and imaging tests more often reported confidence intervals for estimates.

Testing for outliers was identified as a common procedure in studies of laboratory tests but not for studies of physiological and imaging tests.

3.5.6 Limitations

This review was not a systematic review. There will be studies of biological variability not captured by this review and it is likely the searches used identified studies of better quality (due to the journals chosen for the targeted searches and selection of studies that have been previously assessed for inclusion in the Westgard QC database). Only articles written in English were included in the review and a single reviewer extracted information from the studies. The review only considered published work, which again may be of better quality than unpublished work. There is no way of knowing if the searches developed capture all studies meeting the criteria.

The criteria for inclusion in the review meant that studies using calibrated and spiked samples were not included. These studies may be beneficial in early test development and evaluation stages but cannot be used to estimate test variability for patients. The searches used were limited and could be further developed to detect a broader range of studies.

The purpose of this review was to identify methodology issues for biological variability studies with articles identified to understand these issues rather than provide an exhaustive sample of studies.

3.5.7 Further work

This review could be strengthened by improved searches, possibly including key estimates such as CV, II, RCV and ICC. This review has highlighted the need for guidance in certain areas. Some of these areas (sample size, uncertainty of estimates and general reporting) have already been identified and there is guidance for researchers.[22,43,77] As the sample size issue is vital, additional work in this area would be beneficial. Also, the impact of data transformations and outlier identification and removal requires further investigation.

3.6 Conclusions

Due to a lack of specific search terms for biological variability studies, studies were difficult to detect. Search terms should be developed for these studies and test developers, researchers and funders need to be aware of the need for variability estimates and the importance of biological variability studies.

The design, methods of analysis and clarity of reporting for biological variability studies can be improved. Primarily, methods for sample size calculation are required as this was identified as a major deficit of the identified studies. In addition to the work of Røraas et al,[43] guidance is required for sample size justification. Further work investigating sample size for biological variability studies has been conducted, see Chapter 5. The population assessed in these studies needs to be chosen considering the likely use of the test. Studies should be

designed to evaluate all levels of variability and any subgroups should be pre-specified, also considering sample size.

The practice of outlier detection and deletion, and also data transformation is apparent. The impact of data transformation and outlier detection requires investigation, as these processes may be eliminating the variability the studies are aiming to estimate. These procedures were investigated by carrying out empirical analyses, see Chapter 4, and also the impact of these methods, see Chapter 6.

The Fraser-Harris framework requires updating. The design of studies evaluating variability should be fully considered, specifically the sample size (especially if using subgroup analyses), populations evaluated and levels of variability assessed. The methods for analysis should not be so prescriptive, allowing assessment of normality and outliers in a tailored way for each study rather than following a method that may not be appropriate.

The reporting of biological variability studies needs to be detailed and transparent. With the adoption of the Bartlett checklist,[69] this improvement can be achieved. The onus for detailed and transparent reporting should also be with journals, with the word limits for reporting allowing the necessary detail. It should be required that estimates are reported with confidence intervals.

Chapter 4

Analysis of biological variability studies: a case study evaluating glomerular filtration rate (GFR)

This work has, in part, been submitted for publication:
Rowe C, Sitch A, Barratt J, Brettell E, Cockwell P, Dalton N, Deeks J, Eaglestone G, Pellatt-Higgins T, Kalra P, Khunti K, Loud F, Morris F, Ottridge R, Stevens P, Sharpe C, Sutton A, Taal M, Lamb E. Biological variation of measured and estimated glomerular filtration rate (GFR) in patients with chronic kidney disease: the eGFR-C Study. Kidney International (in press).

Summary

This chapter presents an analysis of markers used in a study of glomerular filtration rate of patients with chronic kidney disease (CKD), the eGFR-C study. The methods for analysing

biological variability were described in Chapter 2.

Analysis of the eGFR-C study data using the methods identified in the review of biological variability studies (see Chapters 2 and 3), highlighted the differences when using normality checks and transformation, outlier detection and removal, and the use of 'exact' measures and asymmetric reference change values.

There were differences in results when data were analysed with and without transformation– with some results after log transformation requiring alternative methods. When outliers were detected and removed the estimates of variability at each level decreased, especially at the within and between-individual levels.

4.1 Introduction

Glomerular filtration rate (GFR) is the primary method of detecting chronic kidney disease (CKD) with the reference standard method for measuring GFR being iohexol clearance,[19] a method which is time consuming and unreasonable to undertake for the purposes of disease detection, staging and monitoring of progression. There are also methods for estimating GFR. Estimated GFR, or eGFR, can be calculated using equations requiring characteristics of the patient (age, gender and ethnicity) along with serum creatinine and cystatin C levels ($\text{MDRD}_{\text{creatinine}}$,[96] $\text{CKD-EPI}_{\text{creatinine}}$,[97] $\text{CKD-EPI}_{\text{Cystatin C}}$,[98] and $\text{CKD-EPI}_{\text{Cystatin C creatinine}}$[98]). Estimated GFR is often used in clinical practice as the measures required to calculate eGFR using the equations are easier to obtain.

The eGFR-C study[19] is a prospective longitudinal study designed to investigate the accuracy of the various eGFR equations for the purposes of diagnosis and monitoring of patients with CKD. The eGFR-C study also has a sub-study component where the biological variation of reference (iohexol) GFR, creatinine, and cystatin C (along with the eGFR measures calculated from their use) is assessed, in a population known to have CKD.

Methods for the design and analysis of biological variability studies are sparse with many laboratory based studies using the framework introduced by Fraser and Harris in 1989.[35]

A review of biological variability studies, see Chapter 3, showed studies of laboratory tests generally adhered to the same methods for analysis; with procedures for evaluating normality of data with data transformations if required, outlier detection with data deletion if necessary and one way analysis of variance techniques used to estimate variability at each level, generally expressed as coefficients of variation (CV) and reference change values (RCV). Using the standard methods for assessing biological variability, data obtained from the eGFR-C sub-study can be analysed to not only produce estimates for the iohexol, creatinine and Cystatin C measures but sensitivity analyses allow the standard methods to be assessed using this example; specifically the impact of data transformation and outlier detection were explored with the impact of this methodology evaluated.

4.2 Aims and objectives

There were two primary aims. Firstly, to present a standard analysis of the eGFR-C study and obtain biological variability estimates. Secondly, the impact of certain elements of the standard analysis procedure were further evaluated, namely: data transformation and outlier detection. The method for sample size justification was investigated also.

4.3 Methods

For full details of the eGFR-C study see the published protocol by Lamb et al.[19]

4.3.1 Eligibility criteria

To be eligible to enter the study, participants were required to be:

- aged 18 years or older;
- in stage 3 CKD (GFR 30-59 mL/min/1.73 m^2), diagnosed using eGFR (with at least two consecutive test results in this range at least 90 days apart and the most recent

test in the last 12 months);

- and, treated in primary or secondary care.

4.3.2 Study design

In one study centre, twenty people with stage 3 CKD were recruited to have four iohexol reference measures of GFR along with creatinine and Cystatin C at weekly intervals. In practice, the creatinine and Cystatin C measures were used to estimate GFR using the four estimating equations. Measurements for each individual were taken at the same time of the day (morning), the same day of the week and after participants had consumed only a light breakfast, see Figure 4.1

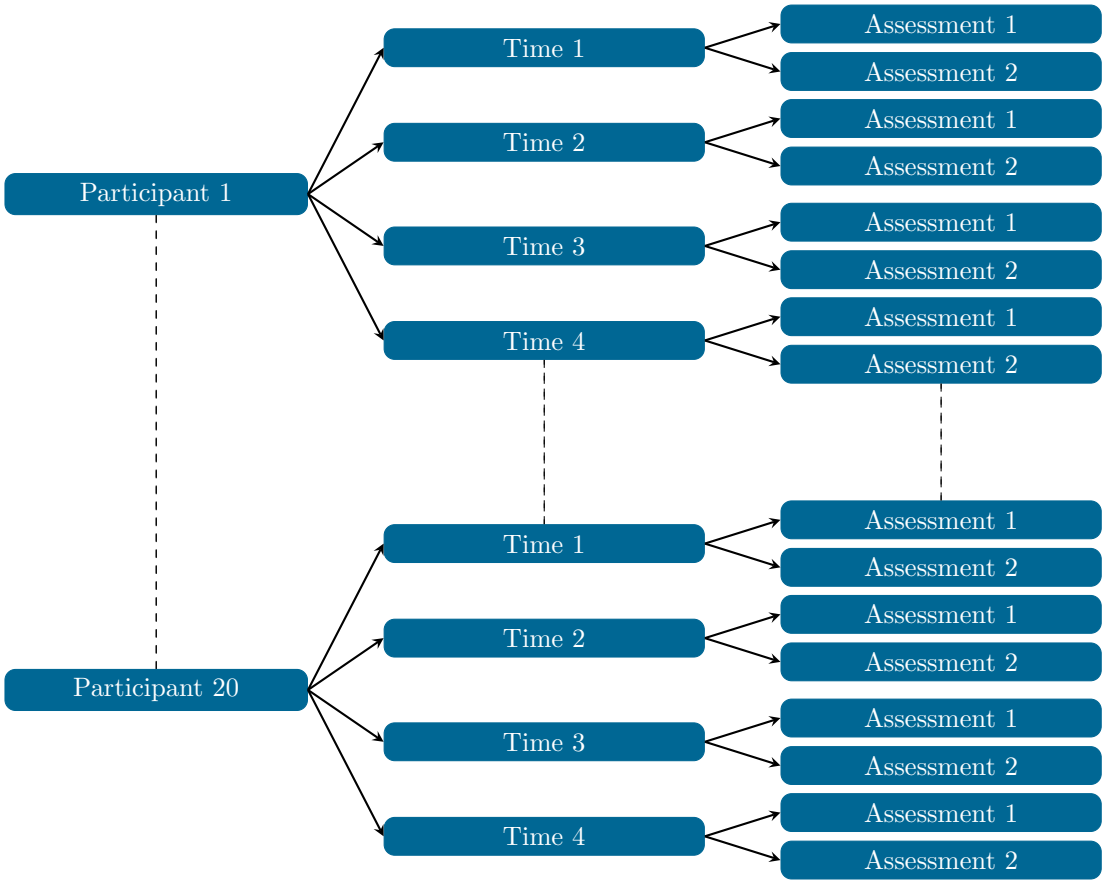


Figure 4.1: eGFR-C biological variability sub-study design.

4.3.3 Sample size

The sample size for the eGFR-C study uses an estimate of precision for within-individual CV. CV_I was estimated to be 10%. With twenty participants recruited and tested on four occasions an approximate 95% confidence interval for CV_I has limits ±2% absolute. This calculation assumes data are log-normally distributed and uses a Chi-squared distribution for the calculation of the confidence interval for CV_I.

4.4 Analysis

The analyses will follow the framework introduced by Fraser and Harris,[35] see Box 4.1 and Chapter 2 for further detail. Analyses will also be conducted to investigate the impact of certain elements of the methodology, specifically data transformation and outlier detection.

Box 4.1: Fraser Harris framework.

Normality

- Data is checked for normality using the Shapiro Wilk test, with the test used to evaluate the distribution of all data points and the distribution of measurements for each individual separately.
- If the data is non-normal for all measurements or most individuals separately, data transformation should be considered.

Outliers

- Firstly Cochran's C test is used to identify outlying measures at the level of duplicate assessments with those identified removed.
- Secondly, Cochran's C test is used to identify outliers at the within-individual level, with individuals detected having the extreme duplicate measures removed.
- Thirdly, Reed's test is used to identify outlying individual means with all measures for detected individuals removed.

Pre-analytical variability was kept to a minimum by standardising test procedures and was not considered when evaluating the results. Analytical variability, within-individual variability and between-individual variability were assessed by allowing for multiple observations within individual and multiple assessments of each observation. A linear mixed model was fitted. The model was a null model with random effects for individuals and observation points

within each individual. All analyses used the 'xtmixed' command in Stata version 15, with restricted estimation of maximum likelihood (REML).

The variability estimates at each level were expressed as coefficients of variation (CV), reference change values (RCV), index of individuality (II) and interclass cluster correlation (ICC). Estimates were presented with corresponding 95% confidence intervals (calculated using the methods of Budick and Graybill),[34] where applicable.

4.4.1 Sensitivity analyses

To evaluate the standard methods for analysis of laboratory tests a series of sensitivity analyses were performed, assessing the impact of the methods for normality testing and data transformation, outlier detection and removal, and reporting of CVs. Combinations of the following analyses were investigated:

- non-transformed and transformed data;
- no outlier detection, Fraser-Harris method for outlier detection and, Cochran C test and Reed's criterion for outlier detection (see §6.3.2);
- and, reporting of CVs and RCVs for log-transformed data.

The Fraser-Harris method uses the Cochran C test and Reed's criterion for detection of outliers, see Box 4.1. The Cochran C test requires all measurements (for an individual) to be removed if detected as an outlier when assessing variances for individuals. The method introduced by Fraser and Harris suggested that only a set of outlying duplicate results for detected individuals be removed. The Cochran C test also requires repeated use after deleting identified values until no additional outliers are detected,[99] which is not specified by the Fraser-Harris method.

Additionally, the method for calculation of geometric CV following log transformation was investigated using a 'raw' calculation of geometric CV(%) ($\sigma \times 100$) and 'exact' calculations of CV (formula for exact geometric CV(%): $\sqrt{(exp(\sigma^2) - 1)} \times 100$,[59,60] and alternative exact

geometric CV(%): $(exp(\sigma)-1)\times 100)$[61]. The calculation of symmetric and asymmetric RCVs was also considered ($RCV_{pos} = [exp(Z \times \sqrt{2}\tau) - 1] \times 100$, and $RCV_{neg} = [exp(-Z \times \sqrt{2}\tau) - 1] \times 100$, where $\tau = \sqrt{(ln(CV_{A+I}^2 + 1))}$, Z is selected from the normal distribution (usually 1.96) and CV_{A+I} the total imprecision $CV_{A+I} = \sqrt{(CV_A^2 + CV_I^2)}$)[62,63], see Chapter 2.

4.5 Results

4.5.1 Study population and completeness of data

Twenty participants were recruited; ten were male and ten female. The median (Q1, Q3) age (years) was 71 (64, 75) and the median (Q1, Q3) BMI (kg/m^2) was 28.2 (25.0, 30.2). All participants were of White/Caucasian ethnicity (see Table 4.1).

Table 4.1: Characteristics of patients recruited to eGFR-C biological variability sub-study.

Characteristic	**Summary**
N	20
Gender (male); n (%)	10 (50)
Age (years); median (Q1, Q3)	71 (64, 75)
BMI (kg/m^2); median (Q1, Q3)	28.2 (25, 30.2)

Of the twenty participants, 19 obtained results in duplicate at all four weekly assessments for all measures, giving eight measures per test per participant. One participant did not attend the fourth week of testing and only had six available results for each test measure (duplicate assessments at three time points). The total available data were 158 measurements from 20 patients for each measure.

All data were considered prior to analysis. For the iohexol measure, clinical colleagues observed eight measurements (four duplicated results for four patients) were the result of the dose not being fully administered or being subcutaneously administered; these eight measurements were removed.

4.5.2 Analyses using standard laboratory based biological variability methods

4.5.2.1 Normality testing and data transformation

Firstly data was assessed for normality. Shapiro-Wilk tests suggested iohexol, creatinine and Cystatin C data were not normally distributed (p-values 0.0004, 0.0010 and 0.0117 respectively, see Table 4.2). When performing tests on the data for each individual separately, for the iohexol measures only one individual had a significant p-value at the 5% level and for creatinine and Cystatin C data two individuals had significant p-values at the 5% level, indicating non-normality.

After log transformation, the p-value from the Shapiro-Wilk test for iohexol measures was 0.1123, suggesting no evidence of non-normality; however, for the log transformed creatinine and Cystatin C values the Shapiro-Wilk test p-values are 0.0106 and 0.0003 respectively, suggesting the log transformed data are not normally distributed. When performing tests on the log transformed data for each individual separately, for the iohexol and Cystatin C measures, the test for one individual produced a significant p-value at the 5% level and for the creatinine data tests for two individuals produced significant p-values at the 5% level, indicating non-normality.

From visual inspection of histograms (see Figure 4.2) there was little difference between the distribution of the original values and the log transformed data. As the normality of all data was marginally improved and there is a benefit regarding the calculation when using log transformed data (calculation of geometric CVs, see Chapter 2 for more details), use of log transformed data was the preferred approach for these analyses. This approach was supported by clinical chemist colleagues and was considered typical decision making for analysis of this type of study.

Table 4.2: Summaries and normality testing for non-transformed and natural log transformed Iohexol, Creatinine and Cystatin C data from the eGFR-C biological variability study.

	Iohexol		**Creatinine**		**Cystatin C**	
	Not transformed	**Transformed**	**Not transformed**	**Transformed**	**Not transformed**	**Transformed**
Measurements	150[a]	150[a]	158	158	158	158
Participants	20	20	20	20	20	20
Mean (SD)	47.6 (8.4)	3.9 (0.2)	127.6 (25.5)	4.8 (0.2)	1.7 (0.3)	0.5 (0.2)
Median (IQR)	47.8 (40.8, 52.3)	3.9 (3.7, 4.0)	124 (109, 144)	4.8 (4.7, 5.0)	1.7 (1.5, 1.9)	0.5 (0.4, 0.6)
Overall Shapiro-Wilk test p-value	0.0004	0.1123	0.001	0.0106	0.0117	0.0003
Individually Significant (at 5% level) Shapiro-Wilk test; n/N (%)	1/20 (5)	1/20 (5)	2/20 (10)	2/20 (10)	2/20 (10)	1/20 (5)

[a]Eight measurements were removed prior to analysis due to issues with test procedure

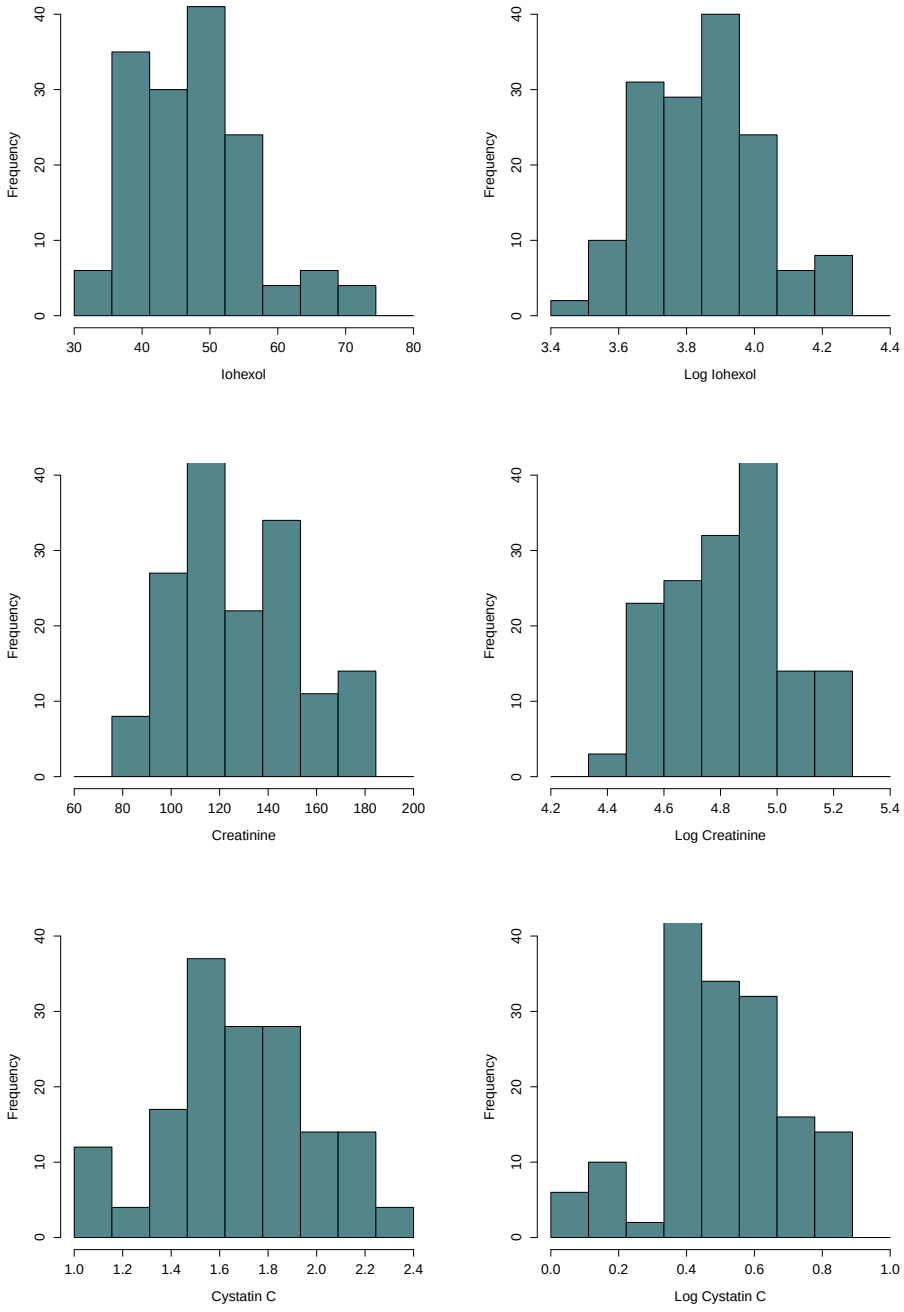


Figure 4.2: Histogram of original and log transformed measures.

4.5.2.2 Outlier detection and data exclusion

Using the log-transformed data and the method for outlier detection suggested by Fraser and Harris, for the iohexol, creatinine and Cystatin C analyses eight, four and six measurements were removed respectively, see Table 4.3.

Table 4.3: Analysis of the eGFR-C biological variability study–outlier detection using the Fraser-Harris method.

Outlier Detection	**Iohexol**	**Creatinine**	**Cystatin C**
Measurements	142	154	152
Participants	20	20	20
Mean (SD)	3.9 (0.2)	4.8 (0.2)	0.5 (0.2)
Median (IQR)	3.9 (3.7, 4.0)	4.8 (4.7, 5.0)	0.5 (0.4, 0.6)
Cochran C duplicates			
Measurements	6	4	6
Duplicates	3	2	3
Participants	3	2	3
Cochran C individuals			
Measurements	2	0	0
Duplicates	1	0	0
Participants	1	0	0
Reed's test			
Measurements	0	0	0
Duplicates	0	0	0
Participants	0	0	0
Total outliers	8	4	6

For the iohexol analysis, six measurements were removed after the first use of the Cochran C test (at the level of duplicate measurements within individuals) and two measurements for the second use of the Cochran C test (at the level of individuals within the whole group). The analysis of creatinine removed four measurements after the first use of the Cochran C test; analysis of Cystatin values removed six measurements after the first use of the Cochran C test. For creatinine and Cystatin C analyses, the second use of the Cochran C test led to no outlier detection. For all measures, Reed's test revealed no outliers for exclusion.

4.5.2.3 Analysis of variance

The analysis of iohexol results following the use of the Fraser-Harris method allowed the exact geometric coefficients of variation (expressed as percentage) and corresponding 95%

confidence intervals to be calculated. The coefficient of variation at the analytical level, CV_A was 2.22 (95% CI: 1.92, 2.63); at the within-individual level, CV_I was 6.67 (5.60, 8.20); and at the between-individual level, CV_G was 16.61 (12.43, 24.53). Positive and negative RCV bounds were calculated, and are again expressed as percentages; the positive RCV bound was 21.49 and the negative was -17.69, the index of individuality was 0.42, see Table 4.4.

Table 4.4: Analysis of the eGFR-C biological variability study–results using the Fraser-Harris method.

Test	**Iohexol**	**Creatinine**	**Cystatin C**
Measurements	142	154	152
Participant	20	20	20
Raw geometric CVs[a]			
CV_A	2.22 (1.92, 2.63)	0.66 (0.57, 0.78)	0.56 (0.48, 0.66)
CV_I	6.66 (5.59, 8.19)	4.34 (3.68, 5.30)	3.99 (3.38, 4.86)
CV_G	16.61 (12.43, 24.53)	19.79 (14.98, 29.00)	18.86 (14.28, 27.63)
CV_{TOT}	18.04 (14.27, 25.53)	20.28 (15.62, 29.34)	19.29 (14.84, 27.92)
Exact geometric CVs[b]			
CV_A (95% CI)	2.22 (1.92, 2.63)	0.66 (0.57, 0.78)	0.56 (0.48, 0.66)
CV_I (95% CI)	6.67 (5.60, 8.20)	4.35 (3.68, 5.30)	3.99 (3.38, 4.87)
CV_G (95% CI)	16.73 (12.48, 24.91)	19.99 (15.07, 29.62)	19.03 (14.36, 28.17)
CV_{TOT} (95% CI)	18.18 (14.34, 25.95)	20.49 (15.71, 29.98)	19.47 (14.92, 28.48)
Alternative geometric CVs[c]			
CV_A (95% CI)	2.25 (1.94, 2.66)	0.67 (0.58, 0.79)	0.56 (0.48, 0.66)
CV_I (95% CI)	6.89 (5.75, 8.53)	4.44 (3.75, 5.44)	4.07 (3.44, 4.98)
CV_G (95% CI)	18.07 (13.23, 27.80)	21.89 (16.17, 33.65)	20.76 (15.35, 31.82)
CV_{TOT} (95% CI)	19.77 (15.34, 29.09)	22.48 (16.90, 34.09)	21.27 (16.00, 32.21)
II[d]	0.42/0.42/0.40	0.22/0.22/0.21	0.21/0.21/0.20
RCV positive[d]	21.47/21.49/22.22	12.95/12.95/13.25	11.81/11.81/12.06
RCV negative[d]	-17.67/-17.69/-18.18	-11.46/-11.47/-11.70	-10.56/-10.56/-10.76
SD_A (95% CI)	0.022 (0.019, 0.026)	0.007 (0.006, 0.008)	0.006 (0.005, 0.007)
SD_I (95% CI)	0.067 (0.056, 0.082)	0.043 (0.037, 0.053)	0.040 (0.034, 0.049)
SD_G (95% CI)	0.166 (0.124, 0.245)	0.193 (0.146, 0.283)	0.189 (0.143, 0.276)
SD_{TOT} (95% CI)	0.180 (0.143, 0.255)	0.198 (0.152, 0.286)	0.193 (0.148, 0.279)
ICC_A	0.015	0.001	0.001
ICC_I	0.137	0.046	0.043
ICC_G	0.848	0.953	0.956

All CV and RCV values as expressed as percentages. [a] $\sigma \times 100$; [b] $\sqrt{(exp(\sigma^2) - 1) \times 100}$; [c] $(exp(\sigma) - 1) \times 100$; [d] calculated using raw CV/ exact geometric CV/ alternative geometric CVs. 95% confidence intervals were calculated using methods of Burdick and Graybill.[34]

When analysing the creatinine results using the Fraser-Harris method, CV_A, CV_I and CV_G were 0.66 (0.57, 0.78), 4.35 (3.68, 5.30) and 19.79 (14.98, 29.00) respectively. The positive

and negative RCV bounds were 12.95 and -11.47, and II was 0.22. Finally for Cystatin C, the analysis using the Fraser-Harris method provided estimates of CV_A, CV_I and CV_G of 0.56 (0.48, 0.66), 3.99 (3.38, 4.87) and 18.86 (14.28, 27.63) respectively. The positive and negative RCV bounds were 11.81 and -10.56, and II was 0.21.

4.5.3 Sample size

The sample size calculation was based on precision of the estimate of CV_I. The estimate of CV_I was 10% whereas estimates obtained were lower and the confidence intervals for these estimates were within the absolute ±2% targeted.

4.5.4 Sensitivity analyses to investigate the impact of methods of analysing standard laboratory based biological variability methods

4.5.4.1 Normality testing and data transformation

When using the transformed and non-transformed data, outliers detected and excluded were the same measurements for analyses of iohexol and creatinine for both the Fraser-Harris method (eight iohexol measurements removed and four creatinine measurements removed) and complete outlier detection (12 iohexol measurements removed and four creatinine measurements removed), see appendix Table B.1 and Table B.2. Outlier detection for Cystatin C differed when using log transformed and non-transformed data; when analysing the log-transformed data fewer measurements were removed for both the Fraser-Harris method (six compared with eight) and complete outlier detection (six compared with 22), see appendix Table B.3.

Analyses of log-transformed data generally gave higher CV estimates compared with non-transformed data, with the exceptions of iohexol CV_G and creatinine CV_A, see Table 4.5-Table 4.7. It should be noted estimates obtained after log-transformation are geometric CVs.

The reference change values from the analyses of log transformed data are asymmetric (dif-

Table 4.5: Analysis of the eGFR-C biological variability study–results of Iohexol analyses.

	Not transformed			**Transformed**		
Outlier Detection	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**
Data removed	0/150	8/150	12/150	0/150	8/150	12/150
Raw CVs[a]						
CV_A (95% CI)	2.16	2.05	1.85	2.39 (2.07, 2.82)	2.22 (1.92, 2.63)	1.99 (1.73, 2.36)
CV_I (95% CI)	6.34	6.13	5.47	6.91 (5.79, 8.49)	6.66 (5.59, 8.19)	5.94 (4.98, 7.30)
CV_G (95% CI)	17.05	17.33	17.68	16.30 (12.17, 24.11)	16.61 (12.43, 24.53)	17.03 (12.79, 25.08)
CV_{TOT} (95% CI)	18.31	18.50	18.59	17.87 (14.19, 25.21)	18.04 (14.27, 25.53)	18.15 (14.24, 25.86)
Exact geometric CVs[b]						
CV_A (95% CI)				2.39 (2.07, 2.82)	2.22 (1.92, 2.63)	1.99 (1.73, 2.36)
CV_I (95% CI)				6.92 (5.80, 8.51)	6.67 (5.60, 8.20)	5.94 (4.99, 7.31)
CV_G (95% CI)				16.41 (12.22, 24.46)	16.73 (12.48, 24.91)	17.15 (12.85, 25.48)
CV_{TOT} (95% CI)				18.01 (14.26, 25.61)	18.18 (14.34, 25.95)	18.30 (14.31, 26.30)
Alternative geometric CVs[c]						
CV_A (95% CI)				2.41 (2.09, 2.86)	2.25 (1.94, 2.66)	2.01 (1.74, 2.38)
CV_I (95% CI)				7.15 (5.96, 8.86)	6.89 (5.75, 8.53)	6.12 (5.11, 7.57)
CV_G (95% CI)				17.71 (12.94, 27.26)	18.07 (13.23, 27.80)	18.57 (13.65, 28.51)
CV_{TOT} (95% CI)				19.56 (15.25, 28.67)	19.77 (15.34, 29.09)	19.90 (15.31, 29.51)
II^d	0.39	0.37	0.33	0.45/0.45/0.43	0.42/0.42/0.40	0.37/0.37/0.35
$RCVpositive^d$				22.42/22.45/23.24	21.47/21.49/22.22	18.94/18.96/19.52
$RCVnegative^d$				-18.31/-18.33/-18.85	-17.67/-17.69/-18.18	-15.92/-15.94/-16.33
RCV	18.50	17.91	15.99			
SD_A (95% CI)	1.031 (0.893, 1.219)	0.974 (0.844, 1.153)	0.884 (0.765, 1.046)	0.024 (0.021, 0.028)	0.022 (0.019, 0.026)	0.020 (0.017, 0.024)
SD_I (95% CI)	3.019 (2.532, 3.709)	2.615 (2.194, 3.213)	2.615 (2.194, 3.213)	0.069 (0.058, 0.085)	0.067 (0.056, 0.082)	0.059 (0.050, 0.073)
SD_G (95% CI)	8.146 (6.109, 12.011)	8.456 (6.369, 12.432)	8.456 (6.369, 12.432)	0.159 (0.118, 0.235)	0.162 (0.121, 0.239)	0.170 (0.128, 0.251)
SD_{TOT} (95% CI)	8.748 (6.889, 12.431)	8.895 (6.941, 12.738)	8.456 (6.369, 12.432)	0.175 (0.139, 0.246)	0.176 (0.140, 0.249)	0.181 (0.142, 0.259)
ICC_A	0.014	0.012	0.010	0.018	0.015	0.013
ICC_I	0.119	0.110	0.086	0.149	0.137	0.107
ICC_G	0.867	0.878	0.904	0.833	0.848	0.881

All CV and RCV values as expressed as percentages. [a] $\sigma \times 100$ for log transformed data; [b] $\sqrt{(exp(\sigma^2) - 1) \times 100}$; [c] $(exp(\sigma) - 1) \times 100$; [d] calculated using raw CV/ exact geometric CV/ alternative geometric CVs. 95% confidence intervals were calculated using methods of Burdick and Graybill[34] with no CIs provided for non-transformed data. Shaded column shows standard data analysis.

Table 4.6: Analysis of the eGFR-C biological variability study–results of Creatinine analyses.

	Not transformed		**Transformed**	
Outlier Detection	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**
Data removed	0/158	4/158	0/158	4/15
Raw CVs[a]				
CV_A (95% CI)	0.70	0.71	0.66 (0.57, 0.78)	0.66 (0.57, 0.78)
CV_I (95% CI)	4.32	4.33	4.31 (3.65, 5.26)	4.34 (3.68, 5.30)
CV_G (95% CI)	19.86	19.71	19.92 (15.08, 29.18)	19.79 (14.98, 29.00)
CV_{TOT} (95% CI)	20.34	20.19	20.39 (15.70, 29.51)	20.28 (15.62, 29.34)
Exact geometric CVs[b]				
CV_A (95% CI)			0.66 (0.57, 0.78)	0.66 (0.57, 0.78)
CV_I (95% CI)			4.31 (3.65, 5.26)	4.35 (3.68, 5.30)
CV_G (95% CI)			20.12 (15.17, 29.81)	19.99 (15.07, 29.62)
CV_{TOT} (95% CI)			20.60 (15.79, 30.16)	20.49 (15.71, 29.98)
Alternative geometric CVs[c]				
CV_A (95% CI)			0.66 (0.57, 0.78)	0.67 (0.58, 0.79)
CV_I (95% CI)			4.41 (3.72, 5.40)	4.44 (3.75, 5.44)
CV_G (95% CI)			22.04 (16.28, 33.88)	21.89 (16.17, 33.65)
CV_{TOT} (95% CI)			22.62 (17.00, 34.32)	22.48 (16.90, 34.09)
II[d]	0.22	0.22	0.22/0.22/0.20	0.22/0.22/0.21
RCV positive[d]			12.84/12.85/13.14	12.95/12.95/13.25
RCV negative[d]			-11.38/-11.39/-11.61	-11.46/-11.47/-11.70
RCV	12.14	12.16		
SD_A (95% CI)	0.897 (0.777, 1.061)	0.901 (0.780, 1.066)	0.007 (0.006, 0.008)	0.007 (0.006, 0.008)
SD_I (95% CI)	5.510 (4.664, 6.722)	5.512 (4.666, 6.725)	0.043 (0.037, 0.053)	0.043 (0.037, 0.053)
SD_G (95% CI)	25.307 (19.159, 37.078)	25.088 (18.992, 36.761)	0.199 (0.151, 0.292)	0.198 (0.150, 0.290)
SD_{TOT} (95% CI)	25.915 (19.956, 37.499)	25.703 (19.796, 37.185)	0.204 (0.157, 0.295)	0.203 (0.156, 0.293)
ICC_A	0.001	0.001	0.001	0.001
ICC_I	0.045	0.047	0.047	0.046
ICC_G	0.954	0.951	0.952	0.953

All CV and RCV values as expressed as percentages. [a] $\sigma \times 100$ for log transformed data; [b] $\sqrt{(exp(\sigma^2) - 1) \times 100}$; [c] $(exp(\sigma) - 1) \times 100$; [d] calculated using raw CV/ exact geometric CV/ alternative geometric CVs. 95% confidence intervals were calculated using methods of Burdick and Graybill[34] with no CIs provided for non-transformed data. Shaded column shows standard data analysis.

Table 4.7: Analysis of the eGFR-C biological variability study–results of Cystatin C analyses.

	Not transformed			**Transformed**	
Outlier Detection	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**
Data removed	0/158	8/158	22/158	0/158	6/158
Raw CVs[a]					
CV_A (95% CI)	0.56	0.52	0.53	0.60 (0.52, 0.71)	0.56 (0.48, 0.66)
CV_I (95% CI)	4.11	3.69	3.32	3.98 (3.37, 4.85)	3.99 (3.38, 4.86)
CV_G (95% CI)	18.05	17.95	18.24	18.82 (14.25, 27.56)	18.86 (14.28, 27.63)
CV_{TOT} (95% CI)	18.52	18.33	18.55	19.24 (14.81, 27.86)	19.29 (14.84, 27.92)
Exact geometric CVs[b]					
CV_A (95% CI)				0.60 (0.52, 0.71)	0.56 (0.48, 0.66)
CV_I (95% CI)				3.98 (3.37, 4.86)	3.99 (3.38, 4.87)
CV_G (95% CI)				18.98 (14.32, 28.09)	19.03 (14.36, 28.17)
CV_{TOT} (95% CI)				19.42 (14.89, 28.41)	19.47 (14.92, 28.48)
Geometric CVs[c]					
CV_A (95% CI)				0.60 (0.52, 0.71)	0.56 (0.48, 0.66)
CV_I (95% CI)				4.06 (3.43, 4.97)	4.07 (3.44, 4.98)
CV_G (95% CI)				20.70 (15.31, 31.74)	20.76 (15.35, 31.82)
CV_{TOT} (95% CI)				21.22 (15.96, 32.12)	21.27 (16.00, 32.21)
II[d]	0.23	0.21	0.18	0.21/0.21/0.20	0.21/0.21/0.20
RCV positive[d]				11.79/11.80/12.04	11.81/11.81/12.06
RCV negative[d]				-10.55/-10.55/-10.75	-10.56/-10.56/-10.76
RCV	11.49	10.34	9.33		
SD_A (95% CI)	0.009 (0.008, 0.011)	0.009 (0.008, 0.010)	0.009 (0.008, 0.010)	0.006 (0.005, 0.007)	0.006 (0.005, 0.007)
SD_I (95% CI)	0.069 (0.059, 0.084)	0.062 (0.052, 0.076)	0.055 (0.046, 0.067)	0.040 (0.034, 0.049)	0.040 (0.034, 0.049)
SD_G (95% CI)	0.303 (0.230, 0.445)	0.301 (0.228, 0.441)	0.300 (0.227, 0.439)	0.188 (0.142, 0.276)	0.184 (0.139, 0.269)
SD_{TOT} (95% CI)	0.311 (0.240, 0.450)	0.307 (0.236, 0.445)	0.305 (0.234, 0.443)	0.192 (0.148, 0.279)	0.188 (0.145, 0.272)
ICC_A	0.001	0.001	0.001	0.001	0.001
ICC_I	0.049	0.041	0.032	0.043	0.043
ICC_G	0.950	0.959	0.967	0.956	0.956

All CV and RCV values as expressed as percentages. [a] $\sigma \times 100$ for log transformed data; [b] $\sqrt{(exp(\sigma^2) - 1) \times 100}$; [c] $(exp(\sigma) - 1) \times 100$; [d] calculated using raw CV/ exact geometric CV/ alternative geometric CVs. 95% confidence intervals were calculated using methods of Burdick and Graybill[34] with no CIs provided for non-transformed data. Shaded column shows standard data analysis.

ferent values indicate a true change depending on positive or negative change between measures);[62,63] whereas, RCVs calculated from the analyses of non-transformed data give a single value for changes in either the positive or negative direction. The RCVs calculated using the log transformed data appeared conservative compared to the non-transformed data (log-transformed and non-transformed data using Fraser-Harris method: iohexol -17.69, 21.49 and ±17.91; creatinine -11.47, 12.95 ±12.16; Cystatin C -10.56, 11.81 and ±10.34).

The index of individuality was stable across analyses of log-transformed and non-transformed data for creatinine and Cystatin C. For the analyses of iohexol the IIs estimated showed greater change when calculated from the log transformed and non-transformed data (Fraser-Harris method: 0.42 and 0.37 respectively).

4.5.4.2 Outlier detection and data exclusion

All outlier testing methods led to the exclusion of data when analysing the three measures. For iohexol, complete outlier detection deleted more measurements (12 measurements for full detection and eight for Fraser-Harris method) due to the Cochran C test leading to the removal of all data for the identified individual, see appendix Table B.1. When identifying outliers in the creatinine data, the outlier detection methods identified the same four measurements, see Appendix B Table B.2. Identification of outliers in the Cystatin C data led to deletion of the same six measurements when using the Fraser-Harris and complete outlier detection methods on the transformed data; however, for the non-transformed Cystatin C data, additional measurements were identified using the complete outlier detection method (22 compared with eight measurements, see appendix Table B.3). The additional measurements were detected when using the Cochran C test for individuals (the complete outlier detection method identified all eight measurements to be removed for an individual and a further eight were removed when another individual was identified using Cochran C testing of individuals for a second time on the remaining data). See Figure 4.3.

Comparison of analyses of data after different outlier detection approaches were used suggested that estimates of CV_A and CV_I decrease as more data were excluded, with results

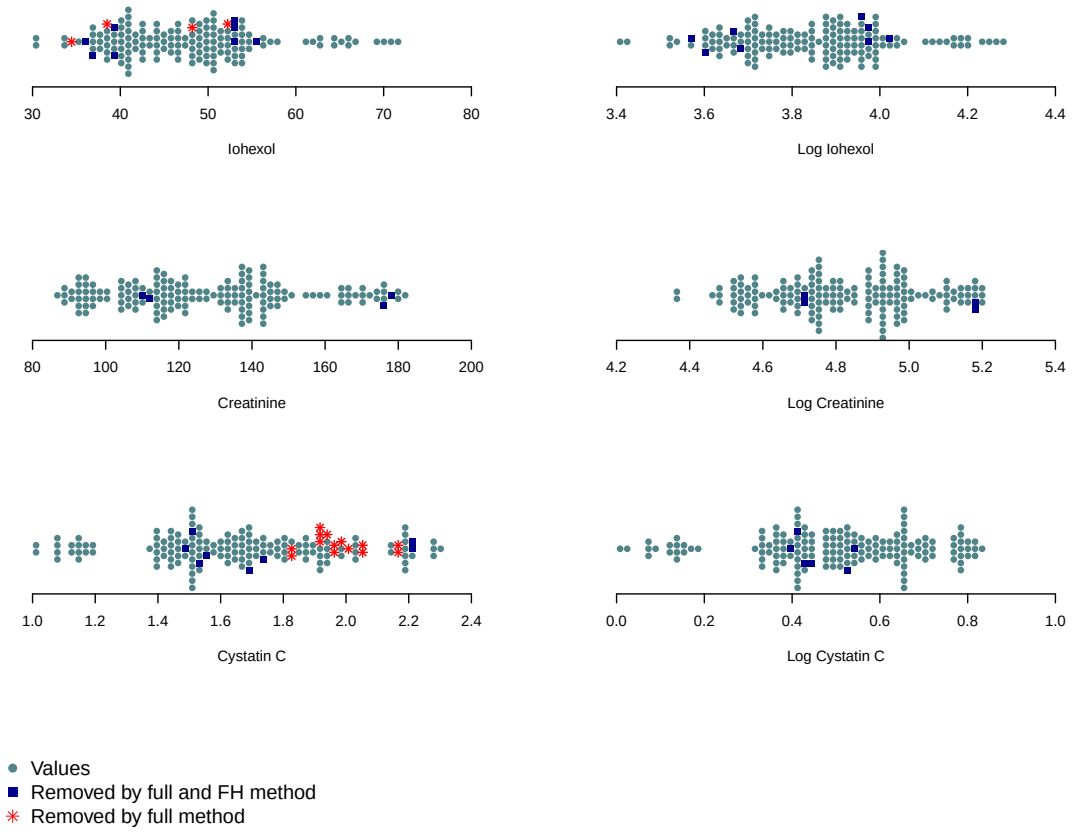


Figure 4.3: Beeswarm plot of data and removed data.

appearing similar when fewer outliers were detected. Comparing estimates of CV_G showed increases with outlier detection for the analysis of iohexol and Cystatin C (see Table 4.5 and Table 4.7) but decreases for creatinine. Similar trends were seen when analysing the non-transformed data.

The reference change value bounds decreased with more stringent outlier detection when analysing the iohexol data. The RCVs for creatinine and Cystatin C appeared similar when using different outlier detection methods, and the RCV values for the non-transformed data decreased with stricter outlier detection methods for the analysis of iohexol and Cystatin C. For the analyses of the non-transformed creatinine data the RCVs remained stable with different methods of outlier detection.

The index of individuality reduced with increased outlier detection for the analyses of log transformed iohexol data but remained stable for the analysis of log transformed creatinine and Cystatin C data. The IIs estimated when analysing non-transformed data decreased as outlier detection increased for iohexol and Cystatin C measures but for creatinine remained similar.

4.6 Discussion

4.6.1 Sample size

The sample size calculation used in this study focused on the estimate of within-individual coefficient of variation only, using the number of participants and observation points to assess precision. The method used for this was simplistic, as the duplicated measures within each observation point were not factored into the calculation. The method of using confidence intervals for a given level of precision is appropriate but further guidance is required to ensure this is correct and the precision of additional estimates are considered. A tool to enable researchers to easily use these methods to plan the sample size of studies is required.

4.6.2 Normality and data transformation

The results of analysing the values recorded for iohexol, creatinine and Cystatin C tests showed estimates were slightly different when using log transformed and non-transformed data. As the estimates generated from these analyses have shown differences depending on whether non-transformed data or transformed data was used, and the interpretation of these estimates differ, it is vital that studies of biological variation report any transformations performed and consider if data is normally distributed prior to transformation and log-normally distributed after log-transformation.

From investigating published biological variability studies (see Chapter 3) it seems log-transformation may be used to allow simplification of calculations primarily and the availability of formulae for calculating confidence intervals for estimates. The true need for transformation of data is often not given.

When using a model to evaluate three levels of variability, the model takes the form of $y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, where μ is the mean value of the measure, $\alpha_i \sim N(0, \sigma_G^2)$, $\beta_{ij} \sim N(0, \sigma_I^2)$, $\epsilon_{ijk} \sim N(0, \sigma_A^2)$ and $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$. The number of participants is n_1, the number of observations for each participant is n_2 and the number of replicate assessments of each observation for each participant is n_3. The analytical, within-individual and between-individual variability, expressed as standard deviations are σ_A, σ_I and σ_G. This model assumes normality of the variability parameters at the analytical, within-individual and between-individual levels. If this assumption is not held results from the model may not be valid. Simply assessing the normality of the test measures may not be sufficient to investigate if the data meets the assumptions of the model. Many of the outlier detection methods (see §6.3.2) rely on the data to be normally distributed, hence the further requirement for normality prior to assessing outliers.

When using normality tests on small samples the tests have limited power; and, with larger samples the results of normality tests may be significant, indicating non-normality, when deviations from the normal distribution are small and alternative approaches (non-parametric testing or transformation) cause limited differences to the results obtained.[100]

Further work is required to investigate issues arising when biological variability data are not normally distributed or log-normally distributed.

4.6.3 Outlier detection and removal

The results of the analyses of iohexol, creatinine and Cystatin C measures showed estimates were different when outliers are detected and removed. In situations with increased outlier detection estimates of variability were generally reduced, meaning these methods were potentially providing optimistic estimates of variability. The risk in using outlier detection methods is that valid data is removed rather than data 'errors' and the consequence is reduced estimates of variability.

The results also showed different types of outlier detection led to different results. The method of Fraser and Harris can be interpreted in different ways leading to differing numbers of outliers detected which can change the results of biological variability studies. The exact use of outlier detection methods need to be carefully explained, if these methods are considered appropriate. It is not clear what the impact of outlier detection is and the most appropriate method to use in biological variability studies.

The Cochran C test is not a perfect test to use even if outlier detection is appropriate and necessary. 't Lam[99] discusses the disadvantages of the Cochran C test: requiring a balanced design; use of the test requires reading of critical values from tables and, the test is not a two-sided test as it uses critical values to identify large variances but does not identify small variances.

4.6.4 Limitations

This chapter displays the analysis of only one data set where the distribution of data and outliers did not appear to have great impact on the analyses performed. For other data sets the difference in analyses could be greater and this requires further investigation.

Differences in results were seen across the analyses but it is difficult to know what magnitude

would be meaningful. CVs are often the reported results of biological variability studies but it may be that estimates of RCV are more meaningful and researchers should be mindful of the precision of this estimate when planning studies.

4.7 Conclusion and recommendations

The methods for analysing biological variability studies can impact on results, this could be using: log-transformed or non-transformed data, different outlier detection methods or different methods of calculation of geometric coefficients of variation. The methods for evaluating biological variability studies require updating and researchers need to be made aware of the methods for analysing these types of studies and the differences in interpretation of results when geometric coefficients of variation are calculated. Only with clear and transparent reporting can the analyses of biological variability studies be used to inform further research.

It needs to be clearly reported if the data has been analysed using log-transformed data or data on the original scale. The decision to transform data should be made given the distribution of study data and prior knowledge of the measure, with these decisions and considerations reported. Researchers should also clearly report which method has been used to give CV estimates when using log-transformed data. The benefits of using the log-transformed data are simplification of the calculation of CV and the ability to calculate confidence intervals for these CVs.

Using outlier detection methods may lead to the identification and deletion of legitimate test measurements and consequently may decrease the estimates of variability. The impact of outlier detection methods on variability estimates in biological variability studies needs further investigation with the methods for identifying outliers compared.

More guidance is required for planning sample sizes for biological variability studies taking account of the variability at all levels assessed, which can be done with the appropriate confidence interval calculations. It would also be beneficial to have guidance for sample size

based on the precision of measures in addition to CVs, such as RCVs. A tool for calculating the precision of estimates for given sample sizes would allow researchers to easily incorporate these calculation when planning studies.

Chapter 5

Sample size guidance and justification for studies of biological variation

This work has been partly presented in the following form:
Sitch, A, Mallett, S, Deeks, J. Sample size guidance and justification for studies of biological variation. EuroMedLab–22nd IFCC-EFLM European Congress of Clinical Chemistry and Laboratory Medicine, Athens, Greece. 11-15 June 2017.

Summary

Biological variability studies aim to measure variability in a biomarker both between and within individuals, allowing the potential for a biomarker to diagnose and monitor disease to be assessed. Sample sizes for these studies state the numbers of participants (n_1), observations

per participant (n_2) and repeat assessments of each observation (n_3). Little guidance exists to compute these values.

Simulation of biological variability data and subsequent analysis allows potential results to be observed for more common measures of variability including the coefficient of variation (CV), reference change values (RCV), and index of individuality (II). Using simulation and observing the results can help researchers plan sample size (the application is available at `https://alicesitch.shinyapps.io/bvs_simulation/`).

Results of simulations showed greater numbers of participants increased the precision of estimated analytical, within and between-individual variability; increasing the number of observations per participant increases the precision of estimates of analytical and within individual variability; and, increasing the number of assessments of observations per participants increases precision of only analytical variability. If the desired precision of variability components are known, values for n_1, n_2 and n_3 can be determined.

5.1 Introduction

Biological variability studies look to estimate variability by assessing participants whilst in a stable disease state. Multiple participants are tested at multiple time points, and each observation for each participant is assessed multiple times.[35] Recruitment of multiple participants allows the variability between participants to be assessed; multiple measures for each participant allows within-individual variability to be assessed; and, the multiple assessment of each measure from each participant allows assessment of analytical variability.[16] Pre-analytical variability is minimised by keeping test procedures constant and is not evaluated.[16]

Analysis of variability is by ANOVA or random effects modelling. Estimates of variability are often expressed as coefficients of variation (CVs), identified in the review of biological variability studies (Chapter 3). Additional estimates provided are index of individuality (II) and reference change value (RCV), for further details see Chapters 2 and 3.

From the review presented in Chapter 3, few studies of biological variability justified the

sample size used. Few studies gave any indication of the uncertainty of the estimates produced with confidence intervals rarely presented. The Fraser and Harris[35] guide for the design and analysis of biological variability studies does not cover sample size justification, stating only that: *'valid estimates of the components of variation can be obtained from relatively small numbers of specimens collected from a small group of subjects over a reasonably short period of time'.*[35]

Røraas et al[43] have provided guidance for both sample size justification and the use of confidence intervals for estimates from biological variability studies. Their work was based on a simulation study using the standard ANOVA or random effects modelling approach to analyse biological variability data, whilst varying the analytical variability, number of replicates, number of samples, and number of individuals, and looked at the effect of varying these study aspects on the confidence interval width for the estimate of within-individual variability. Tables provided enable planning of appropriate sample size and estimation of confidence intervals for within-individual variability estimates. Confidence intervals were calculated using the formula introduced by Burdick and Graybill,[34] see Chapter 2.

McNeish and Stapleton[50] reviewed 'rules of thumb' for the number of clusters to fit multilevel models and achieve unbiased estimates. The authors acknowledge a *'a specific sample size cannot be pinpointed'* with guidance ranging. Kreft[101] suggested 30 clusters with 30 data points within each cluster; Snijders and Bosker[102] suggested 20 clusters were necessary and multilevel models should not be used if fewer than 10 clusters are present;[103] and, Hox[104] suggested 50 clusters with 10 data points in each cluster was necessary for multilevel modelling, with this increased to 100 clusters with 10 datapoints in each if variance parameters were estimated. McNeish and Stapleton[50] used simulation to suggest a minimum of ten clusters to estimate a variance parameter and 50 clusters to estimate the standard error of a variance parameter. A generic multilevel model was used (also with the estimate of fixed effects) with only two levels of data. This example was not tailored to biological variability data but for more general analysis accounting for clustering of data. The validity of multilevel modelling to obtain estimates of variability with small numbers of participants, observations and assessments is unknown.

The work presented in this chapter aims to use simulation to assess the validity of the methods used to produce variability estimates for varying sample sizes and provide guidance to demonstrate the impact of sample size on the precision of various estimates from biological variability studies (including estimates where precision cannot easily be derived, such as CV for non-transformed data, RCV and II). For estimates where estimated confidence intervals can be derived (CV estimates when data is log-normally distributed and log transformed) the simulation will be used to validate these methods. Researchers will be provided with a tool to enable planning and justification of sample size. This tool will not only provide precision of estimates of within-individual variability but also for analytical and between-individual variability, and the subsequent measures generated in biological variability studies.

5.2 Aims and objectives

The aims of this study were:

- primarily, to assess the validity and precision of estimates produced when varying sample size for normal and log-normal data;
- secondarily, to investigate the difference in estimates for a given sample size when changing the variability at the analytical, within-individual and between-individual level.

The simulated scenarios were evaluated by assessing:

- bias, accuracy and coverage of the methods used for analysis when estimating standard deviations;
- and, a range of estimates from biological variability studies (standard deviations, CVs, II, RCV).

This analysis allowed the validity of methods to estimate variability and calculate confidence intervals when varying sample size to be assessed, and provides researchers with likely estimates from planned studies so sample size can be modified to achieve the required precision.

An application was developed enabling sample sizes for biological variability studies to be planned and justified prior to recruitment of participants. In addition an application was developed to calculate confidence intervals given specified variability estimates and sample sizes, using analytical methods where possible.

5.3 Methods

The model simulated data for a given sample size (number of participants, observations and assessments) and test performance (between-individual, within-individual and analytical variability). After simulating the data, standard analyses were performed to estimate the between-individual, within-individual and analytical variability. The simulation and analysis was repeated 1,000 times with the results from the multiple runs analysed to assess the bias, accuracy and coverage of the estimated standard deviations and the precision of estimates of biological variability (standard deviations, CVs, II, RCV). This was repeated for different sample sizes, test performance values and normal and log-normal data. For estimates where approximate confidence intervals can be calculated using the methods of Burdick and Graybill[34] these were compared to the results obtained using simulation.

See Table 5.1 for a guide to the notation used when describing the method.

5.3.1 Number of simulations

With 1,000 simulations the 95% confidence interval for coverage, assuming an estimate of 95% would range from 93.46% to 96.27%. One thousand data simulations were used for each scenario to allow precision of estimates but also efficiency of computing, which would be required for researchers to use the tool when planning studies. A simulation of 10,000 data sets was used with the base-case scenario and results were similar when compared to those generated when using 1,000 data sets.

Table 5.1: Notation description for biological variability study sample size simulation method.

Description	Notation
Simulation inputs	
Sample size	
Number of participants	n_1
Number of observations per participants	n_2
Number of assessments per observation per participants	n_3
Test estimates	
Mean test value	μ
Analytical variability (standard deviation)	σ_A
Within-individual variability (standard deviation)	σ_I
Between-individual variability (standard deviation)	σ_G
Log transformed analytical variability (standard deviation)	σ_A^*
Log transformed within-individual variability (standard deviation)	σ_I^*
Log transformed between-individual variability (standard deviation)	σ_G^*
Simulation and results	
Sample size function	
Data simulation and analysis	
Simulated test value for each assessment of each observation for each participant	y_{ijk}
Mean test value	μ
Between-individual variability; model parameter $\alpha_i \sim N(0, \sigma_G^2)$	α_i
Within-individual variability; model parameter $\beta_{ij} \sim N(0, \sigma_I^2)$	β_{ij}
Analytical variability; model parameter $\epsilon_{ijk} \sim N(0, \sigma_A^2)$	ϵ_{ijk}
Participant number; $i = 1, \ldots, n_1$	i
Observation number; $j = 1, \ldots, n_2$	j
Assessment number; $k = 1, \ldots, n_3$	k
Estimates	
Estimated analytical variability (standard deviation)	$\hat{\sigma}_A$
Estimated within-individual variability (standard deviation)	$\hat{\sigma}_I$
Estimated between-individual variability (standard deviation)	$\hat{\sigma}_G$
Estimated analytical coefficient of variation	CV_A
Estimated within-individual coefficient of variation	CV_I
Estimated between-individual coefficient of variation	CV_G
Estimated index of individuality	II
Estimated reference change value	RCV
Estimated positive reference change value	RCV_{pos}
Estimated negative reference change value	RCV_{neg}

5.3.2 Input values

The simulation required details of the sample size of the biological variability study: the number of participants (n_1), the number of observations for each participant (n_2) and the number of replicate assessments of each observation for each participant (n_3). Also required were estimates of the mean value of the test (μ) and the analytical, within-individual and between-individual variability, expressed as coefficients of variation (CV_A, CV_I and CV_G) or standard deviations (σ_A, σ_I and σ_G).

A previous review of biological variability studies (see Chapter 3) showed the median (Q1, Q3) number of individuals in biological variability studies (n_1) was 25 (15, 40); for the number of observations per participant (n_2) this was 5 (3, 10); and, for the number of assessments per observation point 2 (2, 3). From the same review estimates of CV_A, CV_I and CV_G were extracted with the median (Q1, Q3) of 3.5 (1.4, 6.3), 10.0 (5.0, 18.5) and 26.3 (14.0, 43.9) respectively.

5.3.3 Data simulation

5.3.3.1 Basic simulation model with normally distributed errors

Test data were simulated to follow the model $y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, where μ is the mean value of the measure, $\alpha_i \sim N(0, \sigma_G^2)$, $\beta_{ij} \sim N(0, \sigma_I^2)$, $\epsilon_{ijk} \sim N(0, \sigma_A^2)$ and $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$, see Figure 5.1. This was equivalent to the model proposed by Røraas et al.[43]

This gave $n_1 \times n_2 \times n_3$ observations; n_3 assessments of n_2 observations for n_1 participants. A histogram of the simulated measures and plot of the simulated measures by participant can be seen in Figure 5.2.

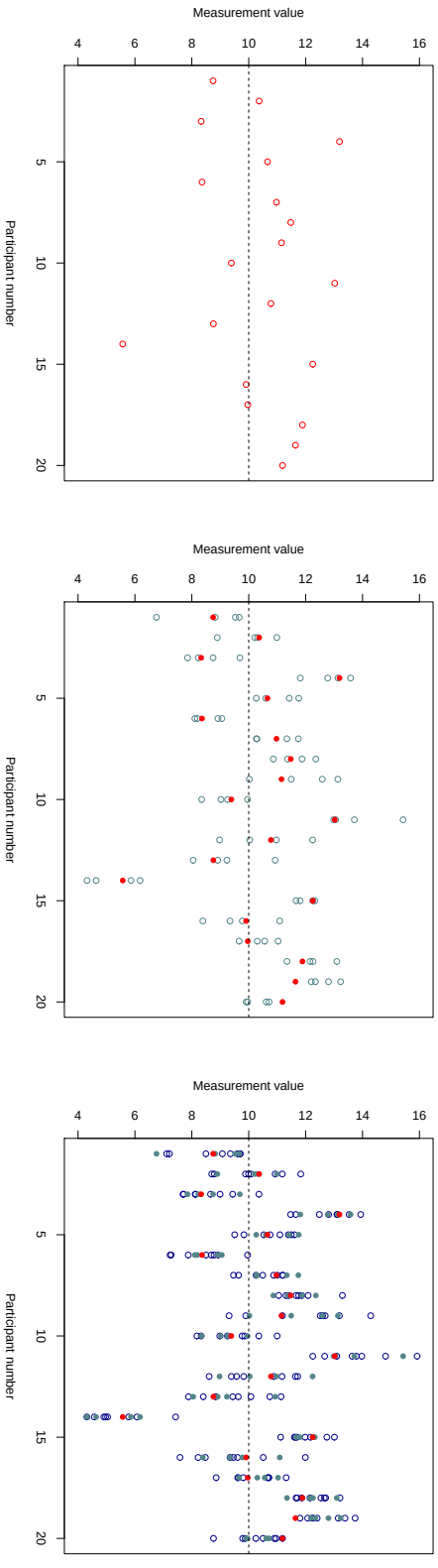


Figure 5.1: Illustration of biological variability data simulation process. Left: measurement value for each individual including between-individual variability ($\mu + \alpha_i$) red points; middle: measurement value for each individual at each time point (4 time points $n_2 = 4$) including within-individual variability ($\mu + \alpha_i + \beta_{ij}$) teal points; and, right: measurement value for each individual at each time point for each assessment (2 assessment points $n_3 = 2$) including analytical variability ($\mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$) blue points.

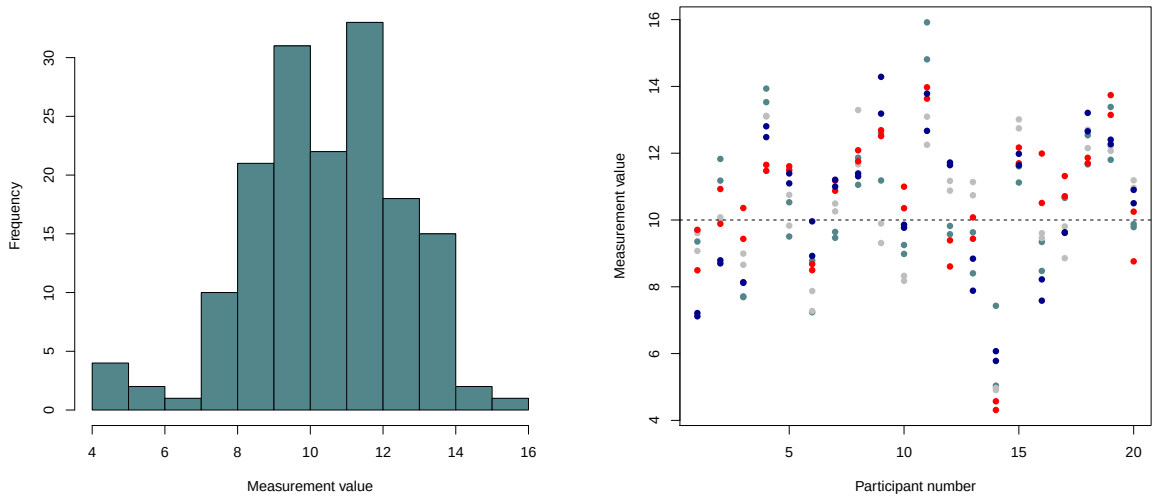


Figure 5.2: Histogram (left) and plot (right) of simulated biological variability data with normally distributed variability ($n_1 = 20, n_2 = 4, n_3 = 2, CV_G = 20\%, CV_I = 10\%$ and $CV_A = 5\%$). Duplicate assessments within observation points shown by points with the same colour.

5.3.3.2 Log-normal data simulation model

An alternative simulation of log-normal data was performed. Data were simulated following the model $y_{ijk} = \mu \times \alpha_i \times \beta_{ij} \times \epsilon_{ijk}$, thus $ln(y_{ijk}) = ln(\mu) + ln(\alpha_i) + ln(\beta_{ij}) + ln(\epsilon_{ijk})$ and $ln(\alpha_i) \sim N(0, \sigma_G^2)$, $ln(\beta_{ij}) \sim N(0, \sigma_I^2)$, $ln(\epsilon_{ijk}) \sim N(0, \sigma_A^2)$ and $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$. This gave $n_1 \times n_2 \times n_3$ observations; n_3 assessments of n_2 observations for n_1 participants. Figure 5.3 shows the distribution of the measurements on the original and log scale.

5.3.3.3 Detailed specification of the log-normal data

The analytical, within-individual and between-individual standard deviations of the log transformed data are σ_A, σ_I and σ_G respectively. The analytical, within-individual and between-individual standard deviations of the data on the original scale are σ_A^*, σ_I^* and σ_G^*.

A log-normal variable y has standard deviation σ^* and mean μ_σ^*, and when log transformed has a normal distribution such that $ln(y) \sim N(\mu_\sigma, \sigma^2)$. The mean of log trans-

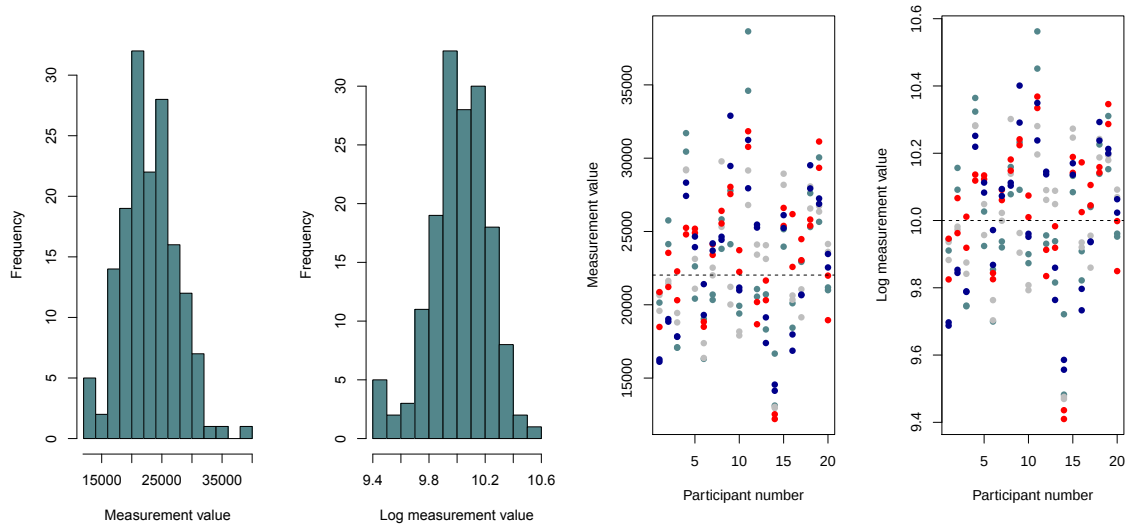


Figure 5.3: Histograms (left) and plots (right) of simulated biological variability data ($n_1 = 20, n_2 = 4, n_3 = 2, CV_G = 20\%, CV_I = 10\%$ and $CV_A = 5\%$) with log-normal distributed variability (left shows original data and right shows log transformed data). Duplicate assessments within observation points shown by points with the same colour.

formed log-normally distributed data (μ_σ) is $ln\left(\frac{\mu^*_\sigma}{\sqrt{1+\frac{\sigma^{*2}}{\mu^{*2}_\sigma}}}\right)$ and the standard deviation (σ) is $\sqrt{ln\left(1+\frac{\sigma^{*2}}{\mu^{*2}_\sigma}\right)}$.[105]

The simulation provided exact geometric estimates of CV_A, CV_I and CV_G. This was achieved by simulating the log transformed data to satisfy $\sigma = \sqrt{ln(CV^2+1)}$, so the exact CV is equal to $\sqrt{exp(\sigma^2)-1}$.[59,60] Log-normal data was simulated using the R function 'lnorm', meaning zero-cell corrections were not necessary.

The mean value of the measure on the original scale is $\bar{y}_{ijk} = \mu \times \bar{\alpha}_i \times \bar{\beta}_{ij} \times \bar{\epsilon}_{ijk}$ (hence $\mu = \frac{\bar{y}_{ijk}}{\bar{\alpha}_i \times \bar{\beta}_{ij} \times \bar{\epsilon}_{ijk}}$) and the mean value of the measure on the log scale is $ln(\bar{y}_{ijk})$. For the purpose of the simulation the value of $ln(\mu)$ was set to equal 10 units. As CV estimates are geometric for this simulation, the value of the mean is independent.

5.3.4 Analysis

5.3.4.1 Normally distributed data

The generated test data (y_{ijk}) was analysed using a linear model with random effects for participants and observation points within participants. $y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, where $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$, y_{ijk} is the test measure for the ith participant at the jth time point and for the kth assessment, μ is the mean value of the measure, $\alpha_i \sim N(0, \sigma_G^2)$, $\beta_{ij} \sim N(0, \sigma_I^2)$ and $\epsilon_{ijk} \sim N(0, \sigma_A^2)$.

Fitting the model allowed estimates of σ_A, σ_I and σ_G to be obtained, $\hat{\sigma}_A$, $\hat{\sigma}_I$ and $\hat{\sigma}_G$. In addition to standard deviations, other measures of variability were produced (coefficient of variation (CV), index of individuality (II) and reference change values (RCV)) using the estimates from the model.

5.3.4.2 Log-normal data

Analyses were performed using the same model but after log transforming the data.

5.3.5 Results for each simulation

Standard deviations were estimated after fitting the random effects model to each simulated data set, at the analytical ($\hat{\sigma}_A$), within-individual ($\hat{\sigma}_I$) and between-individual ($\hat{\sigma}_G$) level. Coefficients of variation (CV), index of individuality (II) and reference change values (RCV) were calculated. $CV_A = \frac{\hat{\sigma}_A}{\mu}$, $CV_I = \frac{\hat{\sigma}_I}{\mu}$, $CV_G = \frac{\hat{\sigma}_G}{\mu}$, $II = \frac{\sqrt{CV_A^2 + CV_I^2}}{CV_G}$ and $RCV = \sqrt{2} \times 1.96 \times \sqrt{CV_A^2 + CV_I^2}$. For the analyses of the log transformed data, exact geometric CVs were calculated using $\sqrt{exp(\hat{\sigma}^2) - 1}$.[59,60] Corresponding 95% confidence intervals for the standard deviations and CVs for the log-transformed data were calculated using the equations presented by Burdick and Graybill,[34] and asymmetric RCV, where $RCV_{pos} = exp(1.96 \times \sqrt{2}\tau) - 1$, and $RCV_{neg} = exp(-1.96 \times \sqrt{2}\tau) - 1$, where $\tau = \sqrt{ln(CV_{A+I}^2 + 1)}$, 1.96 is selected from the normal distribution and CV_{A+I} is the coefficient of variation for the total imprecision,

$CV_{A+I} = \sqrt{CV_A^2 + CV_I^2}$[62,63], as described in Chapter 2.

The application also provides estimates of geometric CVs using $exp(\hat{\sigma}) - 1$[61] and intracluster correlation coefficients (ICCs), however, these estimates are not presented here.

5.3.6 Repeated data simulations and analyses

Standard simulation performance measures were used to evaluate the ability of the methods to estimate the standard deviations for differing sample sizes and test performance. These measures were suggested by Burton et al,[106] see Table 5.2. As the amount of bias considered problematic is not know (Burton et al[106] state this has been estimated between $\frac{1}{2}SE(\hat{\beta})$ and $2SE(\hat{\beta})$) the bias is also considered as a percentage of the estimate (percentage bias) and as a percentage of the empirical standard error of the estimate from the simulations (standardised percentage bias). Burton and colleagues advocate the use of the standardised percentage bias as it: *'can be more informative, as the consequence of the size of the uncertainty'*.[106]

Table 5.2: Performance measures to assess biological variability sample size simulation results.

Evaluation criteria	**Formula**
Bias	
Bias	$\delta = \bar{\hat{\sigma}} - \sigma$
Percentage bias	$\left(\frac{\delta}{\sigma}\right) \times 100$
Standardised bias	$\left(\frac{\delta}{SE(\hat{\sigma})}\right) \times 100$
Accuracy	
Mean squared error	$\delta^2 + SE(\hat{\sigma})^2$
Coverage	
	The proportion of times the $100(1-\alpha)\%$ confidence interval includes σ
	Average $100(1-\alpha)\%$ confidence interval length

σ is the true value of the standard deviation.
$\hat{\sigma}$ is the estimated σ for each simulation.
$\bar{\hat{\sigma}}$ is the sum of the $\hat{\sigma}$ divided by the number of replicate simulations performed (the mean).
$SE(\hat{\sigma})$ is the empirical standard error of the estimate for all simulations (the SD across simulated estimates).

For each of the estimates, the mean, median, 25th percentile (Q1), 75th percentile (Q3), minimum and maximum value were calculated to summarise results. The 2.5th and 97.5th percentiles are available in the application also.

All analyses were performed using the statistical software R with the seed set at the start

of each scenario. Each scenario differs due to the sample size and data drawn from different distributions. For analyses varying sample size, data simulations are not paired. For analyses of the same sample size varying σ_A, σ_I or σ_G the results are for paired analyses changing only the specified component. All models were fitted using restricted estimation of maximum likelihood (REML).

5.3.7 Simulation inputs

5.3.7.1 Base-case

Base-case parameters were kept constant with $n_1 = 20$, $n_2 = 4$, $n_3 = 2$, $\sigma_A = 0.5$ ($CV_A = 5\%$), $\sigma_I = 1$ ($CV_I = 10\%$) and $\sigma_G = 2$ ($CV_G = 20\%$), chosen to reflect standard measures seen in the review of biological variability studies, see Chapter 3.

5.3.7.2 Varying sample size

The input parameters were varied to reflect the range of sample sizes seen in the review of biological variability studies (see Chapter 3), with the number of participants ranging from 5 to 100 (5, 10, 20, 30, 40, 60, 100); the number of observations ranging from 2 to 20 (2, 4, 6, 8, 12, 20); and, the number of assessments ranging from 2 to 10 (2, 3, 4, 6, 10). The log-normal simulation maintained the same CV estimates but the standard deviations used were slightly different to ensure the CV values remained the same.

Sensitivity analyses were performed increasing the variability of the test measures to $\sigma_A = 1$ ($CV_A = 10\%$), $\sigma_I = 2$ ($CV_I = 20\%$) and $\sigma_G = 4$ ($CV_G = 40\%$).

5.3.7.3 Varying test performance

The variability measures were chosen to reflect likely test performance also, with values (across all simulations) for σ_A varying from 0.125 to 1.25 (CV_A 1.25%, 2.5%, 3.75%, 5%, 6.25%, 7.5%, 8.75%, 10%, 11.25%, 12.5%); values for σ_I varying from 0.5 to 2.5 (CV_I 5%,

7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 22.5%, 25%); values for σ_G varying from 1 to 5 (CV_G 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%), reflecting the values of CV reported and assuming a mean test value of 10 units. Variability measures were used in simulations ensuring $\sigma_A \leq \sigma_I \leq \sigma_G$.

Sensitivity analyses were performed increasing the sample size to $n_1 = 40$, $n_2 = 10$ and $n_3 = 3$.

5.4 Results

An application was developed (allowing these simulations to be performed) to simulate different biological variability study scenarios, giving an indication of the precision of estimates. This application can be found at `https://alicesitch.shinyapps.io/bvs_simulation/`. An application was also developed allowing confidence intervals to be calculated for estimates of standard deviation at the analytical, within-individual and between-individual levels, and can be found at `https://alicesitch.shinyapps.io/bvs_cis/`.

All CVs and RCVs are displayed as percentages.

5.4.1 Number of participants, observations and assessments

5.4.1.1 Analysis of normally distributed data

Bias

For σ_A, σ_I and σ_G the bias appeared to be negative for all simulated situations except for some scenarios using larger sample sizes, see Table 5.3. With only five participants, four observations and two assessments the percentage bias–bias as a percentage of the true value–was -0.564, -1.834 and -7.250 at the analytical, within-individual and between individual levels respectively; the standardised percentage bias–the bias expressed as a percentage of the standard error of the estimate–was -3.698, -8.783 and -20.212. When increasing the number of participants to 20, the percentage bias was smaller at the analytical, within-individual and

between-individual levels (-0.175, -0.332 and -1.529); with 100 participants the percentage bias decreased further (+0.072, -0.127, -0.012). The bias was less than 2% for all situations with at least 20 participants. With increases in the number of participants (n_1) the bias for σ_A, σ_I and σ_G decreased; with increases in the number of observations (n_2) the bias for σ_A and σ_I decreased; and for increases in the number of assessments (n_3) the bias for σ_A decreased.

Confidence intervals

As the number of participants, observations and assessments increased the mean width of the Burdick and Graybill[34] 95% confidence intervals for estimated CV_A, CV_I and CV_G all reduced in width. The coverage of 95% confidence intervals for estimates of σ_A, σ_I and σ_G was consistently close to 95%, with two of the scenarios providing coverage estimates for σ_G at the lower bound of what was expected given the number of simulations, see Table 5.4 and Figure 5.4. For some of the smaller sample sizes 95% confidence intervals could not be calculated (when $n_1 = 5, 10$ and $n_2 = 2$). Comparison of the results from the simulated data to the bounds produced by the Burdick and Graybill confidence intervals (see Figure 5.5) suggest the lower bound is underestimated.

Estimates of standard deviations and coefficients of variation

The median estimate of σ_A, σ_I and σ_G, and CV_A, CV_I and CV_G appeared consistent for each number of participants, observations and assessments. For increases in the number of participants (n_1) the range of estimates for σ_A, σ_I and σ_G (and therefore CV_A, CV_I and CV_G) decreased with estimates from the 1,000 replications closer to the true value. For increases in the number of observations (n_2) the range of estimates for σ_A and σ_I (and CV_A and CV_I) decreased; however for all numbers of observations, the range of estimates of σ_G (and CV_G) were constant. For increases in the number of assessments, the range of estimates of σ_A (and CV_A) decreased only; the estimates of σ_I and σ_G (CV_I and CV_G) were similar, see Figures 5.5 and 5.6 and Appendix C Table C.1 and Table C.2.

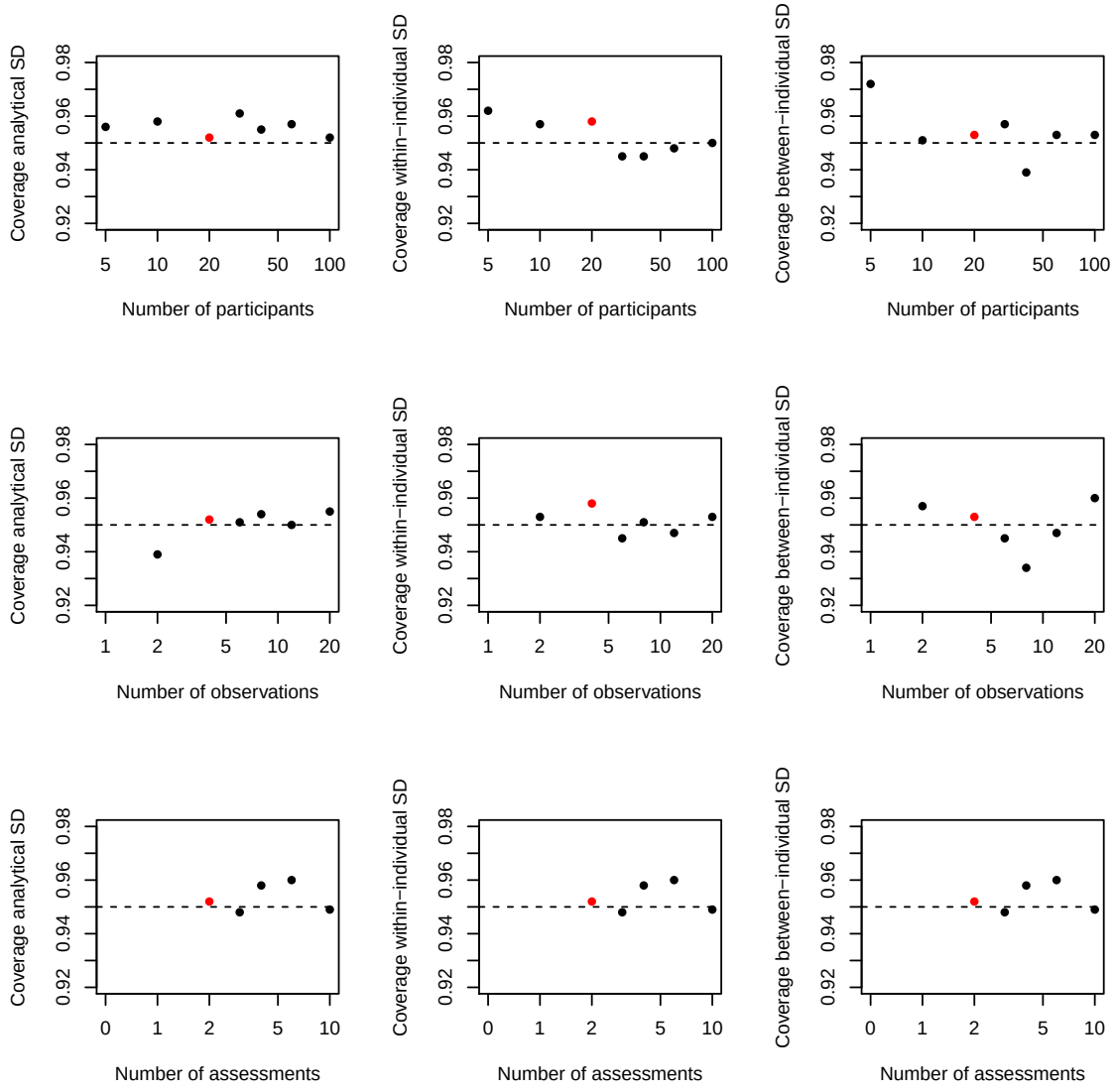


Figure 5.4: Coverage estimates from biological variability data simulations varying sample size: coverage of SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying number of participants (n_1, top row); number of observations per participant (n_2, middle row); and, number of replicate assessments per observation per participant (n_3, bottom row). 95% coverage is shown by horizontal line. Estimates shown in red are for the baseline strategy. 95% confidence intervals were calculated using the methods of Burdick and Graybill.[34] Confidence intervals could not be calculated for all scenarios, see Table 5.4.

Table 5.3: Biological variability study sample size simulation results–bias performance measures varying number of participants, observations and assessments.

Inputs						Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	0.5	1	2	-28.186	-183.365	-1450.061	-0.564	-1.834	-7.250	-3.698	-8.783	-20.212
10	4	2	0.5	1	2	-31.132	-32.756	-648.276	-0.623	-0.328	-3.241	-5.697	-2.299	-12.768
20	4	2	0.5	1	2	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
30	4	2	0.5	1	2	-6.397	-61.652	-169.784	-0.128	-0.617	-0.849	-2.067	-7.232	-5.973
40	4	2	0.5	1	2	-5.405	-15.610	-140.272	-0.108	-0.156	-0.701	-1.968	-2.076	-5.643
60	4	2	0.5	1	2	-8.585	7.622	-62.797	-0.172	0.076	-0.314	-3.861	1.269	-3.290
100	4	2	0.5	1	2	3.608	-12.721	-2.366	0.072	-0.127	-0.012	2.070	-2.791	-0.154
20	**2**	2	0.5	1	2	-15.217	-135.668	-269.144	-0.304	-1.357	-1.346	-2.688	-7.895	-7.199
20	**4**	2	0.5	1	2	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	**6**	2	0.5	1	2	-14.345	-8.598	-175.700	-0.287	-0.086	-0.878	-4.517	-1.073	-5.084
20	**8**	2	0.5	1	2	-11.513	-8.287	-422.331	-0.230	-0.083	-2.112	-4.158	-1.237	-12.109
20	**12**	2	0.5	1	2	-1.105	-35.497	-115.551	-0.022	-0.355	-0.578	-0.494	-6.432	-3.481
20	**20**	2	0.5	1	2	1.062	-1.533	-387.691	0.021	-0.015	-1.938	0.618	-0.380	-12.474
20	4	**2**	0.5	1	2	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	**3**	0.5	1	2	5.704	-53.911	-195.607	0.114	-0.539	-0.978	1.996	-5.384	-5.597
20	4	**4**	0.5	1	2	-1.822	-56.582	-83.150	-0.036	-0.566	-0.416	-0.795	-5.777	-2.302
20	4	**6**	0.5	1	2	-2.362	-39.539	-216.487	-0.047	-0.395	-1.082	-1.359	-4.260	-6.157
20	4	**10**	0.5	1	2	0.326	-49.668	-265.252	0.007	-0.497	-1.326	0.253	-5.474	-7.677

Table 5.4: Biological variability study sample size simulation results–accuracy and coverage performance measures varying number of participants, observations and assessments.

Inputs						**Accuracy and coverage**								
						Mean squared error ($\times 10^{-4}$)			**Coverage**			**Mean 95% CI width**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	0.5	1	2	58.189	439.191	5357.484	0.956	0.962[a]	0.972[b]	0.338	0.917[a]	4.763[b]
10	4	2	0.5	1	2	29.962	203.037	2620.135	0.958	0.957	0.951[c]	0.228	0.608	2.378[c]
20	4	2	0.5	1	2	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
30	4	2	0.5	1	2	9.582	73.055	810.902	0.961	0.945	0.957	0.128	0.336	1.165
40	4	2	0.5	1	2	7.548	56.538	619.875	0.955	0.945	0.939	0.111	0.290	0.990
60	4	2	0.5	1	2	4.950	36.077	364.751	0.957	0.948	0.953	0.090	0.236	0.795
100	4	2	0.5	1	2	3.041	20.784	235.163	0.952	0.950	0.953	0.070	0.182	0.608
20	**2**	2	0.5	1	2	32.064	297.101	1404.812	0.939	0.953	0.957[d]	0.229	0.761	1.620[d]
20	**4**	2	0.5	1	2	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
20	**6**	2	0.5	1	2	10.108	64.244	1197.323	0.951	0.945	0.945	0.128	0.319	1.454
20	**8**	2	0.5	1	2	7.679	44.863	1234.344	0.954	0.951	0.934	0.111	0.268	1.421
20	**12**	2	0.5	1	2	4.999	30.580	1103.206	0.950	0.947	0.947	0.090	0.213	1.424
20	**20**	2	0.5	1	2	2.957	16.290	980.965	0.955	0.953	0.960	0.070	0.162	1.393
20	4	**2**	0.5	1	2	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
20	4	**3**	0.5	1	2	8.167	100.568	1225.333	0.948	0.941	0.954	0.111	0.398	1.483
20	4	**4**	0.5	1	2	5.258	96.263	1305.137	0.958	0.949	0.934	0.090	0.390	1.488
20	4	**6**	0.5	1	2	3.022	86.285	1241.073	0.960	0.958	0.942	0.070	0.383	1.478
20	4	**10**	0.5	1	2	1.666	82.568	1200.736	0.949	0.956	0.944	0.052	0.376	1.473

[a]12 CIs could not be calculated; [b]78 CIs could not be calculated; [c]4 CIs could not be calculated; [d]2 CIs could not be calculated.

Estimates of index of individuality and reference change value

The median estimates of II and RCV were consistent and accurate for all numbers of participants, observations and assessments. As the number of participants and observations increased (n_1 and n_2), the range of estimates for both II and RCV decreased. With fewer participants ($n_1 = 5$) II was overestimated. With increased number of assessments, there was little change in the range of estimates of II and RCV, see Figure 5.7 and Appendix C Table C.3.

5.4.1.2 Analysis of log-normal data

When simulating log-normal data the CVs shown are exact geometric CVs and all CVs and RCVs are displayed as percentages.

When analysing the simulated log-normal data the bias in results along with the coverage was comparable with the normal data simulation. The log-normal data simulation yielded similar trends in estimated standard deviations, CVs, RCVs and IIs for increased sample sizes.

Confidence intervals were calculated for the estimates of CV directly using the equations of Burdick and Graybill.[34] These confidence intervals were appropriate compared to the simulated results and estimates of coverage were mainly within the expected range (coverage for one scenario when estimating CV_G was 93.4%), see figure 5.8.

The range of values from the 1,000 simulations were less varied for the log-normal simulation than the normal data simulation, see Appendix C Tables C.4 to C.9 and Figures C.1 and C.3.

5.4.2 Analytical, within-individual and between-individual variability

Results for scenarios are similar as only the chosen variability estimate is changed in the simulation; analyses are paired.

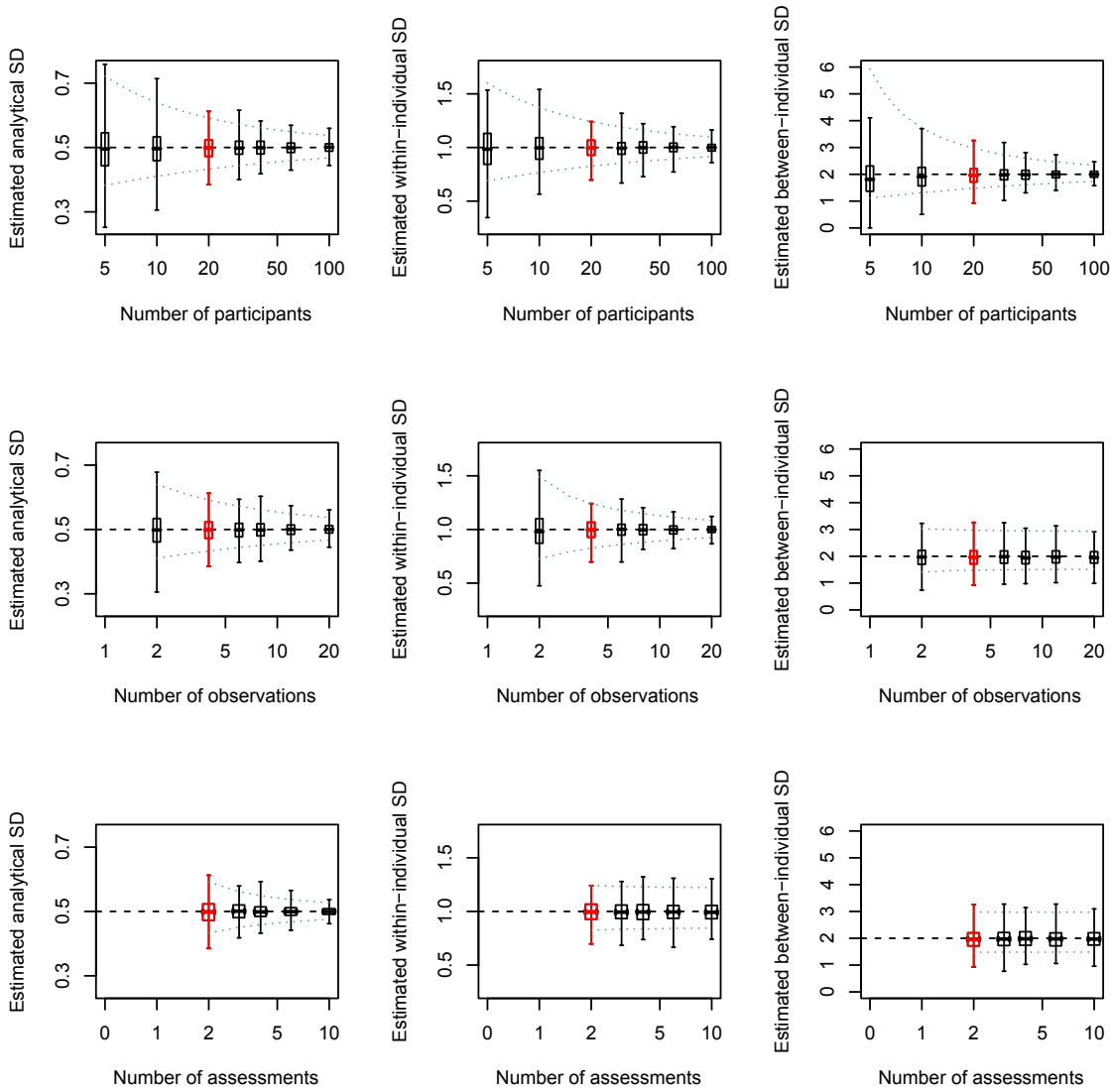


Figure 5.5: SD estimates from biological variability data simulations varying sample size: SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying number of participants (n_1, top row); number of observations per participant (n_2, middle row); and, number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true SD and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

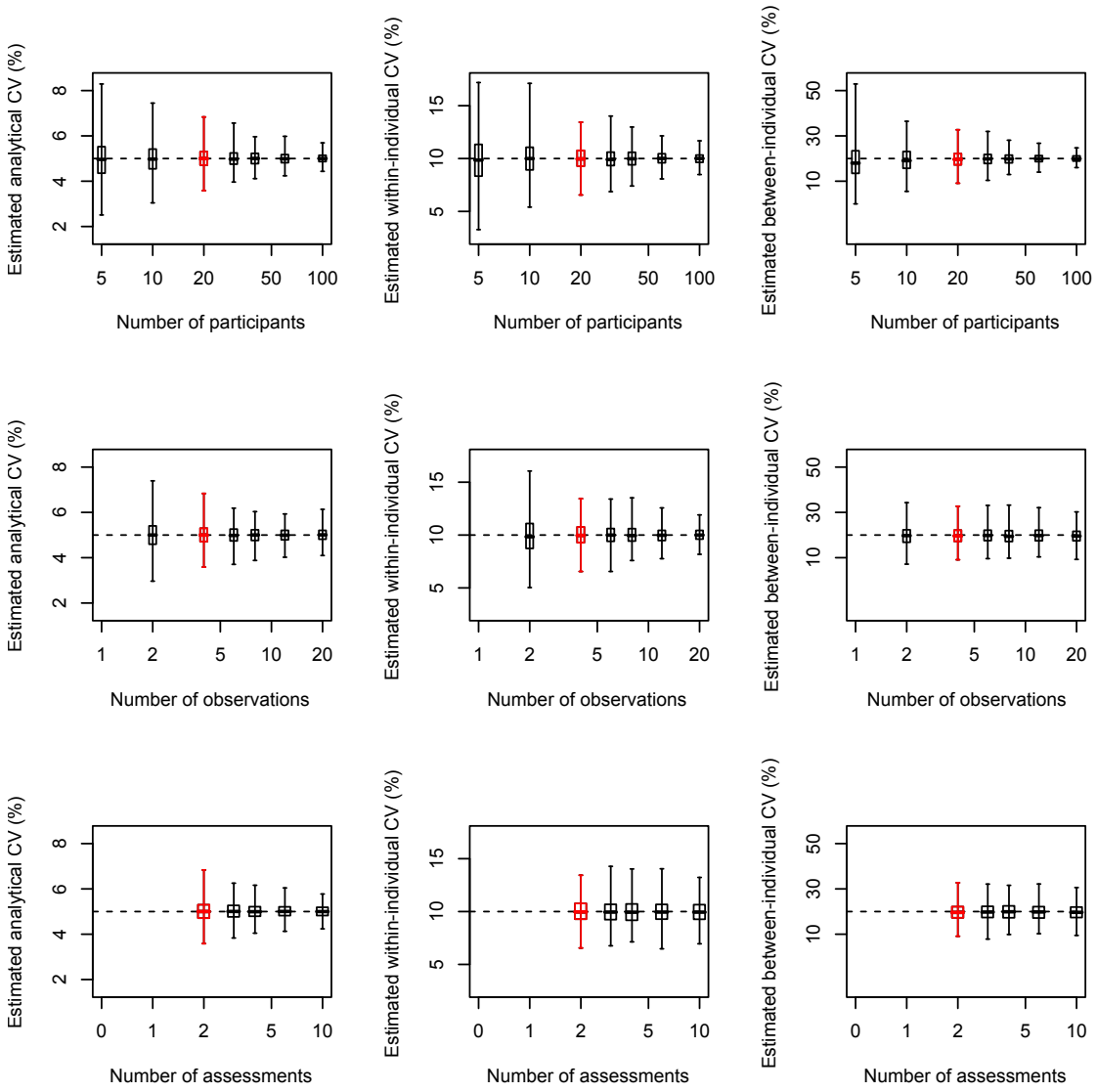


Figure 5.6: CV estimates biological variability data simulations varying sample size: CV_A (left column), CV_I (middle column) and CV_G (right column) estimates when varying number of participants (n_1, top row); number of observations per participant (n_2, middle row); and, number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true CV.

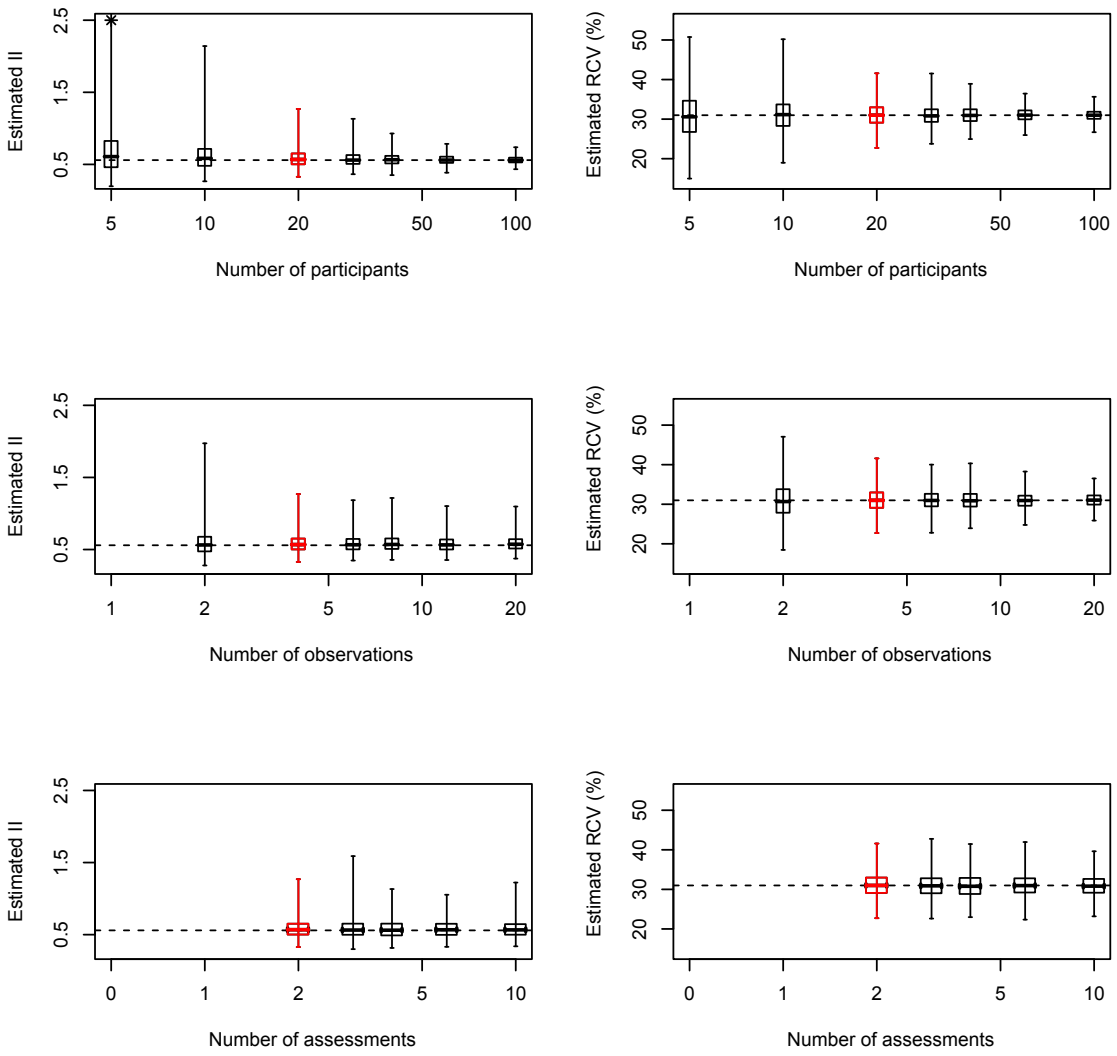


Figure 5.7: II and RCV estimates from biological variability data simulations varying sample size: II (left column) and RCV (right column) estimates when varying number of participants (n_1, top row); number of observations per participant (n_2, middle row); and, number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true II or RCV. *Maximum value not shown.

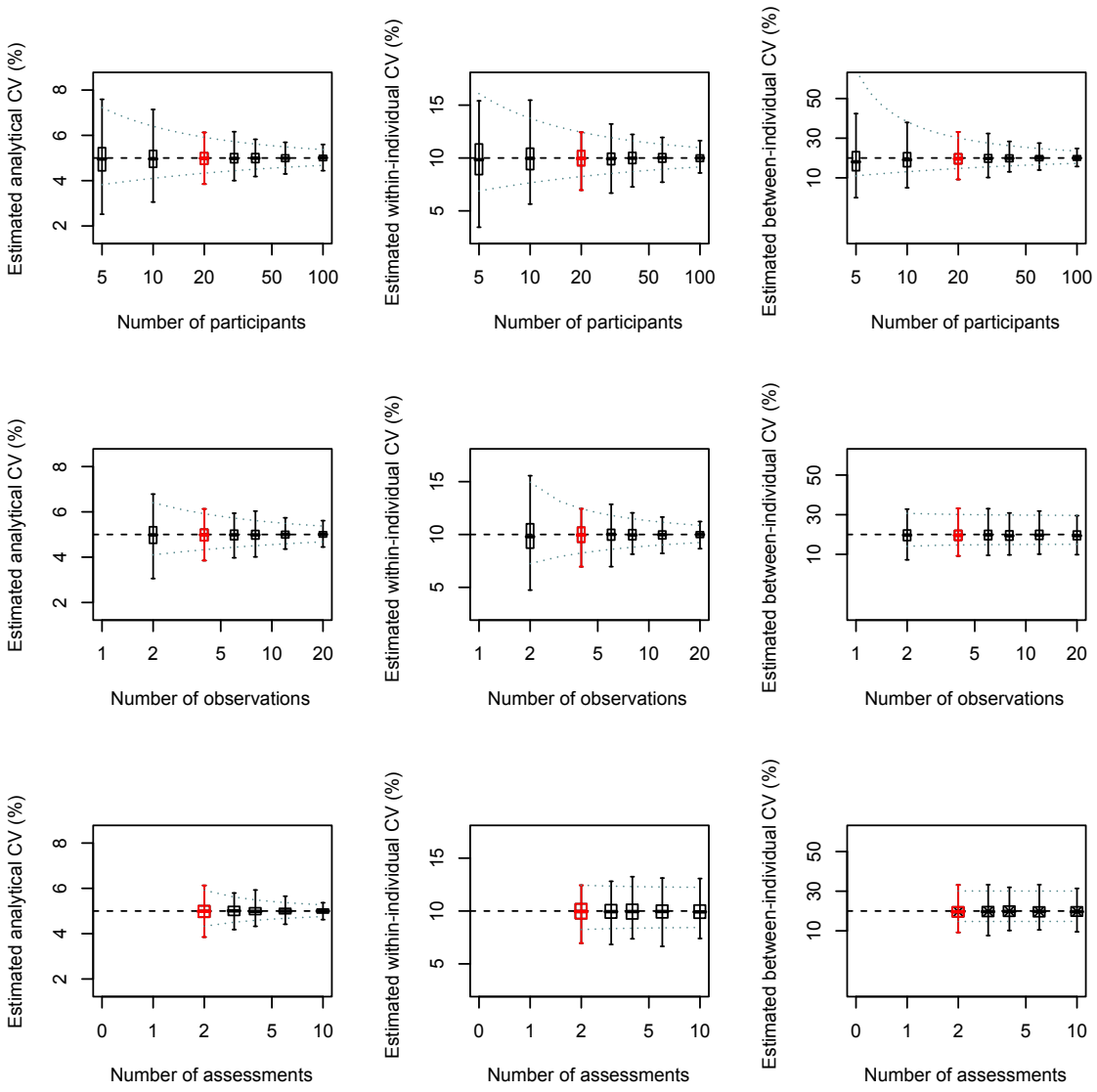


Figure 5.8: Log-normal biological variability sample size simulation: CV estimates from biological variability data simulations varying sample size: CV_A (left column), CV_I (middle column) and CV_G (right column) estimates when varying number of participants (n_1, top row); number of observations per participant (n_2, middle row); and, number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true CV and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

Bias

For σ_A, σ_I and σ_G the bias was negative for all simulated situations, see Table 5.5. The bias was less than 2% for each estimated standard deviation for most scenarios; for three of the simulated scenarios the bias was greater than 2%: 1) $\sigma_A = 0.5$, $\sigma_I = 1.75$ and $\sigma_G = 2$; 2) $\sigma_A = 0.5$, $\sigma_I = 2$ and $\sigma_G = 2$; and 3) $\sigma_A = 0.5$, $\sigma_I = 1$ and $\sigma_G = 1$. With increased analytical variability (CV_A) the bias for σ_I increased; with increased within-individual variability (CV_I) the bias for σ_A and σ_I increased; and for increased between-individual variability (CV_G) the bias for σ_G increased.

Confidence intervals

As the analytical variability (CV_A) increased the mean width of 95% confidence intervals for σ_A, σ_I and σ_G increased; as the within-individual variability (CV_I) increased the width of 95% confidence intervals for σ_I and σ_G increased (with the width of 95% confidence intervals for σ_A unchanged); and for increases in between-individual variability (CV_G) the mean width of the 95% confidence intervals for only σ_G increased in width (with the width of 95% confidence intervals for σ_A and σ_I were constant). The coverage of 95% confidence intervals for estimates of σ_A, σ_I and σ_G was close to 95% and greater than the expected lower bound given the number of simulations, see Table 5.6 and Figure 5.9. For some of the larger variability estimates 95% confidence intervals could not be calculated. Again, comparison of the results from the simulated data to the bounds produced by the Burdick and Graybill confidence intervals (see Figure 5.10) suggest the lower bound is underestimated.

Estimates of standard deviations and coefficients of variation

The median estimate of σ_A, σ_I and σ_G and CV_A, CV_I and CV_G changed as expected for each level of test variability. For increases in analytical variation (CV_A) the range of estimates for σ_A and σ_I (and therefore CV_A, and CV_I) increased, with the range of estimates for σ_G (CV_G) consistent. For increases in within-individual variation (CV_I) the range of estimates for σ_I and σ_G (and CV_I and CV_G) increased; however for all values of within-individual variation, the range of estimates of σ_A (and CV_A) appeared constant. For increases in between-individual variation, the range of estimates of σ_G (and CV_G) increased, with the estimates for σ_A and

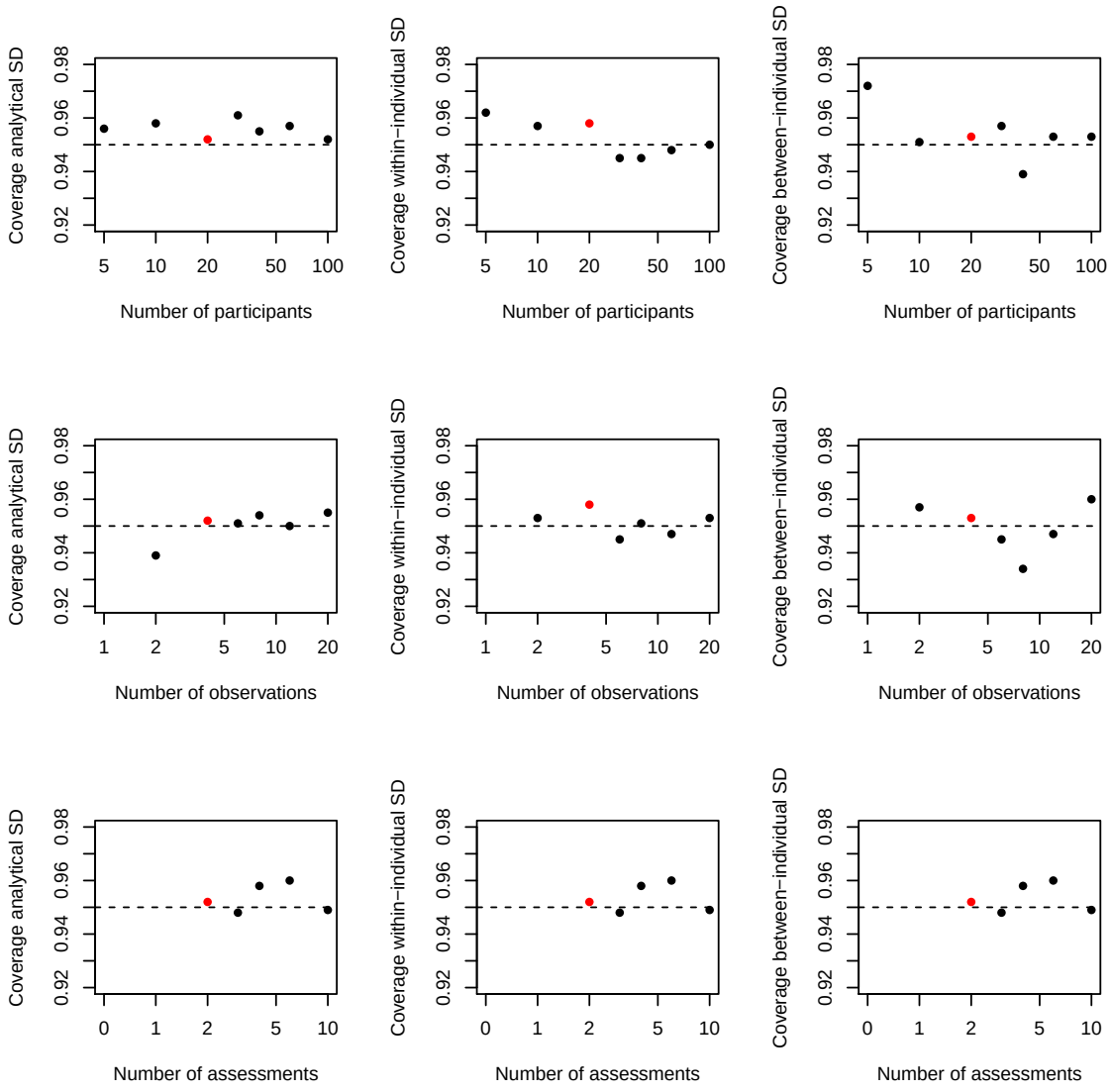


Figure 5.9: Coverage estimates from biological variability data simulations varying SD_A, SD_I and SD_G: coverage of SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying value of SD_A (top row); value of SD_I (middle row); and, value of SD_G (bottom row). 95% coverage is shown by horizontal line. Estimates shown in red are for the baseline strategy. 95% confidence intervals were calculated using the methods of Burdick and Graybill.[34] Confidence intervals could not be calculated for all scenarios, see Table 5.6.

Table 5.5: Biological variability study sample size simulation results–bias performance measures varying CV_A, CV_I and CV_G.

			Inputs			**Bias**								
						Bias ($\times 10^{-4}$)			**Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	1	2	-2.190	-27.958	-308.932	-0.175	-0.280	-1.545	-2.274	-3.092	-9.004
20	4	2	**0.25**	1	2	-4.373	-28.303	-307.152	-0.175	-0.283	-1.536	-2.271	-3.073	-8.942
20	4	2	**0.375**	1	2	-6.553	-29.954	-306.096	-0.175	-0.300	-1.530	-2.269	-3.139	-8.891
20	4	2	**0.5**	1	2	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	**0.625**	1	2	-10.934	-38.495	-306.031	-0.175	-0.385	-1.530	-2.271	-3.576	-8.819
20	4	2	**0.75**	1	2	-13.126	-46.412	-307.109	-0.175	-0.464	-1.536	-2.272	-3.974	-8.801
20	4	2	**0.875**	1	2	-15.303	-57.747	-308.971	-0.175	-0.577	-1.545	-2.270	-4.507	-8.796
20	4	2	**1**	1	2	-17.485	-73.477	-311.615	-0.175	-0.735	-1.558	-2.270	-5.179	-8.804
20	4	2	0.5	**0.5**	2	-8.745	-36.733	-254.363	-0.175	-0.735	-1.272	-2.270	-5.178	-7.659
20	4	2	0.5	**0.75**	2	-8.750	-30.603	-276.900	-0.175	-0.408	-1.385	-2.272	-3.695	-8.207
20	4	2	0.5	**1**	2	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	0.5	**1.25**	2	-8.738	-38.069	-341.382	-0.175	-0.305	-1.707	-2.269	-3.162	-9.579
20	4	2	0.5	**1.5**	2	-8.739	-43.870	-384.610	-0.175	-0.292	-1.923	-2.269	-3.107	-10.384
20	4	2	0.5	**1.75**	2	-8.740	-50.137	-436.704	-0.175	-0.286	-2.184	-2.269	-3.085	-11.264
20	4	2	0.5	**2**	2	-8.739	-56.663	-499.387	-0.175	-0.283	-2.497	-2.269	-3.076	-12.220
20	4	2	0.5	1	**1**	-8.742	-33.204	-252.994	-0.175	-0.332	-2.530	-2.270	-3.302	-12.133
20	4	2	0.5	1	**1.5**	-8.741	-33.198	-266.524	-0.175	-0.332	-1.777	-2.270	-3.301	-9.795
20	4	2	0.5	1	**2**	-8.740	-33.200	-305.758	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	0.5	1	**2.5**	-8.739	-33.207	-352.932	-0.175	-0.332	-1.412	-2.269	-3.302	-8.358
20	4	2	0.5	1	**3**	-8.738	-33.214	-403.707	-0.175	-0.332	-1.346	-2.269	-3.303	-8.063
20	4	2	0.5	1	**3.5**	-8.736	-33.220	-456.451	-0.175	-0.332	-1.304	-2.268	-3.303	-7.870
20	4	2	0.5	1	**4**	-8.735	-33.226	-510.387	-0.175	-0.332	-1.276	-2.268	-3.304	-7.734
20	4	2	0.5	1	**4.5**	-8.733	-33.231	-565.107	-0.175	-0.332	-1.256	-2.267	-3.304	-7.634
20	4	2	0.5	1	**5**	-8.731	-33.236	-620.366	-0.175	-0.332	-1.241	-2.267	-3.305	-7.559

Table 5.6: Biological variability study sample size simulation results–accuracy and coverage performance measures varying CV_A, CV_I and CV_G.

						Accuracy and coverage								
			Inputs			Mean squared error ($\times 10^{-4}$)			Coverage			Mean 95% CI width		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	1	2	0.928	81.861	1186.628	0.952	0.957	0.951	0.040	0.371	1.469
20	4	2	**0.25**	1	2	3.711	84.913	1189.362	0.952	0.959	0.951	0.079	0.379	1.471
20	4	2	**0.375**	1	2	8.348	91.166	1194.741	0.952	0.959	0.954	0.119	0.394	1.475
20	4	2	**0.5**	1	2	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
20	4	2	**0.625**	1	2	23.188	116.003	1213.481	0.952	0.955	0.955	0.198	0.446	1.486
20	4	2	**0.75**	1	2	33.393	136.584	1226.982	0.952	0.956	0.954	0.237	0.485	1.494
20	4	2	**0.875**	1	2	45.456	164.504	1243.306	0.952	0.953	0.954	0.277	0.534	1.504
20	4	2	**1**	1	2	59.364	201.826	1262.531	0.952	0.961[a]	0.954	0.316	0.594[a]	1.515
20	4	2	0.5	**0.5**	2	14.842	50.459	1109.339	0.952	0.961[b]	0.950	0.158	0.297[b]	1.415
20	4	2	0.5	**0.75**	2	14.843	68.705	1146.057	0.952	0.955	0.951	0.158	0.344	1.442
20	4	2	0.5	**1**	2	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
20	4	2	0.5	**1.25**	2	14.840	145.084	1281.835	0.952	0.959	0.959	0.158	0.498	1.530
20	4	2	0.5	**1.5**	2	14.841	199.555	1386.733	0.952	0.962	0.961	0.158	0.583	1.595
20	4	2	0.5	**1.75**	2	14.843	264.430	1522.234	0.952	0.958	0.962[c]	0.158	0.670	1.675[c]
20	4	2	0.5	**2**	2	14.843	339.624	1694.932	0.952	0.959	0.971[d]	0.158	0.759	1.775[d]
20	4	2	0.5	1	**1**	14.842	101.248	441.166	0.952	0.958	0.971[e]	0.158	0.416	0.909[e]
20	4	2	0.5	1	**1.5**	14.842	101.246	747.480	0.952	0.958	0.958	0.158	0.416	1.170
20	4	2	0.5	1	**2**	14.843	101.247	1202.760	0.952	0.958	0.953	0.158	0.416	1.480
20	4	2	0.5	1	**2.5**	14.842	101.247	1795.390	0.952	0.958	0.954	0.158	0.416	1.805
20	4	2	0.5	1	**3**	14.843	101.245	2523.062	0.952	0.958	0.952	0.158	0.416	2.138
20	4	2	0.5	1	**3.5**	14.841	101.247	3385.033	0.952	0.958	0.950	0.158	0.416	2.475
20	4	2	0.5	1	**4**	14.843	101.250	4381.025	0.952	0.958	0.952	0.158	0.416	2.814
20	4	2	0.5	1	**4.5**	14.842	101.246	5510.924	0.952	0.958	0.952	0.158	0.416	3.154
20	4	2	0.5	1	**5**	14.841	101.246	6774.654	0.952	0.958	0.952	0.158	0.416	3.496

[a]8 CIs could not be calculated; [b]8 CIs could not be calculated; [c]2 CIs could not be calculated; [d]10 CIs could not be calculated;
[e]18CIs could not be calculated.

σ_I (CV_A and CV_I) constant, see Figures 5.10 and 5.11 and Appendix C Tables C.10 and C.11.

Estimates of index of individuality and reference change value

The median estimates of II and RCV were consistent and accurate for all test variation values. As the analytical variation increased, there was little change in the range of estimates for both II and RCV. As the within-individual variation increased, the range of estimates for both II and RCV increased. The range of estimates for II, decreased with increases in between-individual variability; whereas the range of estimates for RCV increased with increases in between-individual variability, see Figure 5.12 and Appendix C Table C.12.

5.4.2.1 Analysis of log-normal data

Again similar trends were seen when simulating log-normal data and changing the estimates of variability. See Figures C.4 to C.7 and Tables C.13 to C.18.

5.4.3 Sensitivity analyses

When increasing the base case number of participants in the simulation ($n_1 = 40$, $n_2 = 10$ and $n_3 = 3$) and the base case test variability ($\sigma_A = 1$, $\sigma_I = 2$ and $\sigma_G = 4$) the observed trends remained the same. For simulations with the base case number of participants, observations and assessments increased, the range of results decreased from the original simulation, see Appendix C Tables C.19 to C.28. When the base case reflected a test with increased variability the range of results increased, see Appendix C Tables C.29 to C.38.

5.5 Discussion

The trends seen when varying factors of the simulation (number of participants, observations and assessments, and test variability at the analytical, within-individual and between-individual level) are intuitive but allow planning given the specific purpose of the study and

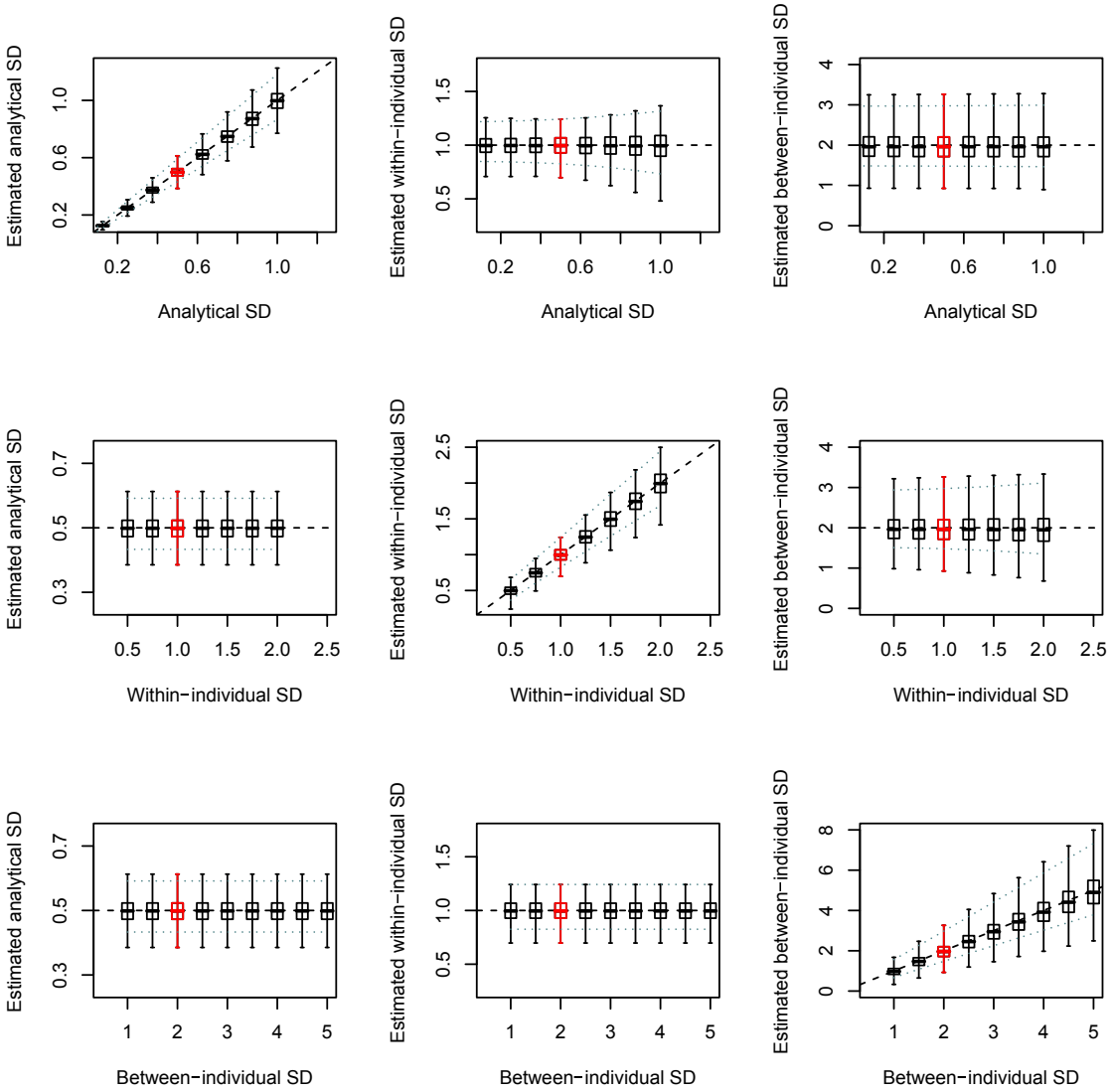


Figure 5.10: SD estimates from biological variability data simulations varying test variability: SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying value of SD_A (top row); value of SD_I (middle row); and, value of SD_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true SD and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

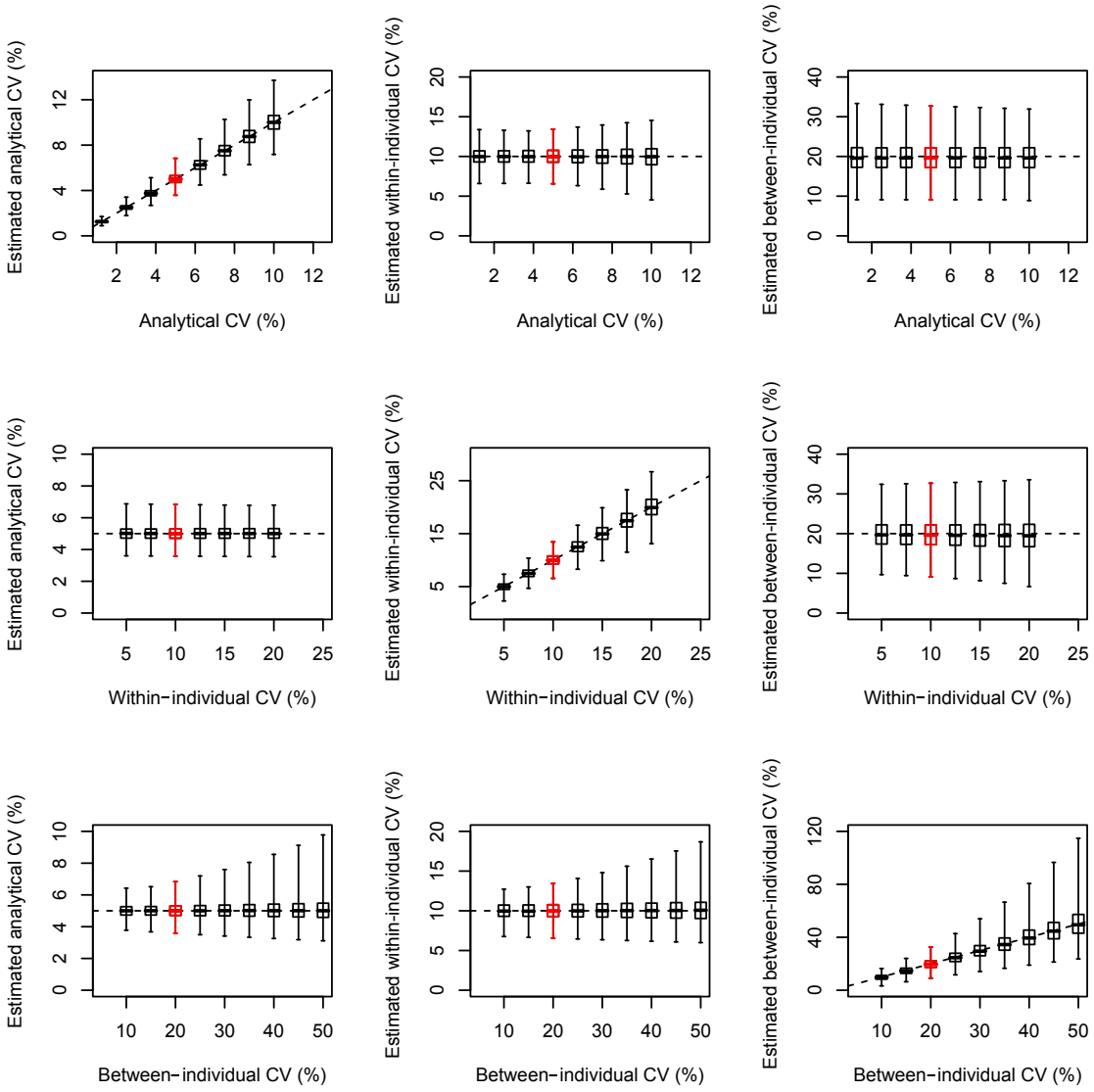


Figure 5.11: CV estimates from biological variability data simulations varying test variability: CV_A (left column), CV_I (middle column) and CV_G (right column) estimates when varying value of CV_A (top row); value of CV_I (middle row); and, value of CV_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true CV.

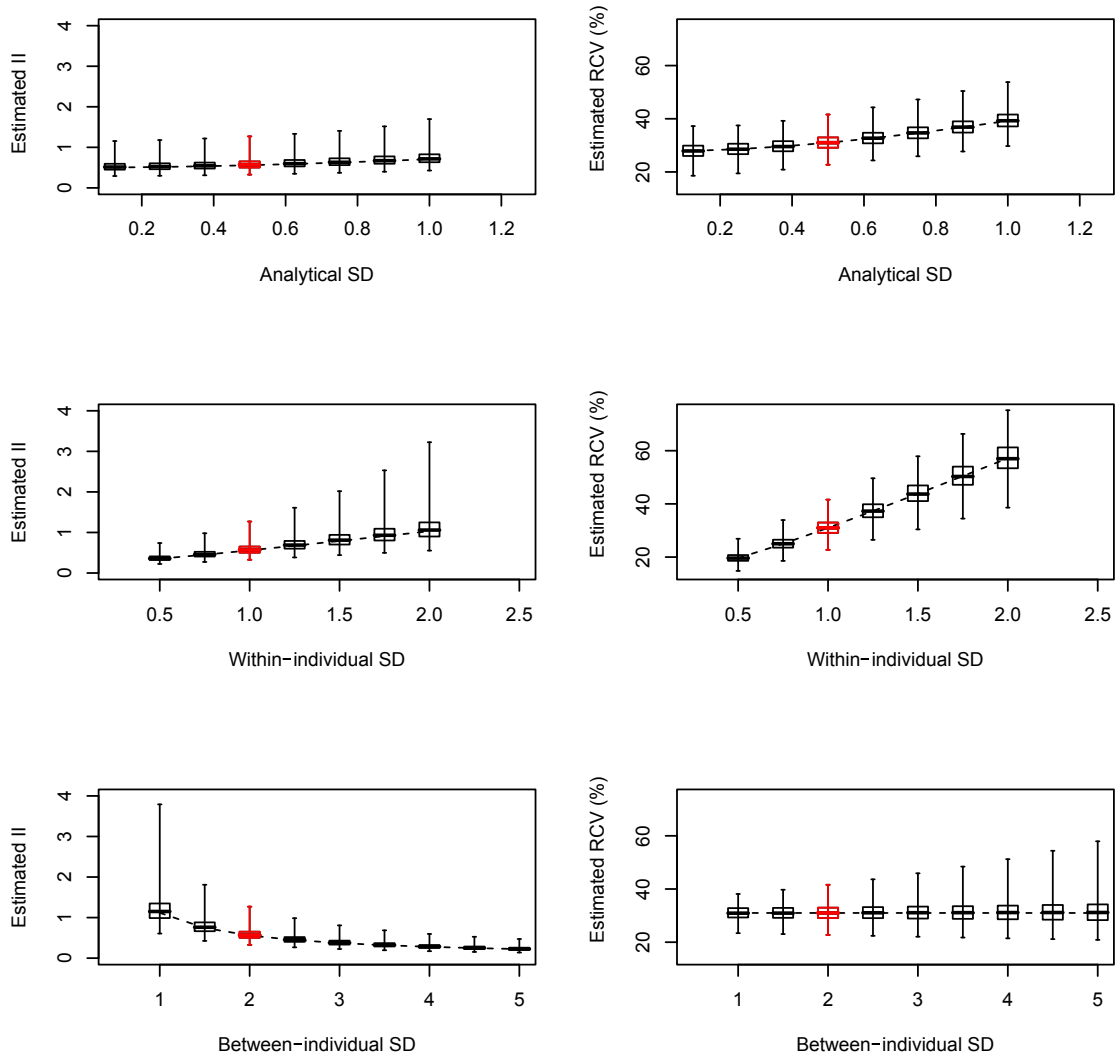


Figure 5.12: II and RCV estimates from biological variability data simulations varying test variability: II (left column) and RCV (right column) estimates when varying value of CV_A (top row); value of CV_I (middle row); and, value of CV_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true II or RCV.

the estimate required, with resource used to give precision of a specific estimate. The negative bias (underestimation) of the variance parameters was expected with the small sample size.[50,107] The bias was less than 2% for scenarios with at least 20 participants. For studies with fewer than 20 participants (of which many were identified, see Chapter 3) the bias may have a large enough impact on results; results from small studies may not be valid and should be interpreted with caution.

The Burdick and Graybill method for calculating confidence intervals performed well with coverage estimates in the expected range for most simulated scenarios. Comparison of results from simulations to estimates of confidence intervals using the equation suggest the lower bound of the interval is consistently underestimated for all sample sizes and variability estimates investigated. This is perhaps due the restriction on the lower bound (the standard deviation cannot take a negative value).

Increases in the number of participants increased precision of all estimates of variability; whereas increases in the number of observations increased precision of estimates of variability at the analytical and within-individual level only, and increasing the number of assessments increased precision of analytical variability estimates only.

The index of individuality (II) includes the coefficient of variation (CV) at the analytical, within-individual and between-individual level with each contributing to the estimate obtained. However, as analytical variation is small this has little impact on the estimate of II. If interest is in the estimate of II, precision of the estimates of variability at the within-individual and between-individual level should be prioritised by increasing the sample size via increasing the number of participants and observations for each participant.

The measure of reference change value (RCV) includes the variability estimates at the analytical and within-individual levels. Again, the variability at the analytical level is generally small and the level of precision of the variability at the analytical level has little impact on the estimate of RCV. Increasing the number of observations per participant will increase the precision of the estimate of within-individual variability and analytical variability. Interestingly, increases in the number of participants increases the precision of RCV also as increases

in the number of participants helps increase the precision of estimates of variability at the analytical, within-individual and between-individual levels.

5.5.1 Limitations

The simulation method used assumes test data are normally distributed or log-normally distributed. Any data that are not distributed in this way may require transformation prior to analysis as the methods used assume this distribution. The sample size tool, like standard sample size calculation tools, requires estimates of variability which may be unknown.

Simulation results are shown for the combinations of variability and sample size that have been selected to demonstrate trends based on the sample sizes and results seen in the review of biological variability studies (see Chapter 3). The results displayed may not reflect results for all combinations of variability and sample size, which would need to be considered prior to planning a study.

Both the normal and log-normal simulation methods used assume the errors have no bias. This was a necessary assumption, allowing the comparison of input values to estimated results given the methods used.

When simulating data for more extreme scenarios (large sample sizes and large variability) the random-effects model failed to converge for a minority of the 1,000 simulations and the result was forced without convergence.

5.6 Conclusions

When sample sizes of at least 20 participants (with repeated observations and assessments) are used the methods to generate estimates of analytical, within-individual and between-individual variability appear valid, with small negative bias. With fewer than 20 participants the bias may impact results and results from studies should be interpreted with caution. The methods to generate 95% confidence intervals for biological variability estimates were

valid.

The use of the specified tool allows simulation of biological variability studies using different sample sizes, for various estimates of variability. From analysing the simulated data and observing the range of results obtained the appropriate sample size required to ensure precision of a given estimate can be inferred.

Increasing the number of participants appears to benefit the precision of all estimates of variability (analytical, within-individual and between-individual) and this subsequently improves the precision of estimates of commonly reported measures of biological variability of CV_A, CV_I and CV_G, and also II and RCV.

Increasing the number of observations for each participant has a positive impact on the precision of estimates of analytical and within-individual variability; this improves precision of the estimates of II and RCV also, but not to the magnitude of increasing the number of participants. Increasing the number of assessments of observations for each individual increases precision of estimates of analytical variability; this improves precision of CV_A but has little impact on measures of II and RCV.

Chapter 6

The impact of outlier detection and removal on studies of biological variability

This work has been partly presented in the following form:
Sitch A, Mallett S, Deeks J. The impact of outlier detection and removal on studies of biological variability (BV). Methods for Evaluation of medical prediction Models, Tests And Biomarkers (MEMTAB), Utrecht, Netherlands. 2-3 July 2018.

Summary

Outlier detection methods are frequently used in the analysis of laboratory based studies of biological variability. The methods for detecting outliers to be removed from a data set prior to evaluation vary. Some methods account for the structure of data (multiple assessments of multiple observations of multiple participants) by comparing variances, and others look at all

data points evaluating differences in measurements in terms of the range, interquartile range or standard deviation with comparison to fixed values or critical values based on distributional assumptions.

Simulation was used to compare the results of analyses using different methods to identify outliers, with these data points removed prior to analysis.

The simulation showed that when outlier detection is used (when data were log-normally distributed and had been log-transformed) and identified values are removed the resulting estimates of analytical, within-individual and between-individual variability are underestimated. The bias in these estimates was greatest when using a detection strategy involving the Cochran C test and Tukey's IQR rule. When data were simulated to include outlying measurements, different outlier detection methods worked best depending on the number of outliers present. The nature of outliers must be understood before an appropriate method can be identified.

6.1 Introduction

In biological variability studies of laboratory tests, it is considered best practice to use methods to identify outliers.[13] Identified data are removed prior to obtaining estimates of variability. Outliers can have a large impact on estimates of variability, especially in small data sets.[108] The methods commonly used are Cochran C test, Reed's criterion, Dixon's test, Grubbs's test, the Tukey IQR rule and checking values are within three standard deviations, see Chapter 3. These methods detect data as outlying by assessing the variance of values in subsets, or evaluating the range, interquartile range or standard deviation with comparison to fixed values or critical values based on distributional assumptions

Aguinis et al[109] offered an extensive review of outlier detection methods defining, identifying and handling outliers. This review considered 14 outlier definitions, 39 outlier identification methods and 20 approaches to handling outliers. Outlier definitions leave three main types of outliers to be considered; 'error outliers', 'interesting outliers' and 'influential outliers'.

'Error outliers' are inaccurate data points; 'interesting outliers' are data points outside of the usual range but offer key information; and, 'influential outliers' are data points affecting model fitting and prediction. The authors offer decision making trees depending on the type of outliers you are looking to identify and the analysis used. The recommendations are: for 'error outliers' remove the data and carefully report this; for 'interesting outliers' further study is appropriate; and, for 'influential outliers' analyses should be performed and reported with and without these outlying measures.

By recommending testing for outliers when analysing biological variability studies[35,69] there is a danger of identifying and removing test values which are plausible (potential interesting and/or influential outliers) rather than only erroneous data ('error outliers'), which would be appropriate and in most circumstances could be identified using basic data descriptions and clinical knowledge. Outlier detection and removal is performed as researchers believe this process is beneficial with the estimates obtained after analysis a better reflection of the truth than if the 'outlying' data were to remain.[13] However, the practice of detecting and removing outliers may lead to underestimated variability.

The Cochran C test is criticised for only applying to data with equal groups (balanced design) and not being two-sided (only large variances are identified).[99] Methods relying on detecting differences from the mean in terms of standard deviations are criticised as the mean and the standard deviation can be strongly influenced by outliers, they require normality of the data and perform poorly in small samples.[110]

A variety of methods are used to identify outliers in biological variability studies, see Chapter 3, with combinations of outlier detection methods used also. Some tests and testing strategies are specific to the study design and account for the clustering of data, identifying outliers at each level, and others simply look at the range of the data in comparison to distributional assumptions. The aim is to understand the impact of using outlier detection methods on the analysis of biological variability studies comparing no outlier detection and removal to different methods of outlier detection and removal.

6.2 Aims and objectives

The aim of this work is to understand and evaluate the impact of outlier detection and removal on obtaining accurate estimates for biological variability studies. The difference between methods was evaluated, considering data without outliers and data with 'error' outliers.

The main research questions were:

- What are the differences between outlier detection methods when simulating data without outliers? How many data points are unnecessarily and inappropriately removed with each method and what is the consequence?

- What are the differences between outlier detection methods when simulating data with 'error' outliers? How many data points are correctly removed and how many are unnecessarily removed in addition to these for each method, and what is the consequence?

The specific objectives were to:

- investigate the difference in the number of measurements detected and removed (correctly and incorrectly) using different methods;

- understand the impact of outlier detection methods on the estimated standard deviations;

- and, evaluate the impact of outlier detection methods on estimates of CV, II and RCV.

The methods of outlier detection evaluated were the Cochran C test, Reed's criterion, the Fraser-Harris method,[35] the Tukey IQR rule, Dixon's Q test, Grubbs's test and restricting the data to be within three standard deviations of the mean. These methods were identified when reviewing biological variabilty studies, see Chapter 3.

6.3 Methods

Biological variability data was simulated and tests for outliers were then used. The detected outliers were removed and the remaining data were analysed.

6.3.1 Data simulation

6.3.1.1 No outlying data simulation

Log-normally distributed data were simulated. Data were simulated following the model $y_{ijk} = \mu \times \alpha_i \times \beta_{ij} \times \varepsilon_{ijk}$, thus $ln(y_{ijk}) = ln(\mu) + ln(\alpha_i) + ln(\beta_{ij}) + ln(\varepsilon_{ijk})$. Data were simulated where $ln(\alpha_i) \sim N(0, \sigma_G^2)$, $ln(\beta_{ij}) \sim N(0, \sigma_I^2)$, $ln(\varepsilon_{ijk}) \sim N(0, \sigma_A^2)$ and $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$.

The simulation was performed with the sample size of 20 participants (n_1), four observation points per participant (n_2), and two assessments of each observation (n_3). Variability estimates were fixed at $CV_A = 5\%$, $CV_I = 10\%$, and $CV_G = 20\%$, and sensitivity analyses were performed for a 'poor performing' test with $CV_A = 7.5\%$, $CV_I = 15\%$, and $CV_G = 30\%$ (increasing each CV separately) and increasing the sample size ($n_1 = 40, n_2 = 4$ and $n_3 = 2$).

See Chapter 5 (§5.3.3.3) for further details on the data simulation.

6.3.1.2 Outlying data simulation

Data were also simulated with a percentage of measurements randomly changed by a factor of 10 or two (multiplied or divided) to give the effect of outliers due to a missed digit or lab error. Simulations were performed with 0.5%, 1% and 2% of data replaced (using a ceiling function to give 0.5%=1 measurement; 1%=2 measurements; and, 2%=4 measurements) and using the multipliers of 10 and two separately.

6.3.2 Outlier detection methods

A review of studies of biological variability (see Chapter 3) identified the reported methods for detecting outliers used in these analyses. The following methods were considered:

- no outlier detection;
- Cochran C test;
- Cochran C test partial;
- Fraser-Harris method (Cochran C test and Reed's criterion for means);
- Reed's criterion for means;
- Reed's criterion for measurements;
- Tukey IQR rule;
- Dixon's Q test;
- Grubbs's test;
- and, ± 3SD.

6.3.2.1 Cochran C test

Cochran C test for outliers is used in two ways for the purpose of biological variability data. Firstly the method is used to identify excessive variances for duplicate results within observations for an individual; and, secondly to identify excessive variances for measurements for individuals between participants.[16]

Cochran C test compares the variance of a subgroup of data points to the sum of variances across all subgroups (subgroups can be duplicate results or observations within individuals). This ratio of variances is then compared to a critical value (obtained using Fisher's F ratio).

If the critical value is exceeded the values within the subgroup generating the largest variance are deleted and the process is repeated until no variances exceed the critical value.[99]

Cochran C test for outliers compares the variance within subgroups to the sum of variances for all subgroups within the larger group. Cochran C values are calculated using:

$$C_j = \frac{\sigma_j^2}{\sum_{i=1}^{N} \sigma_i^2}$$

where there are n assessments within N subgroups. Cochran C values are calculated using the ratio of the variance of measures within each subgroup and the sum of the variances across all subgroups.[99]

For the first assessment, identifying outliers within duplicate assessments, for each individual σ_j^2 is the variance of the jth pair of observations and is divided by the sum of the variances for all duplicates for that individual.[16] For the second use of the Cochran C test, the variance of measures for each individual is divided by the sum of variances for all individuals to calculate the Cochran C value (for this the mean of duplicate assessments is used).

After calculating Cochran C values these are compared to a critical value defined by:

$$C_{UL} = \left[1 + \frac{(N-1)}{F_c(\alpha/N, (n-1), (N-1)(n-1))}\right]^{-1},$$

where F_c is the critical value from Fisher's F ratio. If the Cochran C value is greater than the critical value the data for the subgroup with Cochran C value exceeding the critical value is excluded. Strict use of the Cochran C method would mean exclusion of both duplicates in the first use of the test and all values for an individual in the second use of the test. The procedure of calculating Cochran C values is repeated and results evaluated until no values exceed the critical value.[99]

Partial use of the Cochran C test uses the test only once at each level to identify outliers, as stated by Fraser and Harris,[35] rather than being used repeatedly until no further outliers are detected. The test statistic was calculated using the R command 'C.test'.

6.3.2.2 Reed's criterion

Reed's criterion assesses the difference between the largest and next largest and smallest and next smallest measurements, with these extreme values deleted if the difference is greater than one-third of the range of the data.[16]

Say there are a set of measurements $x_1, \ldots, x_n$ where these measurements are arranged in increasing order, so $x_1 < x_2 < \ldots < x_{n-1} < x_n$. To assess if the minimum and maximum values are outliers, the criterion is:

$$x_2 - x_1 > \frac{1}{3}(x_n - x_1)$$
$$x_n - x_{n-1} > \frac{1}{3}(x_n - x_1).$$

Reed's criterion can also be applied to the mean values for each participant. For each participant, the mean of their measures is calculated. Say there are n_1 participants yielding n_1 mean values. These are $\bar{x}_1, \ldots, \bar{x}_n$ where these means are arranged in increasing order, so $\bar{x}_1 < \bar{x}_2 < \ldots < \bar{x}_{n_1-1} < \bar{x}_{n_1}$. To assess if the minimum and maximum means are outliers, the criterion is:

$$\bar{x}_2 - x_1 > \frac{1}{3}(\bar{x}_n - \bar{x}_1)$$
$$\bar{x}_n - x_{n-1} > \frac{1}{3}(\bar{x}_n - x_1).$$

If a mean for a participant is detected as an outlier, all values for that participant are removed.

6.3.2.3 Fraser-Harris method

Outlier detection and removal is advocated in the framework outlined by Fraser and Harris.[35] The suggested method involves the use of the Cochran C test firstly, to identify outliers at

the level of duplicate measurements and observations within individuals, followed by Reed's criterion using the mean test value for each individual.[16,35]

The use of the Cochran C test as explained by Fraser and Harris appears to be conservative in the second use of the test with only the set of duplicate measures appearing to contribute to the large variance being removed (partial use of the test) rather than all measurements for that individual.[35]

6.3.2.4 Tukey IQR rule

Tukey defined data as outlying if the data did not fall into a region defined using interquartile ranges. Values were considered to be outliers if they were less than the 25th percentile minus 1.5 times the interquartile range (the difference between the 75th and 25th percentile); or, if values were greater than the 75th percentile plus 1.5 times the interquartile range.[111]

If values did not fall into the following region (x) they were declared outliers:

$$25\text{th percentile} - (1.5 \times IQR) \leq x \leq 75\text{th percentile} + (1.5 \times IQR).$$

6.3.2.5 Dixon's Q test

Dixon's Q test was designed to identify a single outlier. The data points are ranked and the differences between consecutive measurements are calculated. These differences are compared to the range of all data to generate Q. The value of Q is then compared to a critical value (depending on the number of measurements and the confidence level); with the extreme value detected as an outlier and deleted if Q is greater than the critical value.[112]

Say there are a set of measurements $x_1, \ldots, x_n$ and these are arranged in increasing order, so $x_1 < x_2 < \ldots < x_{n-1} < x_n$. To assess if x_1 is an outlier Q is calculated as:

$$Q = \frac{x_2 - x_1}{x_n - x_1},$$

providing the ratio of the difference between neighbouring measurements and the range of the dataset. The calculated Q value is compared to a critical value obtained from tables (depending on the confidence level and the number of measurements). If Q is greater than the critical value an outlier has been detected and should be removed from the data set.[112]

Simulations used the 95% level and defaulted to the maximum number of values available if the data set contained more using the inbuilt function in R 'qdixon'. See Table 6.1 for extract from the Q tables.

Table 6.1: Dixon's Q tables.

Number of values:	3	4	5	6	7	8	9	10
$Q_{90\%}$:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
$Q_{95\%}$:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
$Q_{99\%}$:	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

6.3.2.6 Grubbs's test

Grubbs's test again aims to identify a single outlier. Grubbs's test calculates the difference between values and the sample mean in terms of the sample standard deviation. These values are then compared to a critical value (derived from the t distribution), with differences exceeding this critical value declared outliers and removed.[113]

Grubbs's test identifies the maximum difference between a measure (x_i) and the sample mean ($\bar{x}$) in terms of the sample standard deviation (σ). The test statistic is calculated as:

$$G = \frac{max|x_i - \bar{x}|}{\sigma}.$$

When conducting a two-sided test the critical value the G statistic is compared to is calculated as:

$$\frac{n-1}{\sqrt{n}}\sqrt{\frac{(t_{\alpha/2n,n-2})^2}{n-2+(t_{\alpha/2n,n-2})^2}},$$

where n is the sample size and $t_{\alpha/2n,n-2}$ denotes the critical value from the t distribution with $n-2$ degrees of freedom and significance level of $\alpha/2n$. The calculated value of G is compared to the critical value and if G exceeds this value the measure is identified as an

outlier and should be deleted.[113]

6.3.2.7 ± 3SD

The simplest outlier detection method calculates the sample standard deviation and values are declared as outliers and removed if they are not within three standard deviations of the mean.

Note that methods other than the Cochran C test and the Fraser Harris method do not account for the clustering of the data in biological variability studies.

6.3.3 Data analysis

As the simulated data were log-normally distributed, data were firstly log transformed and outlier detection was performed on the log transformed scale. All identified outliers were removed and a random effects model was fitted to the remaining data; this was a null model with random effects allowing for assessments within observations and observations within participants. $ln(y_{ijk}) = ln(\mu) + ln(\alpha_i) + ln(\beta_{ij}) + ln(\varepsilon_{ijk})$, where $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $k = 1, \ldots, n_3$, y_{ijk} is the test measure for the ith participant at the jth time point and for the kth assessment, μ is the mean value of the measure, $ln(\alpha_i) \sim N(0, \sigma_G^2)$, $ln(\beta_{ij}) \sim N(0, \sigma_I^2)$ and $ln(\varepsilon_{ijk}) \sim N(0, \sigma_A^2)$.

Fitting the model allowed estimates ($\hat{\sigma}_A$, $\hat{\sigma}_I$ and $\hat{\sigma}_G$) of σ_A, σ_I and σ_G to be obtained. The coefficient of variation (CV), index of individuality (II) and reference change values (RCV) were additionally produced using estimates from the model.

Exact geometric CVs were calculated using $\sqrt{exp(\hat{\sigma}^2) - 1}$[59,60] (assuming log-normally distributed data) and used to calculate II, RCV, and asymmetric RCV, where $RCV_{pos} = exp(1.96 \times \sqrt{2}\tau) - 1$, and $RCV_{neg} = exp(-1.96 \times \sqrt{2}\tau) - 1$, where $\tau = \sqrt{ln(CV_{A+I}^2 + 1)}$ (1.96 is selected from the normal distribution and CV_{A+I} is the coefficient of variation for the total imprecision, $CV_{A+I} = \sqrt{CV_A^2 + CV_I^2}$.[62,63] Asymmetric CVs are more appropriate estimates of RCV when using log-normal data.

6.3.4 Repeated simulations

Each data simulation, outlier detection method and corresponding analyses were performed 5,000 times. The number of simulations was chosen based on the estimate of coverage; with 5,000 randomly generated data sets an estimate of coverage of 95% (for each of the standard deviation estimates obtained) would have a confidence interval ranging from 94.359% to 95.588%.

For each outlier strategy the number of measurements identified as outliers and the number of individuals with outliers removed for each simulated data set was obtained.

Standard simulation performance measures were used to evaluate the ability of the methods to estimate the standard deviations using different outlier identification procedures and removing identified outliers from the data set. These measures were as suggested by Burton et al,[106] see Chapter 5 Table 5.2.

Additionally, for each of the calculated results the mean, median, 25th percentile (Q1), 75th percentile (Q3), minimum and maximum value were calculated to summarise results.

All analyses were performed using the statistical software R with the seed set at the start of each scenario (unique for each scenario). The same simulated data were used and analysed for each of the outlier strategies, meaning the analysis is paired for each of the 5,000 randomly generated data sets. All models were fitted using restricted estimation of maximum likelihood (REML).

6.4 Results

All displayed results give CVs and RCVs as percentages.

6.4.1 Simulation without outliers

6.4.1.1 Outliers detected and removed

Outlier detection strategies including the Cochran C test identified the most outliers; the median number of measurements removed in each data set was two when using just the Cochran C test and four when using the Cochran C test with Reed's criterion for means. The full use of the Cochran C test alone identified a maximum of 30 measurements and partial use identified a maximum of 14 measurements; when pairing with Reed's criterion for means (Fraser-Harris method) this increased to a maximum of 20. Reed's criterion on mean values for individuals resulted in a median of zero measurements removed but the maximum number of measurements removed was 16. See Table 6.2 and Figure 6.1 for further details.

Tukey's IQR rule identified a median of zero outliers for the 5,000 simulated data sets, but identified a maximum of 19 measurements in one of the simulated data sets. Grubbs's test and the ± 3SD rule identified a maximum of one and seven measurements, respectively; however, the median number identified and removed was zero. Use of Reed's criterion for all measurements and Dixon's Q test did not identify any outliers to be deleted.

Table 6.2: Outliers removed by each detection method for the 5,000 simulations. Maximum number of measurements removed is 160 and maximum number of individuals removed is 20.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	2 (0, 4)[0, 30]	1 (0, 3)[0, 19]	1 (0, 2)[0, 8]	5 (0, 10)[0, 40]
Cochran C test partial	2 (0, 4)[0, 14]	1 (0, 3)[0, 9]	1 (0, 2)[0, 7]	5 (0, 10)[0, 35]
Fraser-Harris method	4 (2, 8)[0, 20]	3 (1, 5)[0, 13]	1 (1, 2)[0, 7]	5 (5, 10)[0, 35]
Reed's criterion for means	0 (0, 8)[0, 16]	0 (0, 5)[0, 10]	0 (0, 1)[0, 2]	0 (0, 5)[0, 10]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	0 (0, 2)[0, 19]	0 (0, 1)[0, 12]	0 (0, 1)[0, 5]	0 (0, 5)[0, 25]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]
± 3SD	0 (0, 0)[0, 7]	0 (0, 0)[0, 4]	0 (0, 0)[0, 2]	0 (0, 0)[0, 10]

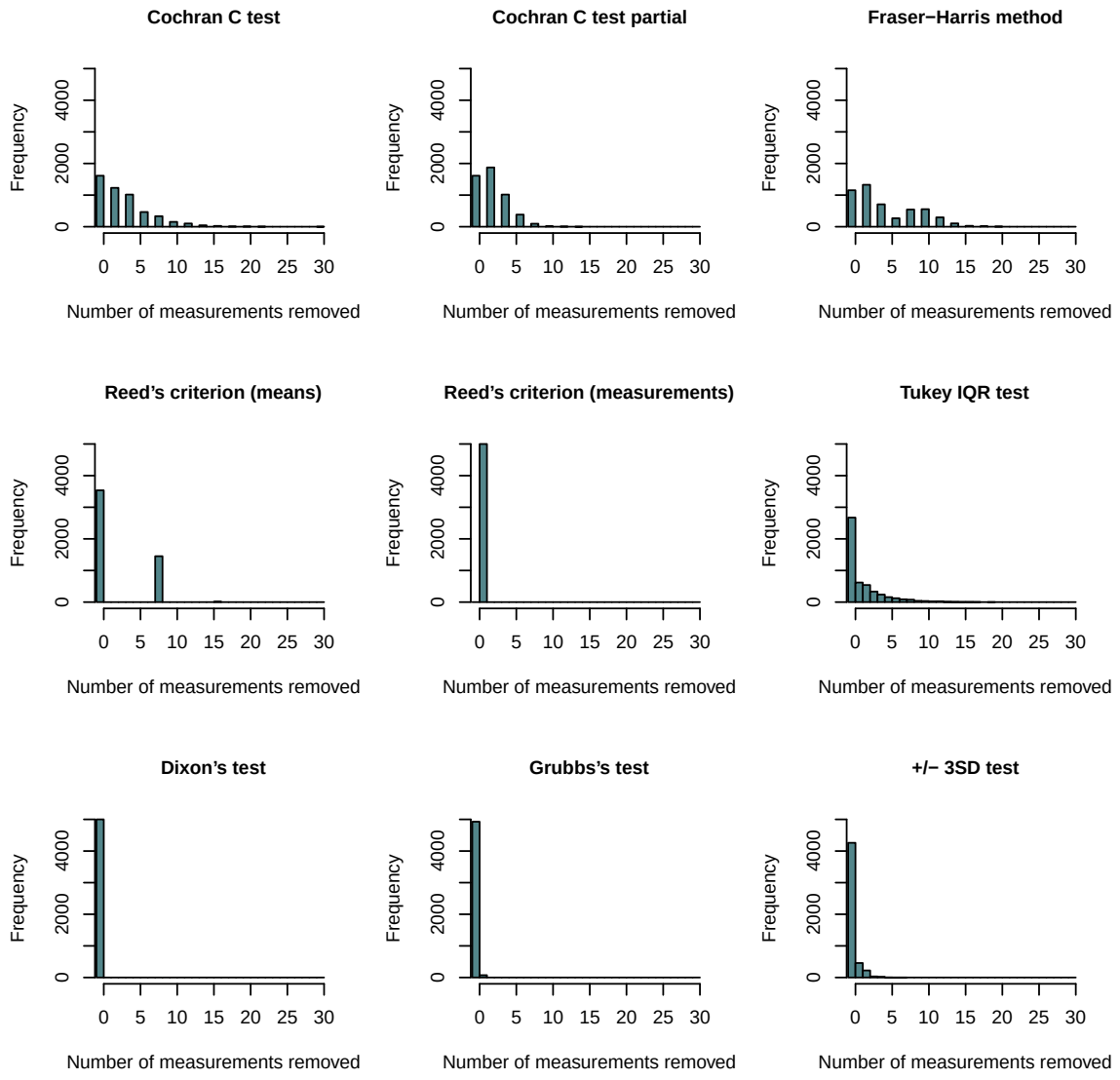


Figure 6.1: Histograms showing the number of measurements removed across the 5,000 simulations when using each outlier detection strategy.

6.4.1.2 Ability of methods to estimate standard deviations

The bias in estimating the analytical standard deviation was larger for the methods including the Cochran C test, with increased negative bias compared to the data with no outlier detection (percentage bias for analytical, within-individual and between-individual SD was -0.256, -0.365 and -1.608 respectively). The negative bias, even for the no outlier detection method was expected due to small sample bias (see Chapter 5).[50,107] Estimates of within-individual SD showed increased bias for the methods including the Cochran C test, especially the Fraser-Harris method (percentage bias for analytical, within-individual and between-individual SD was -2.126, -0.529 and -3.641 respectively). Use of Reed's criterion for means (percentage bias for analytical, within-individual and between-individual SD was -0.290, -0.492 and -3.679 respectively) increased bias, particularly at the between-individual level; and, use of the Tukey IQR rule (percentage bias for analytical, within-individual and between-individual SD was -0.496, -1.549 and -4.305 respectively) increased bias at all levels.

Coverage for estimates of analytical SD was lower when using an outlier detection strategy including the Cochran C test (decrease from 95.2% to 93.5%-94.0%) with the estimates of coverage for the Cochran C test analyses below the limit expected given 95% coverage based on the number of simulations used. Coverage estimates for within-individual SD were also lower with full use of the Cochran C test. When evaluating the estimate of between-individual SD the bias increased when using outlier strategies including Reed's criterion for means and for full use of the Cochran C tests, see Tables 6.3 and 6.4.

Results when using Tukey's IQR rule showed increased bias for all estimates of SD and decreased coverage for estimates of within-individual and between-individual variability. Increases in bias were seen for all estimates of SD when using Grubbs's test and $\pm$ 3SD also, but with the magnitude of the difference to no outlier detection compared with Tukey's IQR rule being less.

For the strategies using Reed's Criterion for measurements and Dixon's Q test no outliers were removed so performance was the same as when using no outlier detection.

Table 6.3: Outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-1.281**	**-3.641**	**-31.842**	**-0.256**	**-0.365**	**-1.608**	**-3.258**	**-3.529**	**-9.163**
Cochran C test	-9.244	-10.068	-31.387	-1.850	-1.009	-1.585	-22.842	-9.316	-9.006
Cochran C test partial	-10.426	-3.985	-31.348	-2.087	-0.399	-1.583	-25.742	-3.816	-9.008
Fraser-Harris method	-10.624	-5.275	-72.105	-2.126	-0.529	-3.641	-25.980	-5.010	-20.336
Reed's criterion for means	-1.447	-4.905	-72.854	-0.290	-0.492	-3.679	-3.651	-4.715	-20.561
Reed's criterion for measurements	-1.281	-3.641	-31.842	-0.256	-0.365	-1.608	-3.258	-3.529	-9.163
Tukey IQR rule	-2.477	-15.455	-85.258	-0.496	-1.549	-4.305	-6.289	-14.825	-23.870
Dixon's Q test	-1.281	-3.641	-31.842	-0.256	-0.365	-1.608	-3.258	-3.529	-9.163
Grubbs's test	-1.339	-3.927	-32.172	-0.268	-0.394	-1.624	-3.407	-3.807	-9.254
± 3SD	-1.622	-7.143	-39.430	-0.325	-0.716	-1.991	-4.119	-6.911	-11.307

Table 6.4: Outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.155**	**1.066**	**12.177**	**0.952**	**0.950**	**0.945**	**0.016**	**0.042**	**0.147**
Cochran C test	0.172	1.178	12.244	0.940	0.938	0.945	0.016	0.041	0.147
Cochran C test partial	0.175	1.092	12.210	0.937	0.947	0.944	0.016	0.041	0.147
Fraser-Harris method	0.178	1.112	13.092	0.935	0.948	0.944	0.016	0.042	0.145
Reed's criterion for means	0.157	1.085	13.085	0.952	0.950	0.946	0.016	0.042	0.145
Reed's criterion for measurements	0.155	1.066	12.177	0.952	0.950	0.945	0.016	0.042	0.147
Tukey IQR rule	0.156	1.111	13.484	0.953	0.944	0.934	0.016	0.041	0.143
Dixon's Q test	0.155	1.066	12.177	0.952	0.950	0.945	0.016	0.042	0.147
Grubbs's test	0.155	1.066	12.189	0.952	0.950	0.945	0.016	0.042	0.147
± 3SD	0.155	1.073	12.317	0.952	0.949	0.943	0.016	0.041	0.146

6.4.1.3 Ability of methods to estimate CVs

When using outlier detection strategies including the Cochran C test prior to analysis the median CV_A was 4.89%-4.91% compared with 4.99% for the no outlier detection method, with the true value of 5%. Reed's criterion for means, Tukey's IQR rule, Grubbs's test and the ± 3SD rule provided median estimates of CV_A of 4.98%, 4.97%, 4.98% and 4.98% respectively.

For the estimate of CV_I, the full Cochran C test strategy had a median value of 9.90% whereas the no outlier strategy had a median estimate of 9.94%; the true value was 10%. The Tukey IQR rule and the ± 3SD rule yielded median estimates of CV_I of 9.83% and 9.91% respectively.

When evaluating estimates of CV_G, the true value was 20%. The median estimate when using no outlier detection was 19.58%, this was similar when using strategies including the Cochran C test only (full and partial use) but for strategies including Reed's criterion for means, the median estimates produced were 19.15%. The Tukey IQR rule gave the worst estimate of 19.00%. When using the ± 3SD rule the median estimate was 19.48%. Results when using Grubbs's test and the Cochran C tests were similar to when no outlier removal was performed.

All estimates for strategies using only Reed's criterion for all measurements and Dixon's Q test remained the same as for using no outlier detection. See Tables 6.5 and 6.6, and Figure 6.2.

6.4.1.4 Ability of methods to estimate IIs and RCVs

The median result when estimating II using no outlier detection was 0.57; the true value was 0.56. Strategies including the Cochran C test provided a similar median estimate of II (0.56-0.58; slightly overestimated (II=0.58) when using the Fraser-Harris method and Reed's criterion for means). The Tukey IQR outlier method also slightly overestimated II (II=0.58). All other methods gave very similar results to the analysis when no outlier detection was

Table 6.5: Outlier detection methods with no outlier simulation–SDs, Median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.1 and σ_G is 0.2.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.05 (0.05, 0.05)[0.04, 0.06]**	**0.10 (0.09, 0.11)[0.06, 0.14]**	**0.19 (0.17, 0.22)[0.08, 0.33]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.07, 0.33]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.07, 0.33]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.07, 0.33]
Reed's criterion for means	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.07, 0.33]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.06, 0.14]	0.19 (0.16, 0.21)[0.07, 0.32]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]

Table 6.6: Outlier detection methods with no outlier simulation–CVs; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**4.99 (4.72, 5.25)[3.53, 6.40]**	**9.94 (9.26, 10.65)[6.29, 13.72]**	**19.58 (17.21, 22.01)[7.55, 34.11]**
Cochran C test	4.91 (4.63, 5.18)[3.52, 6.48]	9.90 (9.18, 10.62)[6.02, 13.98]	19.59 (17.21, 22.03)[7.45, 34.20]
Cochran C test partial	4.90 (4.62, 5.17)[3.50, 6.40]	9.94 (9.25, 10.65)[6.24, 13.83]	19.59 (17.24, 22.03)[7.45, 34.20]
Fraser-Harris method	4.89 (4.61, 5.17)[3.50, 6.40]	9.94 (9.23, 10.65)[6.24, 14.12]	19.15 (16.73, 21.67)[7.45, 34.20]
Reed's criterion for means	4.98 (4.72, 5.25)[3.53, 6.40]	9.94 (9.25, 10.65)[6.29, 14.00]	19.15 (16.72, 21.67)[7.31, 34.11]
Reed's criterion for measurements	4.99 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.58 (17.21, 22.01)[7.55, 34.11]
Tukey IQR rule	4.97 (4.70, 5.24)[3.53, 6.37]	9.83 (9.15, 10.52)[6.29, 13.67]	19.00 (16.57, 21.51)[6.64, 32.74]
Dixon's Q test	4.99 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.58 (17.21, 22.01)[7.55, 34.11]
Grubbs's test	4.98 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.57 (17.21, 22.01)[7.55, 34.11]
± 3SD	4.98 (4.72, 5.25)[3.53, 6.40]	9.91 (9.23, 10.61)[6.29, 13.69]	19.48 (17.15, 21.93)[7.55, 34.11]

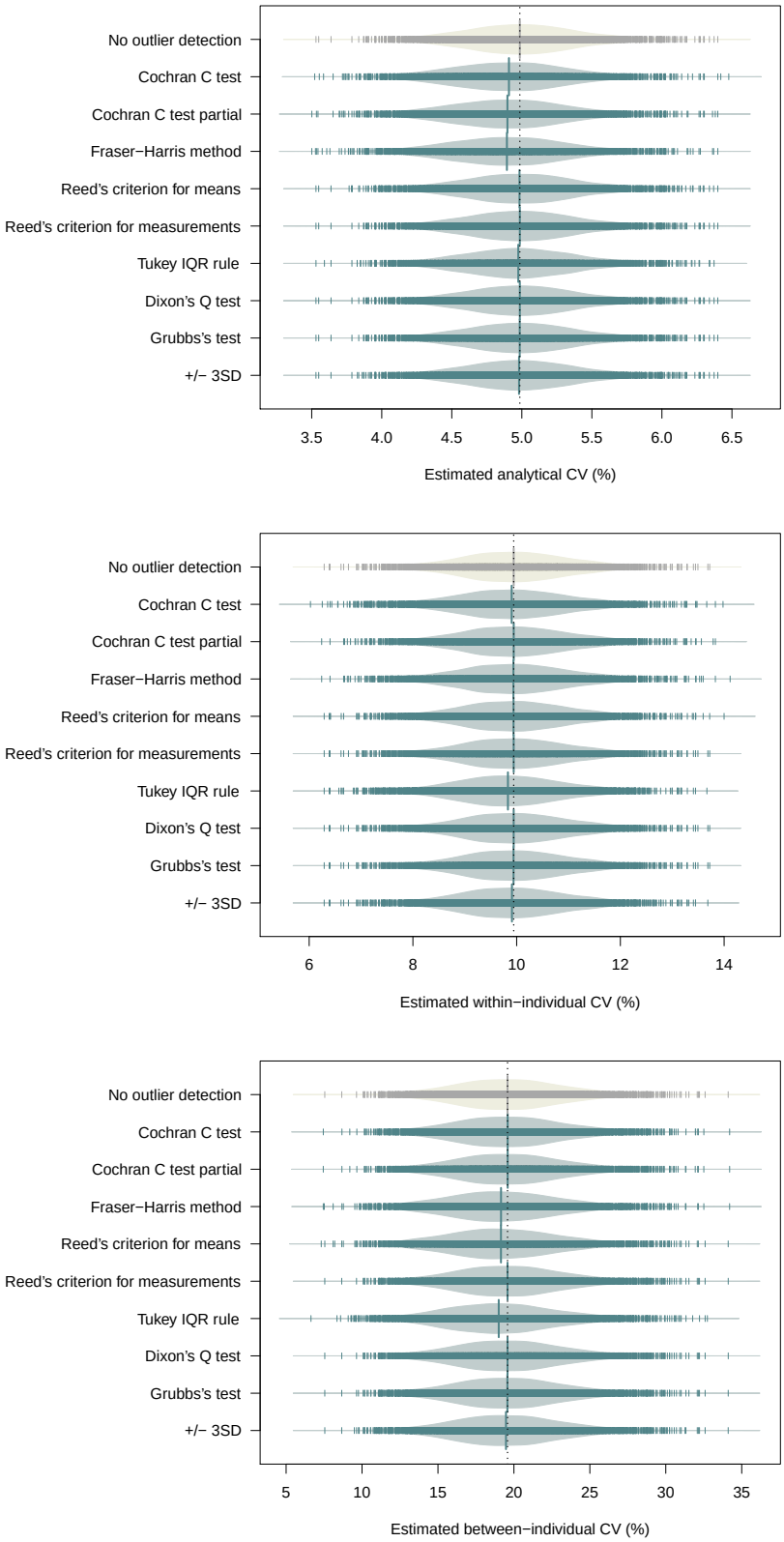


Figure 6.2: Estimates of CVs when using different outlier detection strategies. Beanplots to show distribution; median is shown by bold line and the dashed line reflects the value estimated when no outliers are removed.

used (II=0.57).

The RCV value for this simulation was 30.99%. The median RCV was estimated at 30.86% when no outlier detection was used and this reduced to 30.65%-30.74% when using strategies including Cochran C test. The lowest median estimate was 30.53% when using the Tukey IQR rule. Use of the ± 3SD rule yielded a slight difference in median RCV estimate of 30.76%, see Table 6.7 and Figure 6.3.

Similar results were seen when evaluating the asymmetric RCVs with the lower and upper bound slightly further underestimated, compared to using no outlier detection method, when using the strategies including the Cochran C test and ± 3SD rule; and the underestimation exaggerated further when using the Tukey IQR rule, see Table 6.8 and Figure 6.4.

6.4.1.5 Sensitivity analysis–simulation of test with 'poor performance'

The simulation of data for a test with increased variability showed the same pattern of results as seen in the original simulation. The outlier strategies detecting the most measurements to be removed remained consistent; with slightly fewer measurements removed compared to the analysis of the original data simulation. The absolute bias estimates were generally larger with bias increased (underestimation) for analytical and within-individual estimates of variability when using methods including the Cochran C test, and for estimates of variability at all levels when using methods including the Tukey IQR rule and the ± 3SD rule. Coverage again decreased for estimates of analytical and within-individual SD when using strategies including the Cochran C test. Median estimates of CV_A and CV_I were underestimated when using strategies including the Cochran C test and estimates of CV_I and CV_G were underestimated when using Tukey IQR rule and ± 3SD rule. See Appendix D Tables D.1-D.28.

Table 6.7: Outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10. RCVs are displayed as percentages.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.57 (0.50, 0.65)[0.31, 1.48]**	**30.86 (29.16, 32.63)[22.33, 40.37]**	**9.97 (9.94, 10.01)[9.82, 10.14]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.51]	30.65 (28.88, 32.47)[21.93, 40.59]	9.97 (9.94, 10.01)[9.82, 10.17]
Cochran C test partial	0.57 (0.50, 0.65)[0.30, 1.51]	30.74 (29.02, 32.56)[22.04, 40.59]	9.97 (9.94, 10.01)[9.82, 10.14]
Fraser-Harris method	0.58 (0.50, 0.67)[0.30, 1.65]	30.73 (28.97, 32.52)[22.04, 40.87]	9.98 (9.94, 10.01)[9.80, 10.16]
Reed's criterion for means	0.58 (0.51, 0.67)[0.31, 1.69]	30.83 (29.10, 32.61)[22.33, 40.87]	9.98 (9.94, 10.01)[9.81, 10.15]
Reed's criterion for measurements	0.57 (0.50, 0.65)[0.31, 1.48]	30.86 (29.16, 32.63)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
Tukey IQR rule	0.58 (0.51, 0.67)[0.31, 1.65]	30.53 (28.87, 32.33)[22.33, 39.66]	9.97 (9.94, 10.01)[9.82, 10.15]
Dixon's Q test	0.57 (0.50, 0.65)[0.31, 1.48]	30.86 (29.16, 32.63)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
Grubbs's test	0.57 (0.50, 0.65)[0.31, 1.48]	30.85 (29.15, 32.62)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
± 3SD	0.57 (0.50, 0.66)[0.31, 1.48]	30.76 (29.05, 32.56)[22.33, 39.96]	9.97 (9.94, 10.01)[9.82, 10.14]

Table 6.8: Outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-26.48 (-27.76, -25.23)[-33.07, -19.98]**	**36.02 (33.75, 38.43)[24.97, 49.42]**
Cochran C test	-26.33 (-27.65, -25.02)[-33.22, -19.66]	35.74 (33.37, 38.21)[24.48, 49.74]
Cochran C test partial	-26.39 (-27.71, -25.13)[-33.22, -19.76]	35.86 (33.56, 38.33)[24.62, 49.74]
Fraser-Harris method	-26.38 (-27.68, -25.09)[-33.41, -19.76]	35.84 (33.50, 38.28)[24.62, 50.16]
Reed's criterion for means	-26.46 (-27.74, -25.19)[-33.41, -19.98]	35.98 (33.66, 38.39)[24.97, 50.16]
Reed's criterion for measurements	-26.48 (-27.76, -25.23)[-33.07, -19.98]	36.02 (33.75, 38.43)[24.97, 49.42]
Tukey IQR rule	-26.24 (-27.54, -25.02)[-32.60, -19.98]	35.57 (33.36, 38.02)[24.97, 48.37]
Dixon's Q test	-26.48 (-27.76, -25.23)[-33.07, -19.98]	36.02 (33.75, 38.43)[24.97, 49.42]
Grubbs's test	-26.48 (-27.75, -25.23)[-33.07, -19.98]	36.02 (33.74, 38.41)[24.97, 49.42]
± 3SD	-26.41 (-27.71, -25.15)[-32.80, -19.98]	35.89 (33.61, 38.33)[24.97, 48.82]

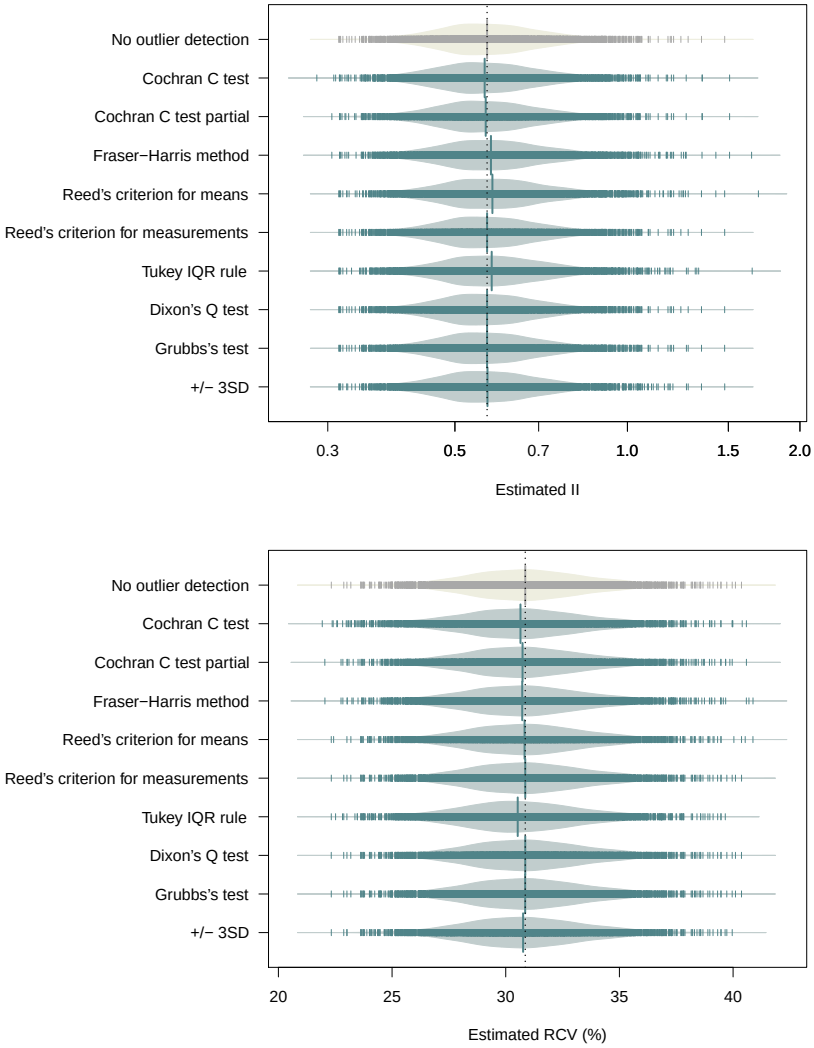


Figure 6.3: Estimates of II and RCV when using different outlier detection strategies. Beanplots to show distribution; median is shown by bold line and the dashed line reflects the value estimated when no outliers are removed.

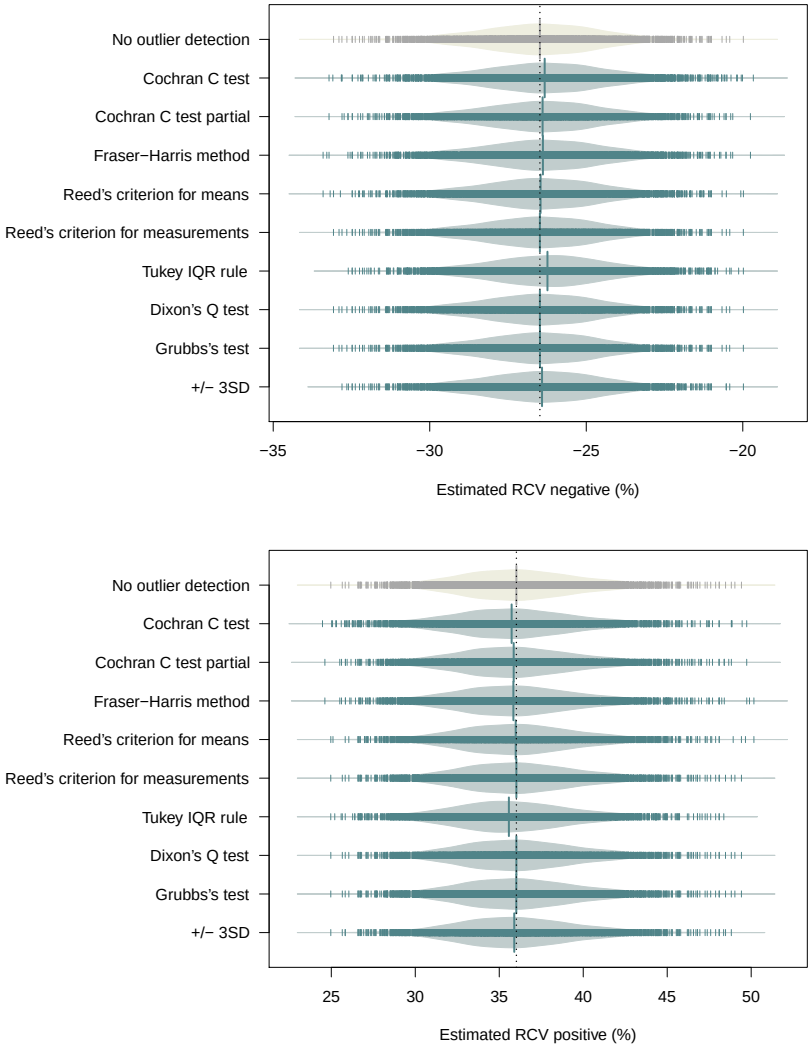


Figure 6.4: Estimates of asymmetric RCV when using different outlier detection strategies. Beanplots to show distribution; median is shown by bold line and the dashed line reflects the value estimated when no outliers are removed.

6.4.1.6 Sensitivity analysis–simulation of data with increased sample size

When increasing the number of participants, the results when using the different outlier detection strategies followed the same pattern. The number of measurements removed was increased, as would be assumed given an increased sample size, and there were more measurements identified by the same strategies (strategies including the Cochran C test, Tukey IQR rule and ± 3SD). The bias in results was consistent, with methods including the Cochran C test having increased negative bias for estimates of analytical and within-individual variability and bias for all estimates was poorer for the strategies including Tukey IQR rule and ± 3SD. Estimates of bias generally were increased for analytical variability and decreased for within-individual and between-individual variability. Coverage for the estimate of analytical SD when using methods including the Cochran C test remained low; with the coverage when estimating within-individual variability still lower than when no outliers were removed but closer to 95% than in the base case simulations. Coverage for the estimate of between-individual SD when using the Tukey IQR rule appeared notably lower also. The effect on median CV estimates was similar to other data simulations. See Appendix D Tables D.29-D.35.

6.4.2 Simulation of outlying data

For full results of simulations including outlying data see Appendix D Tables D.36-D.71.

6.4.2.1 Outliers detected and removed

The Cochran C test (full and partial), the Fraser-Harris method and the Tukey IQR rule identified all outliers in the simulation for all scenarios but with the exclusion of many more non-outliers. Reed's criterion for means performed poorly, not identifying all outliers and removing many valid measurements. Dixon's Q test, Grubbs's test and Reed's criterion for measurements worked appropriately with only a single outlier but could not identify outliers when more than one was present. The ± 3SD rule performed well, especially when the

magnitude of the difference for outliers was smaller and there was only one outlier, see Table 6.9 for full results.

6.4.2.2 Ability of methods to estimate standard deviations

Estimates of bias for the standard deviations showed when no outliers detection was used, for all outlier simulation scenarios, the analytical SD was overestimated and the within-individual and between-individual SDs were underestimated.

When the data included only one outlier (magnitude 2 and 10), Reed's criterion for means performed poorly (overestimating σ_A and underestimating σ_I and σ_G). Results for outlier detection strategies including the Cochran C test and the Tukey IQR rule underestimated the standard deviations at all levels, compared with results using Reed's criterion for measurements, Dixon's test, Grubbs's test and ± 3SD (methods that correctly identified only the outlier).

When more than one outlier was simulated and the magnitude was two, the methods of Reed's criterion for means and measurements, Dixon's test and Grubbs's test performed poorly (overestimating σ_A and underestimating σ_I and σ_G); strategies including the Cochran C test and the Tukey IQR rule did not perform as well as ± 3SD. When simulating data with multiple outliers and with the increase magnitude (10) the ± 3SD rule also performed poorly.

Table 6.9: Outlier detection methods with outlier simulation–outliers removed by each detection method. Maximum number of measurements is 160, maximum number of individuals is 20 and maximum number of outliers removed is 0.5%=1, 1%=2 and 2%=4.

Outlier strategy	**Observations removed**		**Individuals with measurements removed**		**Outliers removed**	
			median (Q1, Q3) [minimum, maximum]			
	n	%	n	%	n	%
Scenario: 0.5% and magnitude 2						
Cochran C test	4 (2, 6)[2, 30]	3 (1, 4)[1, 19]	2 (1, 3)[1, 7]	10 (5, 15)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Cochran C test partial	4 (2, 6)[2, 12]	3 (1, 4)[1, 8]	2 (1, 3)[1, 6]	10 (5, 15)[5, 30]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Fraser-Harris method	4 (4, 10)[2, 24]	3 (3, 6)[1, 15]	2 (2, 3)[1, 7]	10 (10, 15)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 16]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 1]	100 (0, 100)[0, 100]
Reed's criterion for measurements	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Tukey IQR rule	1 (1, 3)[1, 20]	1 (1, 2)[1, 13]	1 (1, 2)[1, 7]	5 (5, 10)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Dixon's Q test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
± 3SD	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Scenario: 1% and magnitude 2						
Cochran C test	6 (4, 8)[4, 32]	4 (3, 5)[3, 20]	3 (2, 4)[2, 8]	15 (10, 20)[10, 40]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Cochran C test partial	6 (4, 8)[4, 16]	4 (3, 5)[3, 10]	3 (2, 4)[2, 8]	15 (10, 20)[10, 40]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Fraser-Harris method	6 (4, 12)[4, 24]	4 (3, 8)[3, 15]	3 (2, 4)[2, 8]	15 (10, 20)[10, 40]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 16]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	50 (0, 50)[0, 100]
Reed's criterion for measurements	1 (0, 1)[0, 2]	1 (0, 1)[0, 1]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	50 (0, 50)[0, 100]
Tukey IQR rule	2 (2, 4)[2, 21]	1 (1, 3)[1, 13]	2 (2, 3)[2, 8]	10 (10, 15)[10, 40]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Dixon's Q test	1 (0, 1)[0, 1]	1 (0, 1)[0, 1]	1 (0, 1)[0, 1]	5 (0, 5)[0, 5]	1 (0, 1)[0, 1]	50 (0, 50)[0, 50]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	50 (50, 50)[50, 50]
± 3SD	2 (2, 2)[2, 2]	1 (1, 1)[1, 1]	2 (2, 2)[2, 2]	10 (10, 10)[10, 10]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]

Outlier strategy	Observations removed		Individuals with measurements removed		Outliers removed	
			median (Q1, Q3) [minimum, maximum]			
	n	%	n	%	n	%
Scenario: 2% and magnitude 2						
Cochran C test	10 (8, 14)[8, 38]	6 (5, 9)[5, 24]	5 (4, 5)[4, 9]	25 (20, 25)[20, 45]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Cochran C test partial	10 (8, 10)[8, 18]	6 (5, 6)[5, 11]	5 (4, 5)[4, 9]	25 (20, 25)[20, 45]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Fraser-Harris method	10 (8, 16)[8, 26]	6 (5, 10)[5, 16]	5 (4, 6)[4, 10]	25 (20, 30)[20, 50]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 16]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	25 (0, 25)[0, 50]
Reed's criterion for measurements	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 5)[0, 5]	0 (0, 1)[0, 1]	0 (0, 25)[0, 25]
Tukey IQR rule	4 (4, 6)[4, 23]	3 (3, 4)[3, 14]	4 (4, 5)[4, 9]	20 (20, 25)[20, 45]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Dixon's Q test	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 5)[0, 5]	0 (0, 1)[0, 1]	0 (0, 25)[0, 25]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	25 (25, 25)[25, 25]
± 3SD	4 (4, 4)[4, 4]	3 (3, 3)[3, 3]	4 (4, 4)[4, 4]	20 (20, 20)[20, 20]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Scenario: 0.5% and magnitude 10						
Cochran C test	4 (2, 6)[2, 28]	3 (1, 4)[1, 18]	2 (1, 3)[1, 7]	10 (5, 15)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Cochran C test partial	4 (2, 6)[2, 14]	3 (1, 4)[1, 9]	2 (1, 3)[1, 7]	10 (5, 15)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Fraser-Harris method	4 (4, 10)[2, 22]	3 (3, 6)[1, 14]	2 (1, 3)[1, 8]	10 (5, 15)[5, 40]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 8]	5 (0, 5)[0, 5]	1 (0, 1)[0, 1]	5 (0, 5)[0, 5]	1 (0, 1)[0, 1]	100 (0, 100)[0, 100]
Reed's criterion for measurements	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Tukey IQR rule	1 (1, 3)[1, 22]	1 (1, 2)[1, 14]	1 (1, 2)[1, 7]	5 (5, 10)[5, 35]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Dixon's Q test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]
± 3SD	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	100 (100, 100)[100, 100]

Outlier strategy	Observations removed		Individuals with measurements removed		Outliers removed	
			median (Q1, Q3) [minimum, maximum]			
	n	%	n	%	n	%
Scenario: 1% and magnitude 10						
Cochran C test	6 (4, 8)[4, 28]	4 (3, 5)[3, 18]	3 (2, 4)[2, 9]	15 (10, 20)[10, 45]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Cochran C test partial	6 (4, 8)[4, 18]	4 (3, 5)[3, 11]	3 (2, 4)[2, 9]	15 (10, 20)[10, 45]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Fraser-Harris method	6 (4, 12)[4, 26]	4 (3, 8)[3, 16]	3 (2, 4)[2, 9]	15 (10, 20)[10, 45]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 16]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	50 (0, 50)[0, 100]
Reed's criterion for measurements	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 5)[0, 5]	0 (0, 1)[0, 1]	0 (0, 50)[0, 50]
Tukey IQR rule	2 (2, 4)[2, 21]	1 (1, 3)[1, 13]	2 (2, 3)[2, 7]	10 (10, 15)[10, 35]	2 (2, 2)[2, 2]	100 (100, 100)[100, 100]
Dixon's Q test	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 1)[0, 1]	0 (0, 5)[0, 5]	0 (0, 1)[0, 1]	0 (0, 50)[0, 50]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	50 (50, 50)[50, 50]
± 3SD	2 (1, 2)[1, 2]	1 (1, 1)[1, 1]	2 (1, 2)[1, 2]	10 (5, 10)[5, 10]	2 (1, 2)[1, 2]	100 (50, 100)[50, 100]
Scenario: 2% and magnitude 10						
Cochran C test	10 (8, 14)[8, 44]	6 (5, 9)[5, 28]	5 (4, 5)[4, 10]	25 (20, 25)[20, 50]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Cochran C test partial	10 (8, 10)[8, 20]	6 (5, 6)[5, 13]	5 (4, 5)[4, 10]	25 (20, 25)[20, 50]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Fraser-Harris method	10 (8, 16)[8, 28]	6 (5, 10)[5, 18]	5 (4, 6)[4, 10]	25 (20, 30)[20, 50]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Reed's criterion for means	8 (0, 8)[0, 16]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	5 (0, 5)[0, 10]	1 (0, 1)[0, 2]	25 (0, 25)[0, 50]
Reed's criterion for measurements	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]	0 (0, 0)[0, 1]	0 (0, 0)[0, 25]
Tukey IQR rule	4 (4, 6)[4, 21]	3 (3, 4)[3, 13]	4 (4, 5)[4, 9]	20 (20, 25)[20, 45]	4 (4, 4)[4, 4]	100 (100, 100)[100, 100]
Dixon's Q test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]	0 (0, 0)[0, 1]	0 (0, 0)[0, 25]
Grubbs's test	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	1 (1, 1)[1, 1]	5 (5, 5)[5, 5]	1 (1, 1)[1, 1]	25 (25, 25)[25, 25]
± 3SD	2 (2, 3)[1, 4]	1 (1, 2)[1, 3]	2 (2, 3)[1, 4]	10 (10, 15)[5, 20]	2 (2, 3)[1, 4]	50 (50, 75)[25, 100]

6.4.2.3 Ability of methods to estimate CVs

With no outlier detection methods used the analysis of the data from the outlier simulations overestimated CV_A, and underestimated CV_I and CV_G. When simulating a single outlier, estimates of CV_A were reasonable for all outlier exclusion methods, except for Reed's criterion for means. When using Reed's criterion for means the median CV_A was still close to the true value but for some simulations CV_A was overestimated.

For the simulations of more than one outlier with magnitude two, the estimates of CV_A were overestimated when using Reed's criterion for means and measurements, Dixon's test and Grubbs's test. Estimates of CV_A were appropriate for strategies including the Cochran C test, the Tukey IQR rule and ± 3SD.

When simulating data with outlier(s) of magnitude 10, there was a general increase in the variability of estimates of CVs. For the 1% outlier simulation, the methods of Reed's criterion for means and measurements, Dixon's test and Grubbs's test were poor. The median estimate of CV_A when using ± 3SD was slightly increased also. When the outliers were increased to 2% the same strategies performed poorly, and the estimates were further overestimated.

When analysing estimates of CV_I and CV_G there was a similar pattern for the outlier detection strategies performing poorly, but with underestimation, and this was not as extreme as for CV_A. See Figures 6.5-6.10.

6.4.2.4 Ability of methods to estimate IIs and RCVs

The II and RCV estimates were extreme, compared to the true values, when analysing the simulated data with outliers using no outlier detection (II: 3.41 to $> 10^4$; lower bound RCV (%): -87.13 to -100.00; and, upper bound RCV (%): 677.30 to $> 10^4$). Estimates when simulating data with only one outlier (regardless of magnitude) were inflated using Reed's criterion for means (II=0.68/0.64; lower bound RCV (%):-28.01/-27.59; and, upper bound RCV (%): 38.91/38.10). When estimating II and RCV with simulations with more outliers, the methods of Reed's criterion for means and measurements, Dixon's and Grubbs's test

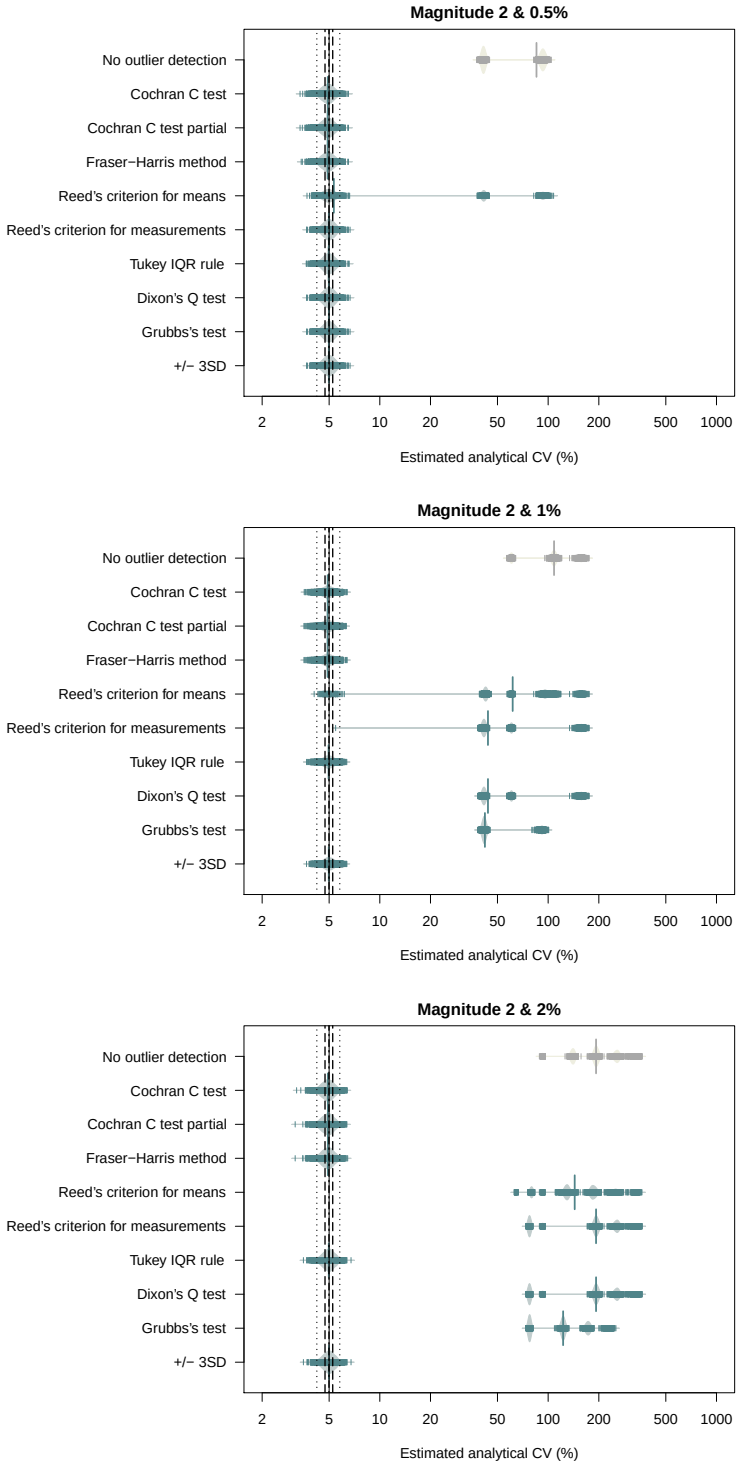


Figure 6.5: Outlier detection methods with outlier simulation (magnitude 2)–analytical CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 5%. * shows estimates continue beyond the range displayed.

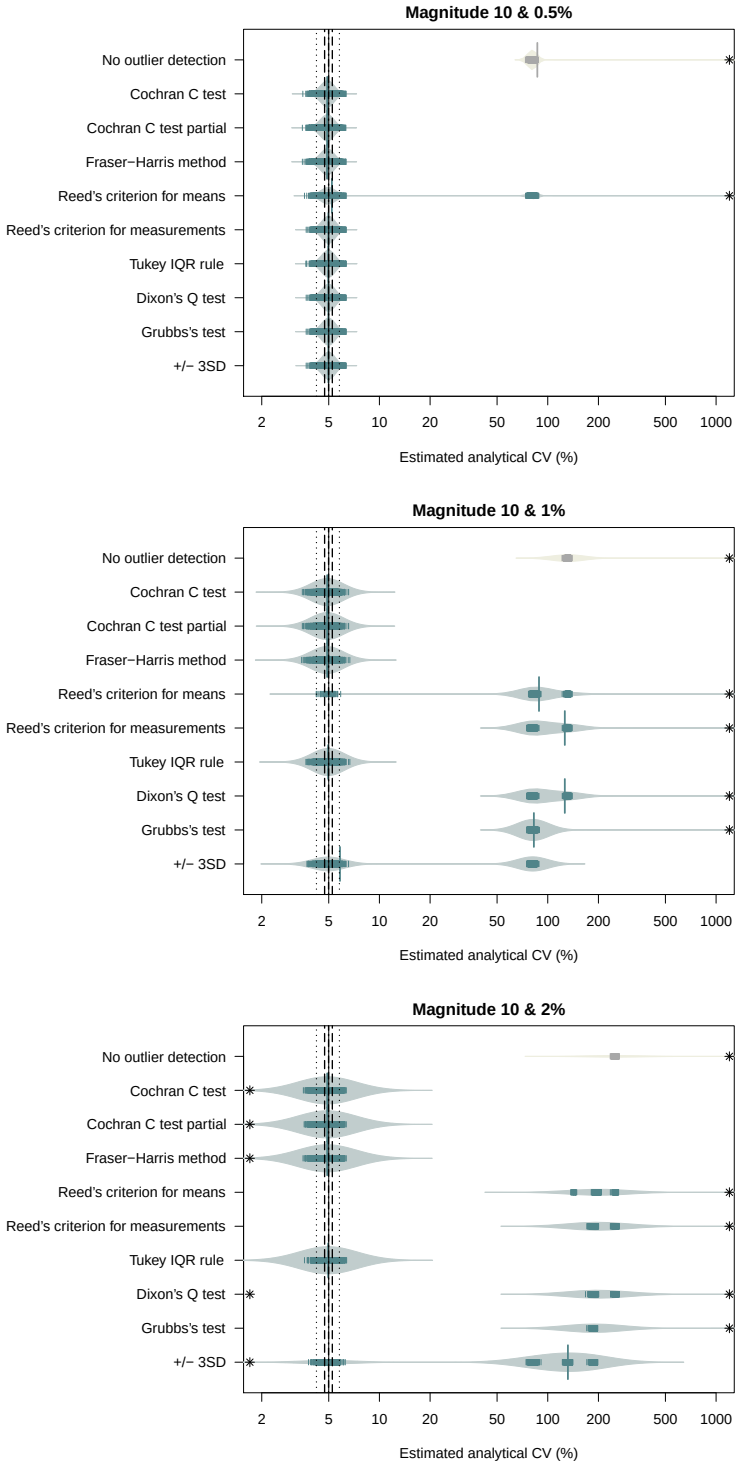


Figure 6.6: Outlier detection methods with outlier simulation (magnitude 10)–analytical CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 5%. * shows estimates continue beyond the range displayed.

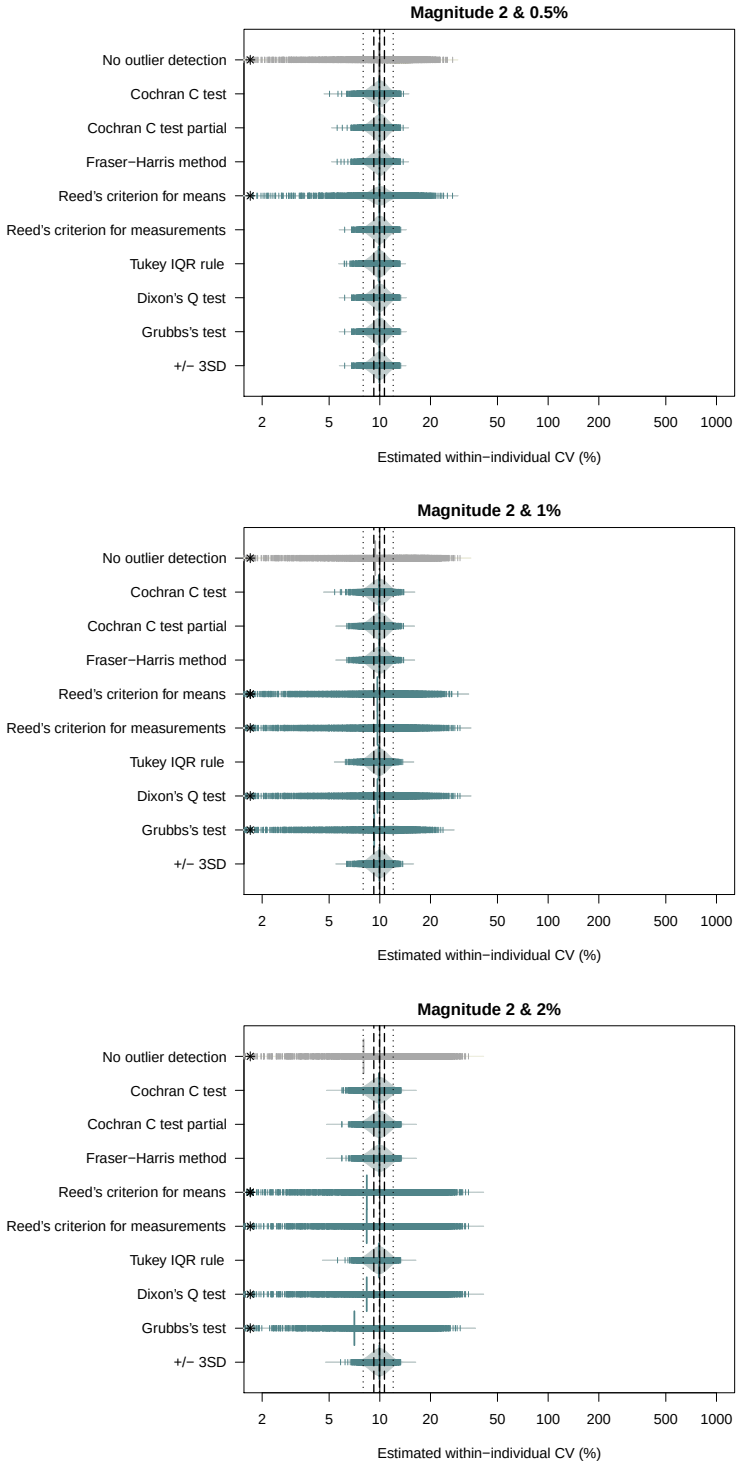


Figure 6.7: Outlier detection methods with outlier simulation (magnitude 2)–within-individual CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 10%. * shows estimates continue beyond the range displayed.

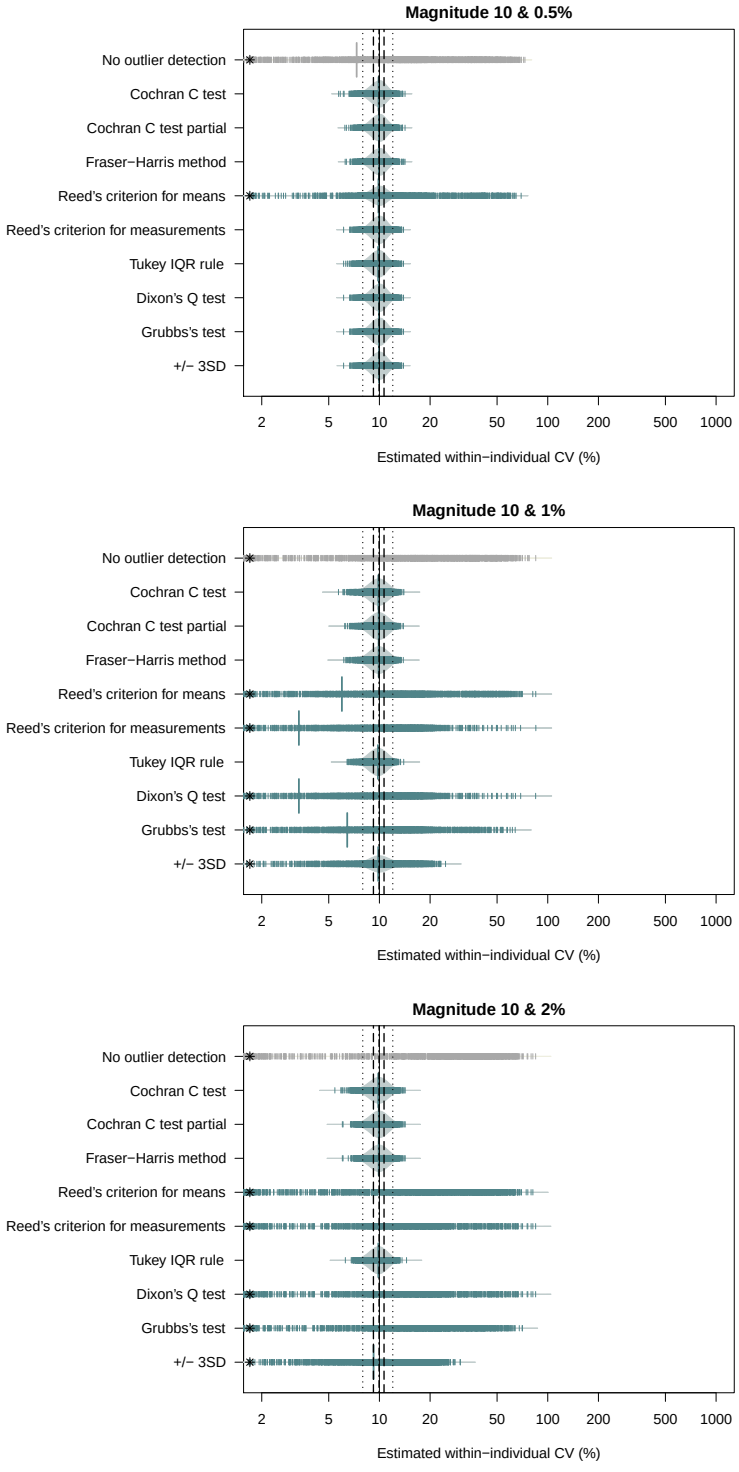


Figure 6.8: Outlier detection methods with outlier simulation (magnitude 10)–within-individual CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 10%. * shows estimates continue beyond the range displayed.

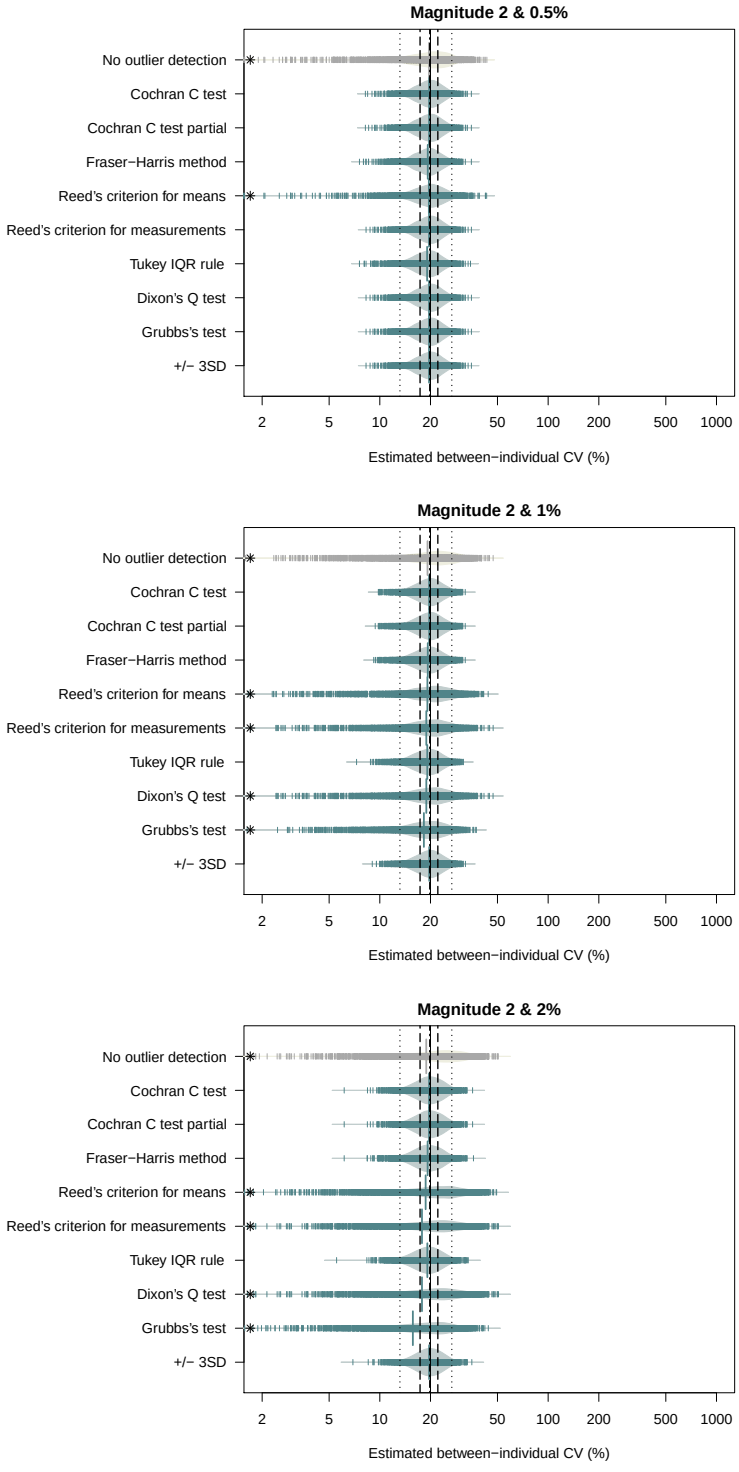


Figure 6.9: Outlier detection methods with outlier simulation (magnitude 2)–between-individual CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 20%. * shows estimates continue beyond the range displayed.

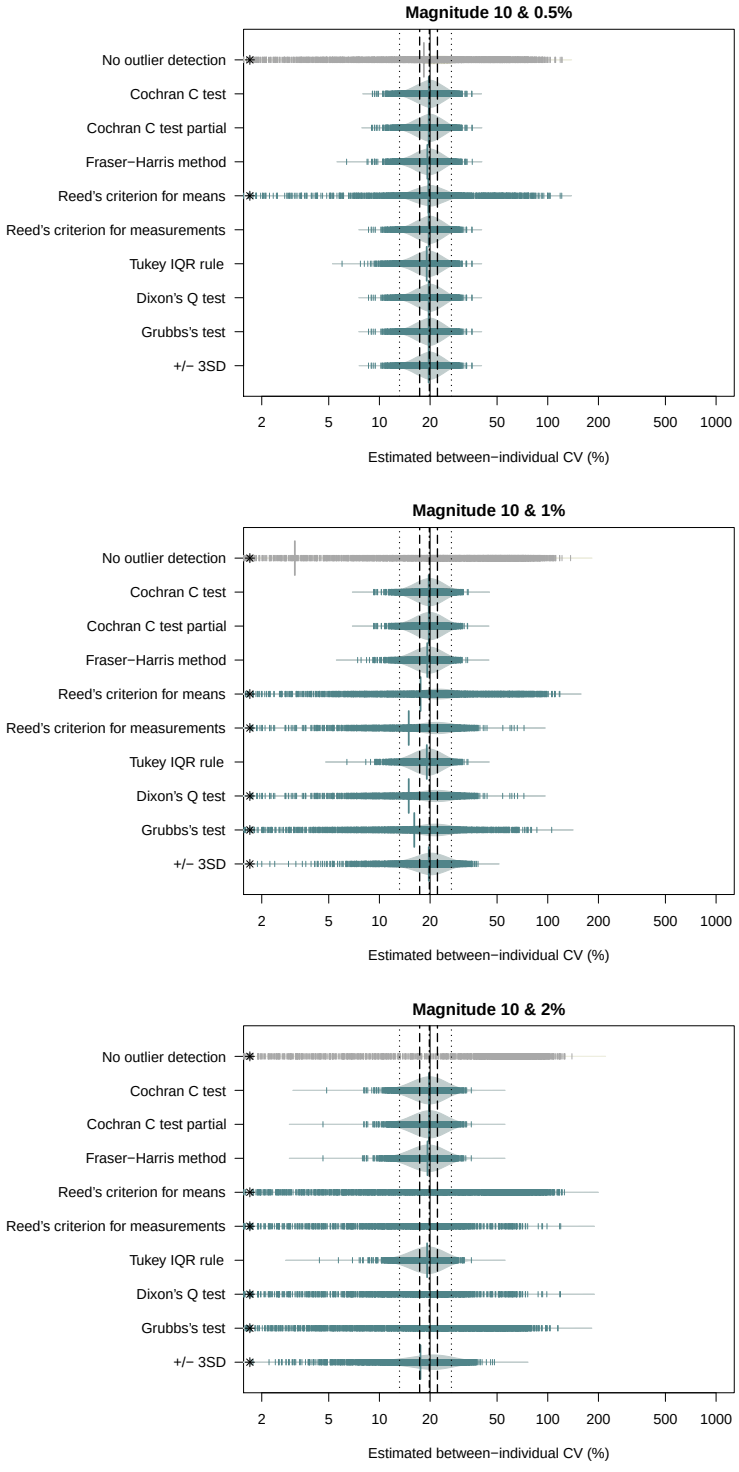


Figure 6.10: Outlier detection methods with outlier simulation (magnitude 10)–between-individual CV. Beanplots to show distribution. Vertical lines indicate values at the 2.5th percentile, 25th percentile, median, 75th percentile and 97.5th percentile of estimates when simulating data without outliers. The true value of CV is 20%. * shows estimates continue beyond the range displayed.

produced extreme results (II and RCV upper bound (%) estimates $> 10^4$ and RCV lower bound (%) estimates of -100); for the high outlier magnitude simulation ± 3SD also produced extreme results (II: 6.94; lower bound RCV (%): -93.85; and, upper bound RCV (%): 1526.72).

6.4.3 Comparison of outlier detection methods

The Cochran C test, Cochran C test partial, Fraser-Harris method and Tukey IQR rule introduced negative bias to estimates of variability when outliers were simulated and not simulated; when outliers were simulated estimated standard deviations were similar to analyses of data when outliers were not simulated. When using Reed's criterion for means, Reed's criterion for measurements, Dixon's test and Grubbs's test the bias was small when analysing data with no outliers but with the outlier simulation (or more than one outlier simulated) the methods failed to identify outliers and the bias was large. The ± 3SD method gave reasonable estimates with small bias for the no outlier simulation and most of the outlier scenarios, the bias only increased for the most extreme outlier scenario, see Figure 6.11.

When estimating the analytical standard deviation using the Cochran C test, Cochran C test partial and the Fraser-Harris method the bias was negative (approximately -2%) when no outliers were simulated and for all outlier simulation scenarios; there was less bias (approximately -1%) when using the Tukey IQR rule. When estimating the within-individual standard deviation the Tukey IQR rule gave the most negatively bias results for all scenarios (no outliers and all outlier simulations); results when using the Cochran C test, Cochran C test partial and the Fraser-Harris method were less biased. When estimating the between-individual standard deviation, bias was larger when using the Fraser-Harris method and Tukey IQR rule for the no outlier and all outlier scenarios; estimates when using Cochran C test and Cochran C test partial were close to results using no outlier detection for the data with no outliers simulated.

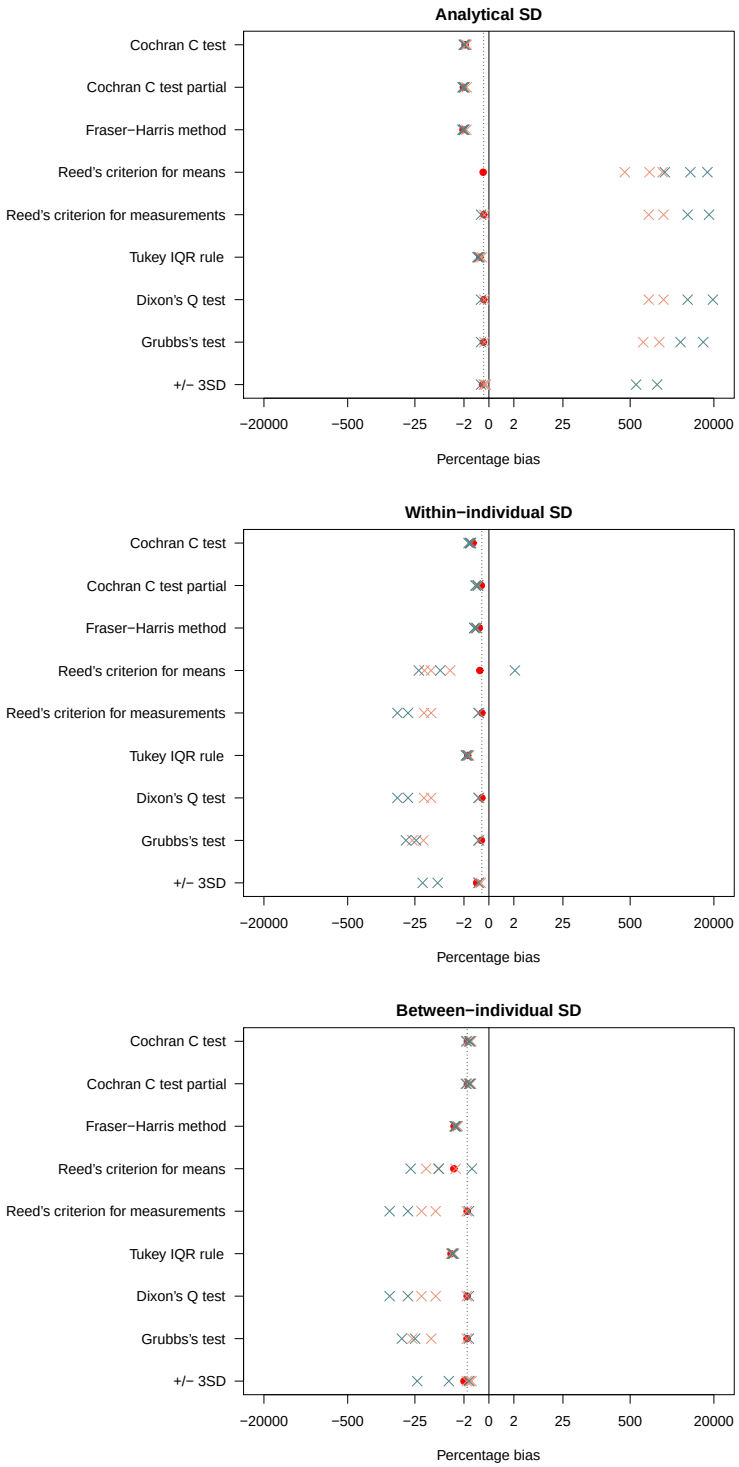


Figure 6.11: Comparison of average percentage bias for analytical, within-individual and between-individual standard deviations between outlier detection methods. Dotted line indicates the average percentage bias when using no outlier detection methods for the data with no outliers. Crosses indicate the average percentage bias for the scenarios with outliers of magnitude 2 (salmon) and magnitude 10 (teal). Red points indicate the average percentage bias for each outlier detection method when analysing data with no outliers.

6.5 Discussion

The analyses using the different outlier detection methods showed, in many cases, none of the log-normally distributed data were considered outlying, leading to no changes to the data and the same results obtained from analysis. The methods of outlier detection using the Cochran C test and Reed's criterion for means caused data to be detected as outlying more often and subsequently the results for this analysis reflected an increase in (negative) bias when estimating analytical and within-individual variation. The Tukey IQR rule, Grubbs's test and ± 3SD method led to outlier detection also, with increases in the bias of variability results (particularly when using the Tukey IQR method). Strategies including the Cochran C test led to reduced estimates of CV_A and CV_I; and strategies using the Tukey IQR rule led to decreased CV_I and CV_G estimates. For all strategies estimates of II were similar with slightly increased results when using the Tukey IQR rule and the RCV bound estimates were closer to zero for methods including the Cochran C test, Tukey IQR rule and ± 3SD rule. The negative bias (underestimation) of the variance parameters was expected with the small sample size.[50,107]

This simulation showed some of the outlier detection methods used may incorrectly identify data as outlying and these data were inappropriately removed prior to analysis. These outlier detection methods are based on reducing the variability within and between groups and the range of data, with the more extreme data points removed. After detecting data as outlying these data were then removed resulting in reduced variability shown by the increased negative bias in the variability results after using outlier detection techniques; some estimated standard deviations at the analytical level had -2% bias; at the within-individual level had -1% bias; and, at the between-individual level -4% bias, which has the potential to impact results and how tests are used. Different techniques affected different levels of bias with the Cochran C test changing the results for analytical and within-individual variability, as the test focussed on variability within the duplicate assessments and the groups; and, Reed's criterion for means and the Tukey IQR test had biggest impact on the estimate of variability at the between-individual level.

The outlier simulation showed methods that identify a single outlier or a minimum and/or maximum outlier worked well when only one outlier was present. Other methods detected more data to be excluded than the just the outlying measurements. Given the magnitude of the difference in the outlying data compared with the unaltered data, the outlying measurements could easily have been detected using descriptive data summaries or plots of the data. It is important for researchers in this area to understand how these outlier detection methods work and for these methods to be used in an appropriate way, tailored to the data set. Outlier detection methods should not be used as standard but only if there is reason to believe extreme data are present with the nature of these outliers understood and the appropriate action identified. The ± 3SD rule worked well, only providing invalid results for the most extreme outlier simulation.

Separate estimates of CV_A, CV_I and CV_G were more robust to incorrect outlier removal than estimates of II and RCV which use combinations of CV_A, CV_I and CV_G. For the analyses where outliers were simulated, the results showed increased CV_A as the outliers introduced were 'error outliers which would increase variability at the analytical level.

Outlier detection methods were used following transformation of data. In this artificial simulation of log-normal data, this approach is known to be required. As many of the outlier methods rely on normality of data, this approach should allow optimal use of these methods. There is guidance for the use of outlier detection and data transformation that may lead to outlier detection and removal prior to transformation.[46] If this strategy is used, outlier detection methods may perform poorly in comparison to the results shown.

The Birmingham Quality group, part of the National External Quality Assessment Scheme (NEQAS), monitor the results across UK labs when assessing control material. When reporting the results of monitoring assessments, NEQAS rank the data and the 2.5th to 97.5th percentile range is reported,[114] similar to the methods relying on data ranges evaluated.

6.5.1 Limitations

Data were simulated from a log-normal distribution which may not be truly reflective of biological variability data seen in practice. The first stage of simulations used data that were normally distributed after transformation and because of this outlier detection was not necessary as the data were simulated without outlying data. This element of the study only allows understanding of the consequences of using different outlier detection methods on data that are normally distributed. To understand the potential performance of outlier detection strategies, simulations were performed with outliers introduced.

This artificial generation of outliers may not be reflective of the type of outlying data seen in biological variability studies. The simulation used added 'error' outliers. These values were random with no link to observation points or individuals. Further work could investigate the impact of outliers with trend relating to individuals and/or observation points.

A limited number of scenarios were used to investigate the performance of outlier detection methods. Sensitivity analyses were used, varying the sample size and test variability estimates; however, the ability of these methods may vary with study design and test performance. The combined effect of small sample size (see Chapter 5) and outlier detection methods has not been fully investigated.

6.6 Conclusions

In the situation when data are known to be incorrect (an 'error' outlier) because of implausible values, be it due to an error in performing the test, storing a sample or recording a result, then these data should be discarded and with remaining values used for analysis and the process for removing data carefully reported. However, if data are plausible, the removal of such data should be treated with extreme caution as these outlying data may be informative (an 'interesting' or 'influential' outlier). These types of outliers require further consideration.

The analysis presented has shown often used methods of outlier detection (particularly the methods advocated by Fraser-Harris) can lead to bias in the results of biological variability studies by underestimating the variability which may impact how tests are used, hence studies using these methods should be interpreted with caution. Also, the analysis of data including outliers showed different methods performed better depending on the number of outlying measurements and the magnitude of the difference between the outlying data and the rest of the data. Different methods of outlier detection may be selected based on the number of outliers considered. The ± 3SD rule worked well, only providing invalid results for the most extreme outlier simulation and introducing less bias than the other methods able to appropriately eliminate outliers (Cochran C test, Cochran C test partial, Fraser-Harris method and Tukey IQR rule) at the analytical, within-individual and between-individual level.

It is recommended the data obtained in biological variability studies is viewed prior to using outlier detection methods, with these methods only employed if visualisation of the data shows cause to investigate. It is also recommended if outlier detection methods are used and this leads to the removal of data, the analysis is performed on both the original and modified data and any differences are commented on or both sets of results are displayed. If an outlier detection method is required, the ± 3SD rule is recommended.

Chapter 7

A review of monitoring-related methodology literature

This work has been partly presented in the following form:
Selby PJ, Banks RE, Gregory W, Hewison J, Rosenberg W, Altman DG, Deeks JJ, McCabe C, Parkes J, Sturgeon C, Thompson D, Twiddy M, Bestall J, Bedlington J, Hale T, Dinnes J, Jones M, Lewington A, Messenger MP, Napp V, Sitch A, Tanwar S, Vasudev NS, Baxter P, Bell S, Cairns DA, Calder N, Corrigan N, Del Galdo F, Heudtlass P, Hornigold N, Hulme C, Hutchinson M, Lippiatt C, Livingstone T, Longo R, Potton M, Roberts S, Sim S, Trainor S, Welberry Smith M, Neuberger J, Thorburn D, Richardson P, Christie J, Sheerin N, McKane W, Gibbs P, Edwards A, Soomro N, Adeyoju A, Stewart GD, Hrouda D. Methods for the evaluation of biomarkers in patients with kidney and liver diseases: multicentre research programme including ELUCIDATE RCT. Chapter 5: A review of monitoring-related methodology literature. Programme Grants for Applied Research 2018.

Summary

The review of monitoring and monitoring related literature yielded few methods directly applicable to monitoring of progressive or recurrent disease.

Many of the methods identified are in the related areas of screening, biomarker development and monitoring for treatment titration purposes. Work in screening may be adapted to identify appropriate monitoring frequency and applications of the signal to noise approach for choosing decision rules and thresholds.

Biases identified in areas related to monitoring also need to be considered when using tests to monitor progression and recurrence of disease.

7.1 Introduction

The purpose of this work is to identify methods for the analysis of data collected from the monitoring of progressive and recurrent conditions and the design of monitoring strategies for subsequent evaluation. There are also related areas that may have methods for analysis and design that can be adapted to be applicable for monitoring of progressive and recurrent conditions.

Research in the area of monitoring has concentrated on treatment titration. Whilst the processes involved in monitoring treatment differ from monitoring patients with potential for progressive or recurrent disease, the general structure of the data is the same and there are lessons to be learnt from the research that has been done. There is some methods literature in the area of screening and, whilst screening is carried out in asymptomatic individuals rather than patients known to have disease that may progress or recur, there are obvious parallels between screening and monitoring. These methods are not currently used when planning screening strategies, with economic evaluations preferred to inform processes, however work is currently ongoing to improve the selection of screening strategies.[115] The literature on reference change values and statistical process controls may also contribute to the development

of methods for monitoring of progressive and recurrent disease.

7.2 Aims and objectives

The aim of this review was to understand the current methods available to design monitoring strategies. To investigate methods in monitoring and related areas searches were developed to identify papers developing and evaluating monitoring and screening strategies, introducing time dependent measures of test performance, explaining methods to evaluate measurement change from test variability and health economic approaches to developing testing protocols.

7.3 Methods

Methodological information related to monitoring was firstly sought from the first edition of the book 'Evidence-based medical monitoring', edited by Paul Glasziou, Les Irwig and Jeffrey Aronson. Key textwords related to monitoring methodology were identified and purposive searches of MEDLINE were undertaken from 2000 to 2010 (searches conducted on 26 March 2010). Reference tracking and citation tracking using Science Citation Index were used to identify additional relevant literature.

A variety of searches were performed to identify relevant literature using various combinations of the following text words:

(i) monitor*;

(ii) measure* or biomarker* or marker*;

(iii) serial or repeat* or periodic or longitudinal or trajectory*;

(iv) recurrence or progression;

(v) rule* or threshold* or trigger;

(vi) statistical process control or control chart* or reference change value or critical difference;

(vii) screen* with frequenc* or intensit* or interval*.

The searches retrieved 2,578 papers for review. These papers were screened by a single reviewer, and only papers reported in the English Language were considered. After reviewing 72 articles were included initially. Summaries from biomarker development papers identified by the search are not presented here. Full details have been published.[116]

The papers selected for review were summarised and full details can be seen in Appendix E E Table E.1.

7.4 Results

Limited methodological literature was identified providing guidance for the design of studies to evaluate monitoring tests. Work has focussed more on analytic techniques to assist with the design of monitoring strategies; primarily through analysis of existing data in order to make recommendations on monitoring frequency or decision rules, or simulation work, with both approaches specific to the disease area researched.

The identified literature can be categorised as:

- Development and evaluation of monitoring strategies
  - Linear mixed effects models and estimation of signal to noise ratio;
  - Joint modelling of longitudinal and outcome data;
  - Non-linear mixed models;
  - Alternative modelling approaches;
  - Simulation studies;

  - Analysis of cohort data;

- Screening;
- Time-dependent ROC curves;
- Differentiating measurement change from measurement variability;
- Health economic approaches
  - Decision analytic models;
  - Real options approaches.

7.4.1 Development and evaluation of monitoring strategies

The literature in the area of monitoring focussed on modelling approaches (linear mixed modelling, joint modelling and non-linear modelling). The review of the literature also identified simulation studies and work on the evaluation of monitoring strategies.

Most identified papers provided methods (or examples of methods being applied) to assess monitoring data via linear mixed modelling, signal to noise ratio analysis and joint modelling. Articles considering design of monitoring strategies were rare and tended to consider test frequency rather than decision rules and test thresholds. See Table 7.1.

Table 7.1: Focus of monitoring papers.

Concept	Number of papers
Design	
Test frequency	8[9,11,29,31,117–120]
Test thresholds	2[9,117]
Decision rules	4[9,29,119,120]
Other	4[121–125]
Analysis	
General data structure	2[12,124]
Linear mixed effects modelling/SNR	14[9–12,14,117,118,123,126–131]
Joint modelling	10[29,31,119,132–138]
Review of methods	2[12,132]
Other	7[124,125,130,131,138–140]

7.4.1.1 Linear mixed effects models and estimation of signal to noise ratio

General description of models

Stevens et al[12] reviewed statistical models used for the control phase of monitoring and explained how models can be fitted to observed monitoring data providing details of maximum likelihood methods, moment-based methods and literature based methods, where parameter estimates are obtained from reviewing the literature. They introduced a generic model for monitoring data, defining Y_{it} as the observed monitoring values including assay noise and variability, and U_{it} as the 'true' underlying and unobserved values. $U_{it} = \alpha_i + \beta_{it}$ and $Y_{it} = U_{it} + \omega_{it}$ where α_i is the true value at time 0, β_{it} is the change in the true value over time and ω_{it} is random error,[12] see Chapter 1.

Stevens et al[12] discussed the analytical methods required to give the proportion of positive tests that are truly positive, normality is assumed allowing the marginal and joint distribution of observed and unobserved values to be used. With more complex models (correlation structures and distributional assumptions) analytical methods become more difficult to use and simulation may be favourable. The intercept and gradient can be simulated using their modelled distributions allowing the unobserved underlying values to be generated. The error term can be simulated also and combined with the underlying values to give the observed values. Calculations are then made by comparing the underlying true values with the observed values.

Signal and noise

Glasziou and colleagues[121] questioned the need for randomised controlled trials of monitoring under certain circumstances, instead promoting understanding the background variation and evaluating the signal to noise ratio when assessing treatment effects. They suggested large estimated treatment effects would be required to demonstrate an effect.

Modelling methods are used with repeated test data in the hope of distinguishing 'signal' from 'noise'. The 'noise' is normal fluctuation in test results for patients (caused by the measurement variability of the test) and the 'signal' is a change in test results signifying a

Box 7.1: Signal and noise monitoring examples.

Thompson and Pocock[126] presented their findings following the analysis of repeated serum cholesterol measurements on 14,600 men and women. Their work focussed on the impact of within-individual variability on screening and monitoring. Using the cholesterol data a single observed measurement did not reflect the true underlying measure. Thompson and Pocock showed how the probability of a measure being classed as 'high' varied with the true underlying value, and whether the classification was based on a single measure or the mean value of multiple measures. The use of multiple measures was shown to improve classification. The authors identified regression to the mean when analysing multiple measures and variability in the measures for untreated individuals over time leading them to doubt whether repeated measures would be able to identify the benefit of treatment. The authors state the use of repeated measures could be 'very discouraging' for some patients.

Buclin et al[9] defined two decision rules that could be used to guide the treatment of patients with HIV infection with antiretroviral treatment based on CD4 cell measurements using a review of longitudinal analyses of CD4 cell trajectories. The first decision rule was a 'snap-shot rule', dependent on a single CD4 measure, and the other was a 'track-shot rule, where multiple CD4 measures are required. The devised rules were tested using clinical data, to minimise false findings, and recommendations were made regarding the frequency of testing.

Bell and colleagues[122] developed a framework to identify when monitoring of initial response to treatment would be beneficial using data from RCTs. The findings showed monitoring of initial response to treatment would only be useful when there is variation in treatment effect between patients and not all treated patients achieved results at the therapeutic level targeted.

Other examples of the use of mixed modelling and signal to noise ratio estimation, to understand when it is appropriate to monitor response to treatment, thresholds or monitoring frequency were: monitoring of cholesterol,[14,122,127] bone mineral density,[123] blood pressure,[10,117,128,129] lipids,[118] and diabetes.[11]

true change in disease state, see Box 7.1 for examples of using these models.

7.4.1.2 Joint modelling of longitudinal and outcome data

Joint latent class models

When fitting a joint latent class model subjects are split into a finite number of latent subgroups. The trajectory of biomarker measurements and the risk of event are specific to each latent class, with the model allowing for the dependency of biomarker values and risk of

event. A multinomial logistic regression model is used to assign subjects to subgroups. A linear mixed model is used to model repeated biomarker measurements given the assigned latent class of the subject; and, a survival model is used to model the time-to-event, again given the latent class of the subject. The model is fitted using maximum likelihood estimation.[132]

Examples of the use of joint latent class models can be seen in work by Proust-Lima and Taylor[133] and Li and Gatsonis,[31] both applied to monitoring with prostate specific antigen (PSA) prostate cancer recurrence, see Box 7.2.

Box 7.2: Joint latent class models example.

Proust-Lima and Taylor[133] discussed the derivation of a posterior probability of recurrence from a joint class linear model to identify a 'dynamic prognostic tool of recurrence'. The posterior probability obtained from the joint latent class model gives the probability of an event occurring between time s and time $s + t$ (where the subject is event free at time s). Estimating the probability of an event after a certain time requires fitting survival models to subjects at each time estimated with only covariates available at time s. As biomarker data are often discrete, imputation techniques are used to allow predictions of an event to be obtained at multiple time points. Proust-Lima et al also discussed the validation of predictive tools and the lack of consensus in this area.

Li and Gatsonis[31] used a joint latent class model to develop a strategy that modifies monitoring intervals. Li and Gatsonis used a two-stage approach when fitting the joint latent class model, where the model used to identify latent classes was fitted separately. Bayesian Information Criteria was used to select the number of classes. The two-stage approach has the advantage of being less computationally intensive. The uncertainty of latent class assignment was evaluated using multiple imputation assuming latent class was missing completely at random. For prospective studies the two-stage procedure is repeated as new information is collected (measures, events and study end). Li and Gatsonis demonstrated the method using simulated PSA measurements for 150 patients with prostate cancer with testing to identify recurrence. Predictions from the model inform a utility function which was used to identify the appropriate monitoring intervals for each patient. The expected value of the utility function used is $\tilde{U}_t = aP(\text{event at time t})$ where a is a negative value if the event occurs and zero otherwise. The optimal monitoring interval can be identified for individuals or groups of patients; the authors advocated optimising by latent class as these intervals can then be adapted for new patients.

Bayesian hierarchical change-point models

Bayesian hierarchical change-point models model the trajectory of test results prior to, at the time of and after the onset of disease simultaneously. Models also allow for the within-individual correlation, as individuals have multiple test measurements, between-subject vari-

ation in trajectories and the random change-point. A piecewise or segmented linear model is used where the parameters of the model are the trajectory of test results prior to the change-point, the test result value at the time of the change point, the time of the change-point and the trajectory of results after the change-point; each of the parameters is a random effect within the model. Non-informative prior distributions are used for the parameters in the model[134,135] see Box 7.3.

Box 7.3: Bayesian hierarchical change-point models examples.

Slate and Turnbull[134] demonstrated the use of Bayesian hierarchical change-point models analysing PSA data. The authors state the advantages of using Bayesian hierarchical change-point models are 'borrowing of strength' when estimating parameters specific to individuals whilst also accounting for the correlation of measures and, by obtaining posterior distributions using Gibbs sampling, the model can give the probability an individual has reached the change-point.

Bellera et al[135] stated additional advantages of this type of modelling are the ability to provide precise estimates compared with simpler models, the parameters used by the model are all of clinical importance, estimates of test measurement variability can be estimated as a function of test result value, and the model is flexible. Bellera and colleagues, however, commented that the model can be influenced by the timing of and the number of test measurements for individuals with the potential for this to cause bias, as participants with more results will provide more information for the model and participants with more test results may be different to those with fewer test results. Subsequent work by Bellera et al[119] used an empirical simulation approach (see Section 7.4.1.5 for further detail).

Inoue et al[136] combined longitudinal PSA measurements from three different studies using a non-linear Bayesian hierarchical model. At the individual level a non-linear model was used to model PSA over time and the hierarchical model component then accounted for the variability between studies.

7.4.1.3 Non-linear mixed models

Non-linear mixed models allow more flexibility in modelling as linearity of the parameters is not necessary which may be appropriate for modelling longitudinal test data in some conditions. Multiple measures for each individual can also be accounted for by non-linear mixed models with the incorporation of random effects, see Box 7.4 for examples.

Box 7.4: Non-linear models examples.

Subtil and Rabilloud[137] used a non-linear mixed model to evaluate PSA measures to identify recurrence of prostate cancer after initial treatment. To model the trajectory of PSA measures a non-linear exponential decay growth model was used. This model contained parameters for the intercept of PSA trajectory and the trajectories before and after recurrence, with the parameters included as random effects. The model was empirically derived and was used to model the exponential increase in PSA occurring at the time of recurrence. The non-linear model was preferred to the change-point model, in this case, as the non-linear model was thought to be more flexible and allowed for exponential increase of PSA measures after recurrence. Subtil and Rabilloud discussed the potential issues introduced by data showing random fluctuation in PSA measures and suggested using the student t-distribution and a Dirichlet process assumption to both reduce the weighting of outlying PSA measures used in the model and allow the distribution of the random effects to be non-normal.

Taylor et al[138] demonstrated a modelling approach, again in the area of PSA monitoring after treatment for prostate cancer. The approach used a model for cure, a model of PSA measures and a model of clinical events. For the modelling of serial PSA measurements a non-linear hierarchical model, similar to the model used by Subtil and Rabilloud,[137] was used.

7.4.1.4 Alternative modelling approaches

Alternative modelling approaches were used by Thiébaut et al[130] and Wolbers et al[131] in the area of CD4 cell monitoring to understand when to initiate treatment for patients with HIV, see Box 7.5. These approaches combined piecewise linear models with Cox models.

7.4.1.5 Simulation studies

With knowledge of the progression of disease and the variability of the test used to monitor the disease, data can be simulated allowing the evaluation and comparison of decision rules and testing frequency, see Box 7.6.

The approaches outlined by Sölétormos et al[120] and Bellera et al[29] used the biomarker values with measurement error excluded as the underlying disease state without any link between biomarker values and true disease state, as measured by the gold standard. The ability of decision rules were assessed without accounting for the meaning of the underlying biomarker values in terms of disease state.

Box 7.5: Alternative modelling approaches examples.

Thiébaut et al[130] modelled time to an AIDS event or death after the initiation of highly active antiretroviral therapy for patients with HIV, using a two-stage approach. Measures of CD4 cell count and HIV RNA were firstly modelled using a piecewise linear mixed model with a slope estimated from the initiation of treatment until two months and from two months after treatment onwards. This model was used to account for measurement error and avoid making further assumptions of the observed measures. The results of this model were then used as time dependent covariates in a Cox proportional hazards model.

Wolbers et al[131] modelled the survival of patients with HIV after the initiation of combination antiretroviral therapy with emphasis on the prognostic value of CD4 cell count change prior to beginning treatment. For patients with two or more CD4 cell measurements, a slope was estimated using linear mixed effects modelling (sensitivity analyses also use joint modelling of CD4 measures and separate estimates for each patient) with the value of the slope then used as a prognostic factor in a Cox proportional hazards survival model, modelling time from initiation of combination antiretroviral therapy to AIDS event or death.

7.4.1.6 Analysis of cohort data

Existing longitudinal data sets can allow comparison of monitoring approaches. In some disease areas data sets from multiple cohort studies exist and it is possible to combine the data and analyse the pooled group of patients. With information on the patients within the cohort studies it is possible to use the combined data as a proxy for trial data and compare potential monitoring strategies or components of strategies, see Box 7.7 for details.

7.4.1.7 Evaluation of monitoring strategies

Measuring the performance of a monitoring strategy is different, and more complex, compared to testing at a single time point, due to repeated testing and the potential for patients to change disease state.

DeLong et al[139] discussed the sensitivity and specificity of a monitoring strategy in the situation where patients were permitted to change disease state between monitoring test applications and patients with disease are removed from the sample after their first result whilst diseased, as the purpose of the test would be to detect the onset of disease. DeLong and

Box 7.6: Simulation studies examples.

Sölétormos et al[120] used simulation to compare rules used for monitoring of cancer antigen 15.3 (CA15.3) and carcinoembryonic antigen (CEA) to identify metastatic breast cancer. The approach simulated observed monitoring results for individuals with progression and those in a stable state. Simulations were produced using 48 permutations of background variation value and cutoff/starting concentration value; 12 permutations of background analytical and biological variation, and progression rate given concentrations start in the middle of a reference interval; and, 12 permutations of background variation value and progression rate, given concentrations start above the criteria cutoff. Previously established criteria were then compared using the simulated data. Sölétormos and colleagues stated there is great potential for using simulation to compare progression criteria, although the approach demonstrated used only specific values of starting concentration, analytical and biological variation, and progression rate. The results showed the extremes of the abilities of the guidelines rather than how these would perform in the population they would be used to monitor, and, furthermore, only established guidelines were tested.

Bellera et al[29] used an empirical simulation approach to estimate the sensitivity and specificity of a decision rule stating three consecutive rises in PSA measure should constitute a positive result indicating possible recurrence of prostate cancer. Firstly, Bellera and colleagues used Bayesian hierarchical change-point modelling (for further detail see Section 7.4.1.2) to analyse longitudinal PSA measures for a cohort of patients. The result of this modelling process was considered the true underlying trend in PSA and the modelling also provides an estimate of the variability of PSA measurements, allowing empirical simulation of observed results. Sensitivity and specificity were then estimated by comparing the simulated results with the modelling results. The authors stated an advantage of this approach to be the flexibility in evaluating decision rules based on consecutive increases or decreases in results as the performance of decision rules can be evaluated for different frequencies of observations, underlying progression and variability. The limitations of this study were that the process requires a longitudinal data set for analysis that may not be available for all situations, only decision rules based on trends (increases and decreases) are investigated, and this approach would not be appropriate in cases where increases (or decreases) in the test measures are expected and the objective of monitoring is to identify progression at the point of a clinical event rather than distinguishing stable disease from non-stable disease.

colleagues used partial likelihood to estimate test performance summaries. Other approaches of estimating time-dependent sensitivity and specificity for ROC curve analysis are described in Section 7.4.3.

Li and Gatsonis[31] provided guidance on the evaluation of monitoring strategies, specifying the reporting of:

Box 7.7: Pooled analyses examples.

The When To Start Consortium[124] used data from multiple cohort studies to investigate when to start combination antiretroviral treatment for patients with HIV-1. Treatment was started either when the CD4 cell count of a patient was at a higher value (threshold 1) or treatment would not be given until the patient reached a lower CD4 cell count (threshold 2). As there was limited evidence comparing the use of the two strategies analysis of cohort studies was sought. The data available to the research team consisted of cohort studies prior to and after the use of combination antiretroviral therapy. Kaplan-Meier estimates of the probability of progression to AIDS or death based on time from initiation of treatment and starting CD4 cell count were obtained using the data from treated patients. Simple hazard ratios of AIDS and death were also calculated using Cox regression, with a more complex method involving imputation also used (method introduced by Cole et al[140]) allowing for lead time bias–extra time from initiation of treatment to AIDS or death due to patients receiving treatment earlier–and unseen events for those not receiving treatment until reaching a lower CD4 cell count (threshold 2). The data collected prior to the use of combination antiretroviral treatment was used to estimate the distribution of time from reaching threshold 1 to worsening to reach threshold 2. The probability of progressing to AIDS or death before having a CD4 cell count low enough to reach threshold 2 was modelled also. Progression rates were compared and random effects models used to estimate the decrease in CD4 cell count for the period prior to antiretroviral therapy and for the period where antiretroviral therapy was used.

Ahdieh-Grant et al[125] used the cohort approach, again, as a proxy for a trial looking to identify the appropriate CD4 cell level at which to begin antiretroviral treatment for patients with HIV. The authors draw attention to the disadvantage of using data obtained by cohort studies as there is no randomisation. Ahdieh-Grant and colleagues also discussed the issue of lead time bias, which they accounted for by splitting the time to progression to AIDS into time from CD4 cell count being in a certain range to beginning antiretroviral treatment and beginning antiretroviral treatment to the onset of AIDS. Time to event analysis was used to analyse the cohort data, comparing the patients starting antiretroviral treatment at high and low CD4 cell count levels.

(i) the total number of tests required for each patient;

(ii) the difference in how early the new strategy can detect an event compared to the comparator strategy;

(iii) the detection rate (events detected by monitoring/events) and error rate (symptomatic detected events/events);

(iv) the cost of monitoring examination and any required confirmatory tests;

(v) and, the percentage of monitoring detected length

$$PMLD = \frac{\text{monitoring detected time} - \text{time at change point}}{\text{time at symptomatic event} - \text{time at change point}} \times 100,$$

where the time at change point is the time the biomarker begins to change (a proxy for the time of event).

7.4.2 Screening

The aim of screening is to benefit patients by detecting disease prior to the onset of symptoms as is the case with monitoring. Figure 7.1 describes the development of disease and the screening process. The detectable pre-clinical stage of disease is the time when screening may detect asymptomatic disease; this is also known as the sojourn time. The delay time is the period of the sojourn time when the screening has not detected disease and lead time is the period of sojourn time after screening has detected disease. The greater the lead time, the greater the potential benefit of screening.

Walter and Day[141] discussed the biases that need to be considered when analysing screening data. Firstly, the population that participate in screening may vary from the population that do not participate in screening, as they may be at higher or lower risk of having the disease the screening process aims to detect. This is likely to be less of an issue for monitoring populations, although it is conceivable that there will be differences between those who do participate in monitoring and those who either drop out (perceiving themselves to be at low risk of the event in question) or who demand some form of treatment (perceiving themselves to be at high risk of the event in question). Other biases that may affect monitoring studies include length-biased sampling and lead time bias. Length-bias sampling occurs as patients with more aggressive disease will be in the pre-clinical phase of disease where screening will detect disease (sojourn time) for a shorter length of time than those with less-aggressive disease. Screening is most likely to detect cases with a longer sojourn time, hence cases of less-aggressive disease which will likely have a better prognosis. Lead time bias is when survival times for screened cases appear to be greater than survival times for cases identified

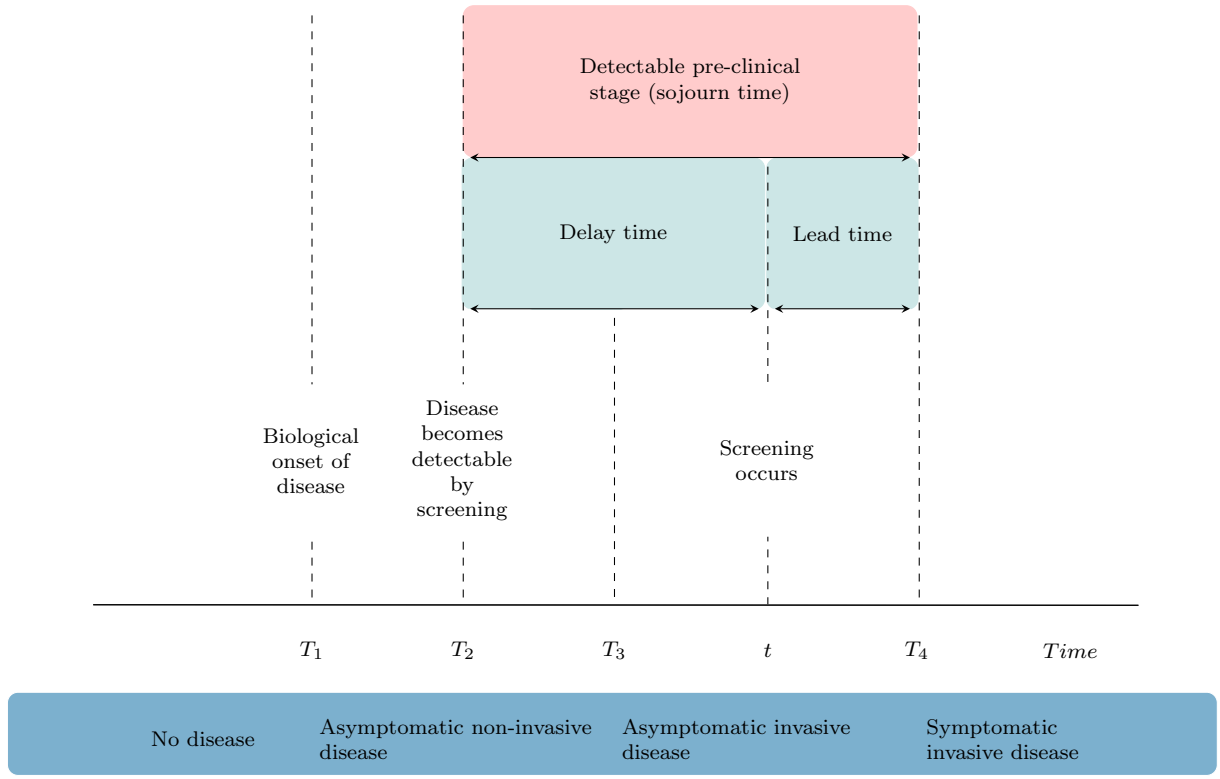


Figure 7.1: Process of screening patients (after Walter and Day[141]).

by different means when there is no difference in survival; the only difference is cases identified by screening are detected earlier.

Work has focussed on estimating the duration of the pre-clinical stage of disease (Walter and Day,[141] Day and Walter[142] and Etzioni and Shen[143]), which enabled further work into the optimal frequency of screening (Zelen[144], Lee and Zelen,[145] Frame and Frame[146] and Lee et al[147]). Others have considered how to set the optimal decision rule for a screening strategy using a new test when the length of the sojourn period is not known (McIntosh and colleagues[148] and McIntosh and Urban[149]), see Boxes 7.8 to 7.10.

7.4.3 Time-dependent ROC curves

When a test gives a binary result (positive or negative) the performance of the test is usually assessed by calculating the sensitivity and specificity. When the result of a test is a continuous value the performance of the test is evaluated for various cut offs by calculating the sensitivity

Box 7.8: Estimating the duration of the pre-clinical stage of disease.

Walter and Day[141] presented a method that uses the prevalence of disease at the time of screening and the incidence of disease in the time between screening visits to estimate a distribution for the time in a pre-clinical state and the sensitivity of screening. The benefit of this approach is length biased sampling is accounted for. Walter and Day discussed the effect of screening on the incidence rate of a population. The authors assumed that prior to screening commencing the incidence rate will remain at a constant level. After a screening test has taken place the incidence rate reduces and initially the majority of new cases of disease will consist of individuals that have a false negative result from the previous screening test. Until the next screening point the incident rate will then gradually increase. Fewer cases will be seen at each successive screening point, with the incidence rate between screening points also being lower as screening continues. If screening were to stop the incidence rate would gradually increase to the level prior to screening. Using these observations Walter and Day derived an expression for the incidence at time t given the previous false negative results of the population and the sojourn time, and from this the distribution of lead time can be obtained. Later, Day and Walter[142] published further detail of these expressions which allowed lead time distributions to be estimated for screening strategies other than the strategy used when the data set was collected. The authors used these expressions in analysing the Health Insurance Plan of Greater New York (HIP) study of breast cancer where approximately 62,000 women were randomised to be offered screening or receive no screening.[141,142]

Etzioni and Shen[143] developed the work of Walter and Day[141] by using the false negative rate, as this varies with the sojourn time, to estimate the asymptomatic period of cancer. Estimates were produced using non-parametric methods with the EM algorithm utilised.

and specificity of the test at each possible result value and plotting sensitivity against 1-specificity, a receiver operating characteristic (ROC) plot. The ROC plot and the areas under the curve produced can then be used to assess the performance of the test and identify optimal thresholds for the use of the test in practice. When allowing for time in ROC analysis, time-dependent ROC methods are used.

Pepe and colleagues[30] undertook a review of time-dependent ROC curves. The definition of sensitivity of a test is dependent on the time when the test is performed. As it is assumed diseased cases will present with positive test values early in the testing process it is thought sensitivity will decrease with time. Pepe et al also discussed cumulative sensitivity, which would provide the sensitivity for a test for an interval of time, and how this can be derived. The false positive fraction, or 1- specificity, is problematic to define as the disease status of individuals can change over time making it difficult to classify individuals as diseased or non-

Box 7.9: Optimal screening frequency.

Zelen[144] expanded the ideas of Day and Walter[142] to calculate the optimal frequency of screening. Zelen described the three health states of screening; patient free of disease, patient with pre-clinical disease and patient with clinically diagnosed disease. The screening period in which a patient enters the pre-clinical disease phase was then used in deriving the incidence and the probability of being in the pre-clinical disease phase at time t. The benefit of Zelen's approach over Day and Walter's approach is the effect of cases present at the initial screening point is accounted for. To use the derived expressions to identify optimal screening frequencies Zelen used a weighted utility function assigning different weights to cases observed at the first screening point, subsequent screening points and between screening points (cases not identified by screening). A different weight was used for the initial screening point as the initial screening point will be most affected by length biased sampling. Using the assigned weights an equation was given to calculate the optimal frequency of screening observations. Zelen also provided a proof of equal time between screening only being optimal if the sensitivity of the screening test is equal to 1.

Lee and Zelen[145] developed the work on optimal screening schedules by Zelen[144] and proposed threshold and sensitivity methods to dictate appropriate screening times. The threshold method required screening to be undertaken at time points such that the probability of being in the pre-clinical detectable state is at a pre-assigned level. The sensitivity method defines the screening schedule based on the ratio of the number of cases the screen is expected to detect to the number of cases that will be detected over the specified duration of screening. However, Frame and Frame[146] argued the risk of disease should not be a factor in evaluating appropriate screening frequencies. Instead, Frame and Frame advocated determining screening schedules based on knowledge of the progression of disease and the sensitivity of the screening test used and gave a mathematical expression for the error of a screening strategy (the proportion of cases of pre-clinical disease not identified by screening, E), $E = (1 - S)^{\frac{W}{F}}$, based on the relationship between the sensitivity of the screening test (S), the sojourn time (W) and the frequency of screening (F).

Lee and colleagues[147] further developed the methods of Lee and Zelen[145] by incorporating mortality into the model for selecting optimal screening frequency strategies. To predict the difference in mortality for different screening strategies the model used the idea of stage-shift with the key assumption that survival is improved for cases identified by screening as they are detected in at an early stage of disease compared with cases identified by usual care. The model predicting mortality assumes the natural history of disease is progressive and benefits of early detection are due to stage-shift in diagnosis.

diseased, especially in situations where all individuals will have an event at some point. One approach is to choose a time point specific to the context assessed and individuals are treated as non-diseased if they are free of the event at the specified time (the static false positive fraction). Another approach is to allow the false positive fraction to vary with the time

Box 7.10: Optimal decision rules for novel tests.

McIntosh and colleagues[148] and McIntosh and Urban[149] provided a method for deciding the optimal decision rule for a screening strategy when using a new test, when the length of the sojourn period is not known. An algorithm using parametric empirical Bayes (PEB) theory was discussed. Test values from healthy individuals collected over time were used to estimate total, between-individual and within-individual variance and the trajectory of test values. Using the estimates of trajectory and allowing for heterogeneity in the population to be screened the PEB algorithm can produce person specific thresholds to maintain a specified false negative rate (FNR) to be used at the next screening point. The authors state the use of the PEB algorithm to generate person specific thresholds allows earlier detection of disease without increasing the burden on healthy individuals.

since the test was performed (the dynamic false positive fraction). When using the dynamic false positive fraction test performance may be misleading as a positive result will be falsely positive shortly before an individual develops disease. For tests with continuous results time-dependent ROC curves compare individuals with and without disease at each time point. If using the dynamic false positive fraction ROC curves are difficult to interpret due to the non-diseased group changing over time (see Boxes 7.11 and 7.12 for examples).

7.4.4 Differentiating measurement change from measurement variability

The variability of repeated test measures for an individual can be broken down into three components: pre-analytical variability, analytical variability and individual variability.[16,56] The coefficient of variation (CV) is calculated by dividing the standard deviation by the mean and is commonly used in place of standard deviations as this allows for a reference change value (RCV) to be calculated to reflect percentage changes rather than absolute changes. Given the values of analytical and within-individual variation, a difference between two results greater than the RCV suggests a real change in condition.[155] For further detail see Chapter 2.

Statistical process control methods (first developed by Shewhart) are often used in manufacturing and can be used for medical applications when a process can be measured directly or via a biomarker. Statistical process control procedures measure variability across time, where variability can be split into common cause and special cause variability (or assignable

Box 7.11: Sensitivity and specificity.

Cai et al[150] presented equivalent time dependent definitions of sensitivity and 1- specificity, but with the emphasis on the time of an event occurring, defining sensitivity and 1- specificity to be functions of time relative to the time of disease or time of an event. The authors stated most research in the area assumed the test and assessment of disease status are carried out simultaneously raising the issue of the predictive accuracy of a test dependent on the time it is carried out in comparison to the onset of disease, assuming an increase in accuracy if the test is used closer to the time of an event. Cai and colleagues also fitted semi-parametric models using longitudinal test data to separately estimate sensitivity and 1- specificity.

Zheng and Heagerty[151] discussed sensitivity and specificity for time-dependent ROC analysis as functions of both the time of testing and the time of event. Zheng and Heagerty also discussed the difference between estimating incident and prevalent ROC curves, restricting their work to incident ROC curves.

Subtil et al[152] discussed how incident sensitivity requires a test to be performed a given number of days prior to the onset of disease and offers a way of taking the variation of time between individuals receiving a test and developing disease into account. Subtil and colleagues introduced a Bayesian method to allow for the interval-censored measurements. Results of using this method compared with the method without adjustment suggested the 'crude' method underestimates sensitivity.

Parker and DeLong[153] provided a method to convert estimates of sensitivity and specificity for monitoring tests for ROC curve analysis. The estimates of sensitivity and specificity used are those introduced by DeLong et al,[139] which were derived using partial likelihood estimation under the assumption that diseased participants can have at most one test result when in the diseased state.

cause variability). Special cause variability is akin to signal and signifies true change in the disease state of an individual. Common cause variability, as noise, reflects random variability in measures.[155]

X-bar charts are used to display measurements over time for an individual. If a process is stable, measurements are expected to fluctuate around the mean, and the standard deviation of observed measures is expected to be constant over time. Estimates of the mean (μ) and standard deviation (σ) can be taken from stable processes, with an unbiased estimate of the standard deviation obtained using a moving range (the difference between consecutive measures) and dividing the mean of the moving range estimates by a constant ($d_2 = 1.128$). Estimates of the mean and standard deviation of a stable process can then be used to identify

Box 7.12: Modelling to produce time-dependent ROC curves.

Slate and Turnbull[134] reviewed methods used to analyse repeated test data when the test is used to screen or monitor the onset of disease in a population. These methods are used to estimate the ROC curve for each test and the resulting ROC curves are compared. The review discussed the use of time dependent Cox proportional hazards modelling, joint modelling of longitudinal test data and time of diagnosis, Weibull methods to model two time events, random effects models and integrated Onstein-Uhlnbeck (IOU) stochastic processes, multi-state models and Markov models, and change-point models.

Zheng and Heagerty[151] discussed a semi-parametric regression approach used to estimate ROC curves and an approach based on asymptotic distribution theory, which will allows covariates to change the distributional shape of test results.

Etzioni et al[154] introduced and demonstrated two methods for modelling the effect of lead time on the ROC curve. The first approach required modelling of longitudinal test data and then using parameter estimates from the model the ROC curve can be estimated at varying time points. The second approach directly modelled the ROC curve as a function of covariates including time of the test relative to the time of diagnosis. Etzioni and colleagues discussed how the methods can be adapted to compare two tests. The first method required separate fitting of models using data for the two tests with comparison of the derived ROC curves after; whereas, the approach of modelling the ROC curve directly more easily allowed for comparison of tests. Other advantages of the direct modelling approach were: fewer distributional assumptions with the method using the ranking of data points, robustness and flexibility and ease of implementation.

control limits.

The control limits can be identified using many criteria and should be modified depending on the situation; it may be target values are safety driven. Moving range charts and exponentially weighted moving average charts (moving averages are calculated with greater weight given to the most recent observations) are also used in similar ways. The variability of a process can be quantified using the capability index, the difference between the upper and lower limit divided by 6σ. The off-target ratio, $S_T = (\mu - T)/\sigma$ where T is the target value, measures how far the process is from the specified target value in terms of standard deviations. Process control charts use the assumption of independent normally distributed outcomes and generally require at least 20-25 observations.[155]

See examples in Boxes 7.13 and 7.14, and issues with these methods in Box 7.15.

Box 7.13: Methods accounting for test variability.

Sölétormos et al[28] used a rule based on an RCV in a computer model for monitoring of progression to metastatic breast cancer with cancer antigen 15.3 (CA15.3), carcinoembryonic antigen (CEA) and tissue polypeptide antigen (TPA).

Smellie[57] questioned the use of the 5% significance level when evaluating differences between results, as 10% or higher levels may be more appropriate in some situations.

Petersen[156] commented on how the use of RCV only considers the type I statistical error rate and not power, which is linked to the magnitude of the change in values for an individual. Petersen et al[157] discussed the importance of distinguishing between reference intervals and decision limits.

Klee[158] reviewed methods for setting analytical performance goals; these methods include the use of guides produced by regulations and external assessment, biological variation limits, needs of clinicians, subsequent testing and medical decision models. Klee identified differences in the performance limits identified by the different approaches.

7.4.5 Health economic approaches

7.4.5.1 Decision analytic models

Decision analytic modelling evaluates the cost, outcomes and cost effectiveness of interventions. In the case of repeated testing appropriate techniques need to be used for this evaluation, see Box 7.16.

7.4.5.2 Real options approaches

Palmer and Smith[169] introduced real options approaches, inspired by methods used in financial markets, which aim to include the uncertainty around the use of a new technology along with health economic evaluation. The approach uses the potential to delay introducing a new technology (akin to a change in management) and the irreversibility of using a new technology. Analyses factor in deferring using a technology and the better evidence that may be available after deferral using expected value of perfect information methods (EVPI), see Box 7.17.

Box 7.14: Statistical process control and statistical rules for interpretation of sequential tests.

Tennant et al[159] reviewed studies where patients were monitored using statistical process control methods and compare the use of statistical control methods with currently used rules and guidelines. Clinical areas found to use process control methods were: peak flow measurements for patients with asthma, blood pressure measurements for patients with hypertension and serum creatinine measurements for patients after undergoing a kidney transplant.

Thor et al[160] also reviewed studies using statistical control processes to monitor patients and highlight the disadvantages of using these methods. Thor and colleagues discussed that in some studies methods had been employed where there was a clear lack of understanding. The authors also commented on issues with auto-correlated measures, collection of data and application of the methods.

Gavit et al[161] discussed a slightly different approach to process control in change point analysis. Change point analysis uses cumulative sum charts of the difference between the mean value and the recorded value. Change points were then analysed as bootstrapping methods are used to generate a confidence interval for the change point. The change point method can also be used to identify differences in variability. An advantage of the change point method is the ability to analyse non-normal data due to the lack of distributional assumptions. Gavit and colleagues also claimed the change point method is able to identify subtle changes that would not be picked up by control charts.

7.5 Summary and conclusions

This review revealed limited methodological literature around the design of monitoring strategies. Work has focussed primarily on analysis of data where subsequently recommendations of monitoring frequency or decision rules could be made or simulation work performed, with both approaches being specific to the disease area researched. The area of screening has developed methods with the focus being identifying the optimal frequency of screening which could be used for designing monitoring strategies. There was some work on the design of biomarker development studies which could potentially be adapted to allow for the evaluation of a monitoring strategy using previously collected specimens. It appeared thresholds are often developed by analysis of the variability of the test being used, identified by the literature describing signal to noise ratio, biomarker development studies, statistical process control and reference change values.

The study by Buclin et al[9] showed an approach where decision rules were devised by a review

Box 7.15: Issues for methods using test variability.

Omar et al[162] discussed how useful reference change values are. Omar and colleagues commented on issues regarding the timing of observations, as the CV_I can increase as the interval between tests increases, and auto-correlation of serial measurements. Omar and colleagues also discussed that within-individual variability is commonly the largest component of RCV and how estimates of CV_I are usually available in the literature. Omar et al also explained the issues with these studies as estimates of CV_I are often for healthy participants and this may not reflect the variability of participants in stable disease.

Biosca et al[163] reported a study of biological variability to identify the appropriate RCV to use for long-term monitoring of renal post-transplantation. The research group had already conducted a biological variation study to determine the RCV values required for short-term monitoring. These studies showed little difference between RCVs required for short and long term monitoring; however, the RCVs obtained from studies using an appropriate population were very different to those detected using healthy participants.

A review of the accuracy and prognostic literature of cardiac natriuretic hormone (CNH) assays by Clerico and Emdin[164] highlighted differences in analytical sensitivity across studies carried out in differing populations. The results of accuracy measures were difficult to compare due to different gold-standards.

Box 7.16: Decision analytic models examples.

Karnon et al[165] reviewed models for measuring the cost effectiveness of screening regimes for breast, cervical and colorectal cancer. The review covered approaches including: decision trees, Markov models, MIcrosimulation for SCreening ANalysis (MISCAN), discrete event simulation (DES) and more complex approaches, such as the Baker[166] and Parmigiani[167] approaches.

Sutton et al[168] introduced comprehensive decision modelling, a combination of evidence synthesis and decision modelling, to integrate the analysis of diagnostic test performance and cost-effectiveness.

of the literature and then using an obtained data set and analysis of signal and noise the rules were refined to minimise false results. Following this, recommendations of the decision rule and frequency of monitoring could be made. Takahashi et al,[117] Takahashi et al[118] and Oke et al[11] also used signal and noise methods when analysing data and subsequently recommendations can be made for future monitoring strategies.

A number of applications of the signal and noise approach[9–11,14,117,118,121–123,126–129] were identified, largely in the area of treatment titration. The limitations of this approach for

Box 7.17: Real options approaches examples.

Driffield and Smith[170] further explained how these methods can be used to understand the benefit of 'watchful waiting' with some monitoring of disease progression rather than immediate treatment. The method was demonstrated using the example of the management of abdominal aortic aneurysms. Meyer and Rees[171] further developed the approach by allowing the incorporation of patient aversion by using a Poisson process.

Shechter et al[172] used a real options approach with a Markov decision model incorporated to identify the optimal time to cease monitoring and begin treatment. Shechter and colleagues also presented an example, looking at treatment for patients with HIV, of the errors made when the development of future technologies is not taken into consideration.

Whynes[173] introduced a similar approach looking at identifying the optimal time to move from monitoring to treatment considering the problem as a cost-minimisation exercise. Lasserre et al[174] used this type of method to investigate when to begin antiretroviral treatments in patients with HIV-infection.

monitoring disease progression or recurrence are that rules and thresholds are devised purely by analysing the variability of test measures and the minimisation of false findings rather than detection of disease at the earliest point possible and the impact on patients.

The simulation approach proposed by Li and Gatsonis[31] uses a joint latent class model which combines predictions from the model along with a utility function to identify optimal monitoring frequencies. The results of a simulation study reported by Li and Gatsonis appeared promising; however, the approach has not been widely adopted perhaps due to the complex nature of the model. Other simulation approaches may also have potential under certain circumstances, particularly if measurement error, and a link between biomarker values and true disease state can be included.

The biases that are well documented in the screening literature are applicable to the area of monitoring also. Length-time bias and lead time bias should be considered when analysing monitoring data and when designing monitoring studies. There is also the issue of post-screening noise which is again important to take into consideration when evaluating a monitoring strategy; the time point at which monitored and non-monitored patients are compared should be selected to minimise the issue of incidence after the final testing point and should also consider the number of likely events. Harm to patients is vitally important in screening

and monitoring as this harm may be at several time points and this must be thought of when designing strategies.

A further consideration for the analysis of monitoring data concerns the number of test measurements and the timing of test measurements:[135] people with more results will contribute more data to the model but they may be very different to those with fewer results. Measurement error and particularly biological variability also requires consideration. Studies have shown that reference change values from biological variability studies of healthy participants are not necessarily reflective of the true RCV for a diseased population. As methods to derive test thresholds used in monitoring rely heavily on the variability of test results it is important that estimates from biological variability studies are accurate. Also, the quality of studies undertaken when developing new biomarkers is not always rigorous; however, there is literature concerning the design of these studies and the evaluation of quality of these studies that may in time improve quality.

7.5.1 Limitations

This review was not systematic, meaning this search will not have identified all methods papers relevant to the area of monitoring. The searches were performed in 2010, so new methods may have been published and missed. To limit missing new developments the literature has been evaluated as it becomes available (published articles and conferences) and included if relevant, and sources working in screening have been able to provide information regarding current practice.[115]

7.5.2 Application to thesis: methods used

A simulation approach using the general model of Stevens and colleagues[12] was used in subsequent work (see Chapter 8) to investigate optimal monitoring strategies (decision rule, threshold and duration of monitoring) for a trial evaluating a monitoring test.

A simulation approach was chosen as data were required to fit the unique situation of the

corresponding trial (see Chapters 1 and 8). This approach used a model with a random intercept and a random slope. Simulation of the latent test values and observed test values allowed multiple candidate strategies to be compared, evaluating all aspects of the monitoring strategy (test frequency, decision rule, test threshold) simultaneously. Simulation approaches had been used to compare monitoring strategies, as identified in this review. The approach used in this thesis allowed the latent and observed values to be modelled, taking account of the measurement error of the test. The importance of measurement error estimates used in the developed model were investigated through sensitivity analyses, but the analyses were driven by differences in potential patient outcomes rather than simply controlling the false findings, as used in other monitoring studies. The chosen approach also allowed the link between true disease state and the biomarker to be modelled and fully incorporated.

For the evaluation of monitoring strategies, the criteria introduced by Li and Gatsonis[31] was modified.

The methods identified for screening were considered but not utilised fully as they only allowed optimisation of individual components of the monitoring strategy. The biases identified in the screening literature were taken into account when developing the monitoring model and analysing the data. The issues raised by Bellera et al[135] regarding the difference in participants contributing large and small amounts of data to models was also considered when analysing the data from the trial.

Chapter 8

Simulating monitoring data and evaluating monitoring strategies

This work has been partly presented in the following form:
Selby PJ, Banks RE, Gregory W, Hewison J, Rosenberg W, Altman DG, Deeks JJ, McCabe C, Parkes J, Sturgeon C, Thompson D, Twiddy M, Bestall J, Bedlington J, Hale T, Dinnes J, Jones M, Lewington A, Messenger MP, Napp V, Sitch A, Tanwar S, Vasudev NS, Baxter P, Bell S, Cairns DA, Calder N, Corrigan N, Del Galdo F, Heudtlass P, Hornigold N, Hulme C, Hutchinson M, Lippiatt C, Livingstone T, Longo R, Potton M, Roberts S, Sim S, Trainor S, Welberry Smith M, Neuberger J, Thorburn D, Richardson P, Christie J, Sheerin N, McKane W, Gibbs P, Edwards A, Soomro N, Adeyoju A, Stewart GD, Hrouda D. Methods for the evaluation of biomarkers in patients with kidney and liver diseases: multicentre research programme including ELUCIDATE RCT. Chapter 7: Simulating monitoring data and evaluating monitoring strategies. Programme Grants for Applied Research (2018).

Summary

There is a need to develop monitoring strategies based on evidence of effectiveness in the monitored population.[2] An example was presented using the Enhanced Liver Fibrosis (ELF) biomarker in managing patients with known liver fibrosis, specifically the ELUCIDATE trial.

Existing data and expert opinion were used to estimate the progression of disease and the performance of repeat testing. Knowledge of the true disease status in addition to the observed test results for a cohort of simulated patients allowed various monitoring strategies to be implemented and evaluated. The modelled data was validated by comparing to data from the trial.

The monitoring strategy utilising a prediction from a linear regression model in the decision rule and the monitoring strategy using a simple threshold decision rule performed similarly well. The results of sensitivity analysis showed the importance of accurate data to inform the simulation. Monitoring data can be simulated and strategies compared given adequate knowledge of disease progression and test performance. The simulated data compared well to the trial data.

8.1 Introduction

Although patient monitoring is a fundamental function of healthcare, incurring considerable cost to health care providers, the underlying methodology of monitoring is under researched[2,175] and there is an increased need for monitoring strategies to be systematically developed incorporating known likely progression of disease and the performance of the monitoring test to be used. Dinnes et al reviewed the evidence base for prostate specific antigen (PSA) monitoring to identify recurrence of prostate cancer.[1,176] The review identified the lack of a systematic approach in developing a monitoring strategy, with monitoring intervals based on standard follow-up schedules and limited evidence of consensus for the thresholds used to initiate treatment.[1]

Stevens et al[12] discussed various statistical models of the transition between the maintenance and re-established control phases of monitoring (the process of detecting when a disease is out of control leading to a change in management, for example treatment or more intensive monitoring) and identified a general statistical model for the evolution of monitoring data over time outlining possible sources of variation.[12] This general statistical model proposes the form of monitoring data based upon the observed values of sequential monitoring tests, the values of measurement error and other sources of variability, and the true disease state, which can be modelled based on epidemiological evidence but never observed, see §1.3.2.

This general model along with existing data, and evidence gathered from the literature, can be used to simulate monitoring data and allow the evaluation of strategies for a given target condition. The potential effect of monitoring strategies can then be evaluated and ranked, prior to full-scale investigation.[5]

The example presented investigates the use of the Enhanced Liver Fibrosis (ELF) biomarker in monitoring patients with known liver fibrosis. This modelling work was done alongside a prospective multicentre randomised trial (the Enhanced Liver fibrosis (ELF) test to Uncover Cirrhosis as an Indication for Diagnosis and Action for Treatable Events (ELUCIDATE) trial[177]). The ELUCIDATE trial evaluates ELF for the early detection of progression from liver fibrosis to liver cirrhosis compared to routine care, with the aim of enabling earlier treatment and potentially improved patient outcomes. Participants with known liver fibrosis and meeting a minimum ELF level at registration were recruited to the study and randomised to receive either standard care or standard care and monitoring with the biomarker Enhanced Liver Fibrosis (ELF). In the ELUCIDATE trial the ELF biomarker was used to detect progression in patients with liver disease so patients could be managed appropriately to prevent complications of cirrhosis. Participants randomised to the monitoring arm received an ELF test every six months and had a positive monitoring test if the result was 9.5 or above. Those with a positive monitoring test received a change in patient management and those with a negative result went forward to the next monitoring point with their management unchanged, see Figure 8.1. As the monitoring trial was performed simultaneously, this provided the opportunity to use information to inform the model building and allowed validation of

the model, by comparing the simulated data to the real study data. Designing a monitoring strategy using all available evidence should be performed prior to the conduct of a monitoring study evaluating the strategy.

8.2 Aims and objectives

The aims of the study were to:

- develop a model for simulating and evaluating monitoring data.
- use these data to identify the optimal monitoring strategy, from candidate monitoring strategies, for patients known to have liver fibrosis receiving repeated testing using the biomarker ELF. Candidate strategies were selected and evaluated to:
  - compare the alternative frequencies of monitoring (6 month or 12 month intervals).
  - evaluate the benefit of using targeted retesting compared to no retesting.
  - compare decision rules (positive results based on crossing a threshold determined by a single value (snapshot simple threshold rule), and track-shot rules based on absolute or relative increases from first test value, absolute or relative increases from last test value and prediction from a linear regression model).
- assess the validity of the model by comparing the estimated performance of the strategies, using the simulation, to results from the trial.

8.3 Methods

The method used to simulate data followed the model introduced by Stevens et al.[12] The process involved simulating the 'true' underlying and unobserved data and, including measurement error to generate the observed data.

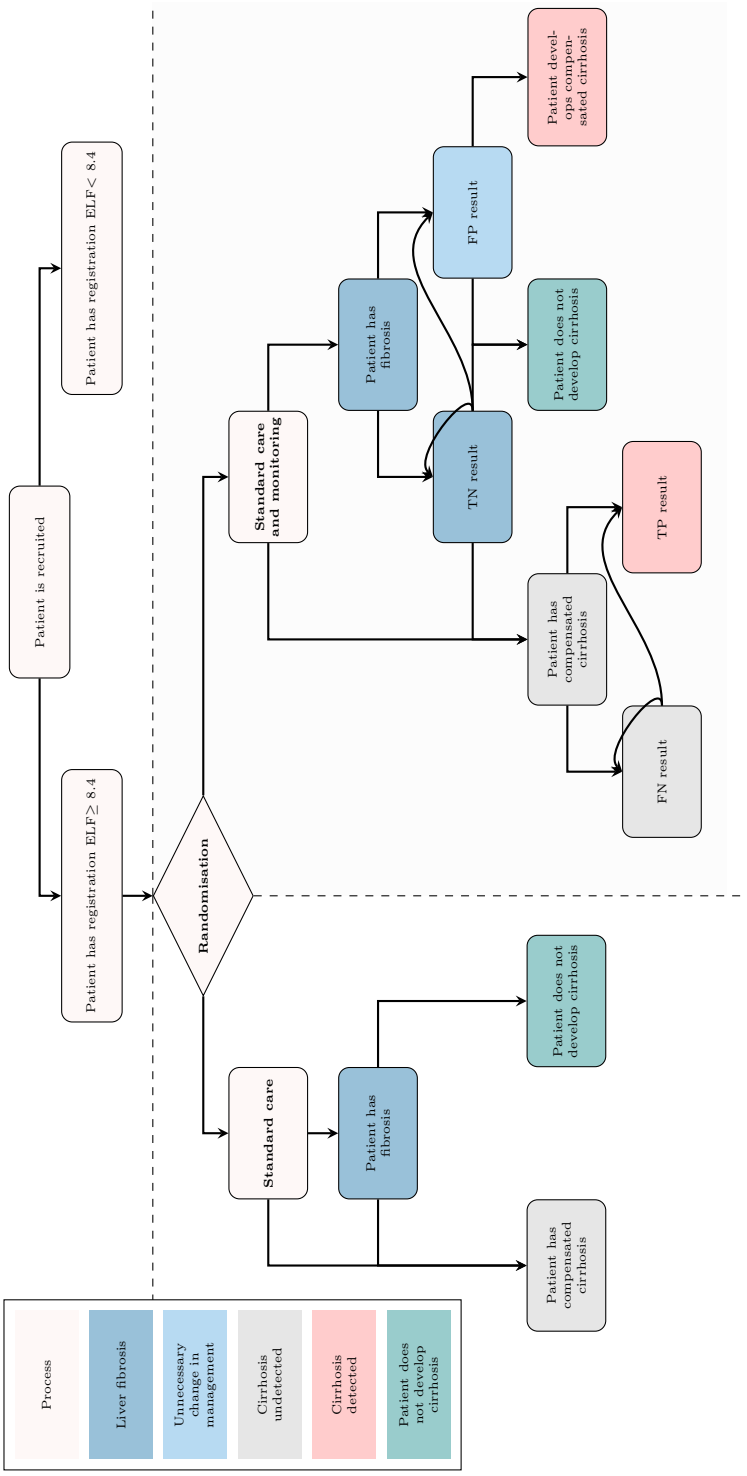


Figure 8.1: The design of the ELUCIDATE study.

- A model was used to generate the underlying and unobserved disease progression (giving 'latent' biomarker values), incorporating estimates of disease progression and the variability of these estimates, for a cohort of simulated individuals.

- Observed test result values were generated using the latent disease progression values and estimates of test performance.

- Selected monitoring strategies were then evaluated and compared, using both the observed test values and the true disease status.

Explanation of the notation used in the model can be seen in Table 8.1.

Table 8.1: Monitoring simulation model notation.

Description	Notation
Individual number	i
Number of initially simulated individuals	n
Time across fibrosis stages	t
Fibrosis progression rate	p_i
Mean fibrosis progression rate	μ_p
Standard deviation of fibrosis progression rate	σ_p
Fibrosis stage	s
Starting fibrosis stage	S_i
ELF value each stage of fibrosis	E_s
Mean ELF value at each fibrosis stage	μ_s
Standard deviation of ELF value at each fibrosis stage	σ_s
Observed mean ELF value at each fibrosis stage	μ_{Y_s}
Observed standard deviation of ELF value at each fibrosis stage	σ_{Y_s}
ELF at the beginning of each fibrosis stage	E_{is}
Gradient of ELF progression	β_{is}
Latent 'true' ELF	U_{it}
Measurement error	ω_{it}
Standard deviation of measurement error	σ_{A+I}
Observed ELF	Y_{it}

8.3.1 Simulation of true disease progression

The model simulated true disease progression, generating a random slope and random intercept in terms of fibrosis stage for each patient.

8.3.1.1 Fibrosis progression–random slope

The rate at which each individual patient progresses through the fibrosis stages was assumed to be constant, throughout the stages of fibrosis. The progression rates between patients varied; this was assumed to be normally distributed, $p_i \sim N(\mu_p, \sigma_p^2)$, where $i = 1, \ldots, n$ and n is the number of simulated individuals; μ_p is the mean and σ_p is the standard deviation of fibrosis progression rate. Fibrosis progression rate was restricted to only positive values, by fixing p_i at 0.01 if $p_i \leq 0$, meaning only increases in fibrosis stage were simulated; however, as the increase is just 0.01 fibrosis units per year this effectively means these patients are in a stable fibrosis state. The data source used to provide an estimate of fibrosis progression is given in §8.3.3.1.

8.3.1.2 Fibrosis stage at entry–random intercept

Patients recruited to a trial would be in varying stages of disease at entry. Using data on the likely distribution of fibrosis stage for a population with known liver fibrosis, a multinomial distribution was used to simulate a starting stage for each individual S_i. The data source used to provide an estimate of fibrosis stage at entry is given in §8.3.3.2.

8.3.1.3 ELF value link to fibrosis stage

For each stage of fibrosis, the distribution of latent ELF values within fibrosis stage was assumed to follow a normal distribution, $E_s \sim N(\mu_s, \sigma_s^2)$, where s is the fibrosis stage and $s = 0, \ldots, 4$; μ_s is the mean value of ELF at each fibrosis stage and σ_s is the standard deviation of ELF at each fibrosis stage. The data source used to provide estimates for the ELF link to fibrosis stage is given in §8.3.3.4.

8.3.1.4 ELF progression between fibrosis stages

The model used fibrosis stage as a continuous value. To generate ELF values for each patient at all stages of fibrosis, ELF progression between consecutive integer fibrosis stages was assumed to be linear. It was assumed that patients would have ELF values at the same point of the normal distribution for each fibrosis stage (patients would remain a given number of standard deviations from the mean). To randomly select the point of the normal distribution that patients would follow, a value from the standard normal distribution was generated for each patient (z_i). The ELF value for each participant, at each stage of fibrosis was $E_{is} = \mu_s + (z_i \sigma_s)$, see Figure 8.2 (left).

8.3.1.5 ELF progression–random slope

The ELF values at the beginning of each fibrosis stage for each individual (E_{is}) and the rate each simulated participant progressed through fibrosis (p_i), were combined to calculate the increase in ELF per year. The gradient of ELF progression was $\beta_{is} = (E_{i,s+1} - E_{i,s})\, p_i$, for $s = 0, \ldots, 3$. The gradient of ELF progression after stage 4 was assumed to be the same as the gradient between stages 3 and 4, $\beta_{i3} = \beta_{i4}$. β_{is} is the random slope in terms of ELF progression. The latent ELF progression values for each stage and all time points from the onset of fibrosis was:

$$U_{it} = \begin{cases} E_{ij0} + \beta_{i0} x_t & \text{for} \quad \text{time in stage 0} \\ E_{ij1} + \beta_{i1} x_t & \text{for} \quad \text{time in stage 1} \\ E_{ij2} + \beta_{i2} x_t & \text{for} \quad \text{time in stage 2} \\ E_{ij3} + \beta_{i3} x_t & \text{for} \quad \text{time in stage 3} \\ E_{ij4} + \beta_{i4} x_t & \text{for} \quad \text{time in stage 4} \end{cases}$$

where t is time across all stages. This allowed the simulation of life time progression data for a cohort of patients; see Figure 8.2 (centre). ELF values were truncated at 0 if a negative value was simulated.

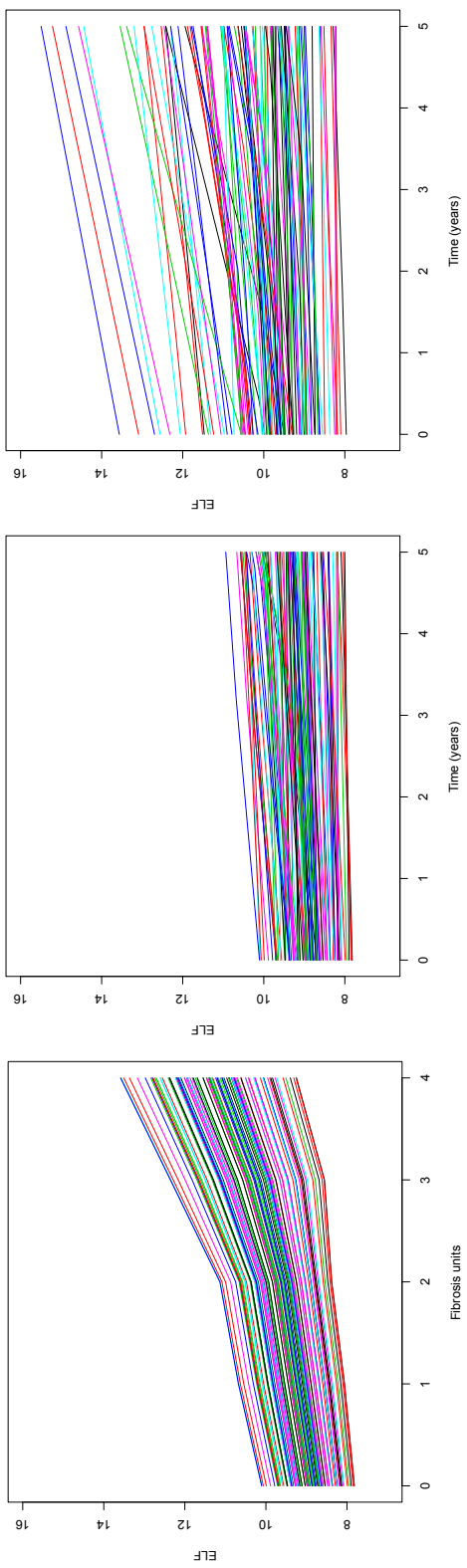


Figure 8.2: Illustration of the monitoring data simulation process. Left: fibrosis units linked to ELF value; centre: ELF progression through time, and right: starting stage adjusted ELF progression through time. Estimates of fibrosis progression are given as Scheuer fibrosis units per year, where Scheuer scores range from 0 to 4 and measure severity of liver disease with stage 0 showing no fibrosis and 4 showing liver cirrhosis.[178]

8.3.1.6 ELF value at entry–random intercept

The simulation was conducted to prevent patients from inclusion if they would be confirmed as having cirrhosis, or at a point of fibrosis they would not have reached in their lifetime due to their simulated progression rate. A maximum value was used to restrict the time patients had liver fibrosis and cirrhosis.

The time at registration for each participant was a randomly selected time point from the time period when the individual was in their generated fibrosis stage at study entry. If the participant was in stage 4 of fibrosis at entry, a random value from the interval of stage 4 starting to a point two years after was selected. The maximum time participants could have fibrosis was 20 years, see Table 8.2

In the ELUCIDATE trial, a registration ELF test was given to each patient to assess eligibility. The first ELF test included in the trial data was taken at the point of randomisation. The start of the trial was assumed to occur three months after registration.

Table 8.2: Trial consideration estimates used in monitoring simulation modelling.

Description	Estimate used in simulation modelling
Maximum time in cirrhosis before trial entry	To avoid simulating patients that are in advanced cirrhosis, the maximum amount of time a patient has been cirrhotic for is set to 2 years.
Maximum time in fibrosis before entry to the trial	To avoid patients being simulated at a point of disease they would not have reached given their fibrosis progression rate the maximum amount of time a patient has had fibrosis for at the time of entering the trial is set at 20 years.
Time between registration and randomisation	The time between registration ELF test and randomisation ELF test was estimated to be 3 months.
Trial duration	The duration of the trial used in all simulations was 5 years.
Time between test and retest measures	The time between a patient having a test and retest (if the original test was in the target range) was estimated to be 1 week.

8.3.1.7 Random slope and random intercept model in terms of ELF

The underlying disease progression for the simulated individuals, U_{it}, was used for the time period starting at study entry and continuing for the duration of the trial (5 years was used)

for simulated patients with an eligible registration ELF test value (see §8.3.2.2), see Figure 8.2 (right).

8.3.2 Simulation of observed values

The latent underlying ELF measurements were converted to observed ELF measures by the addition of measurement error.

8.3.2.1 Measurement error

The measurement error at each observation point (ω_{it}) was formed of within-individual variation and analytical variation. Measurement error was assumed to be normally distributed with a mean of zero, $\omega_{it} \sim N(0, \sigma^2_{A+I})$. The observed ELF measurement at any given time (Y_{it}) was the underlying measurement plus the measurement error $Y_{it} = U_{it} + \omega_{it}$. Values were adjusted to equal 0 if a negative observed ELF value was simulated. The data sources used to provide an estimate of ELF measurement error are given in §8.3.3.3.

8.3.2.2 Entry criteria

In order to fulfil trial entry criteria the observed ELF measurement at registration had to be greater than or equal to the pre-set value of 8.4. An example of simulated observed ELF measures can be seen in Figure 8.3.

8.3.3 Data sources

The data sources used to estimate fibrosis progression rate, fibrosis stage at trial entry, measurement error and ELF value link to fibrosis stage are described below and also in Table 8.3.

Table 8.3: Data used in monitoring simulation model.

	Data	Estimates used in model
Fibrosis progression rate	Poynard et al:[179] estimate of median fibrosis progression (units per year) 0.133 (95% CI 0.125, 0.143).	Estimate calculated from Poynard et al:[179] $p_i \sim N(0.13, 0.17^2)$ Estimate after adjustment (to be used in sensitivity analyses): estimate of fibrosis progression was increased to reflect expert opinion. $p_i \sim N(0.27, 0.17^2)$
Fibrosis stage at entry to trial	Cross sectional data set:[180] estimated proportion of patients in each stage. Stage 0- 0.25; stage 1- 0.35; stage 2- 0.13; stage 3- 0.15; stage 4- 0.12.	The cross-sectional data set was used:[180] $\rho_0 = 0.25, \rho_1 = 0.35, \rho_2 = 0.13, \rho_3 = 0.15, \rho_4 = 0.12$
Measurement error	Longitudinal data set: estimate of the standard deviation of measurement error of 0.81. Siemens: estimate of the standard deviation of total measurement error of 0.11.[181] ELUCIDATE registration and randomisation data: estimate of standard deviation of total measurement error 0.47.	Estimate obtained from ELUCIDATE was used. $\omega_{it} \sim N(0, 0.47^2)$
ELF link to fibrosis stage	Cross sectional data set:[180] estimates of ELF mean (SD) at each fibrosis stage. Stage 0- 8.82 (0.87); stage 1- 9.18 (0.96); stage 3- 9.55 (1.00); stage 4- 11.32 (1.47).	After adjustment: measurement error is accounted for to give the latent unobserved ELF values and modified to represent values for each stage. $E_0 \sim N(8.63, 0.73^2)$, $E_1 \sim N(9.00, 0.84^2)$, $E_2 \sim N(9.36, 0.89^2)$, $E_3 \sim N(9.91, 1.22^2)$, $E_4 \sim N(10.80, 1.39^2)$

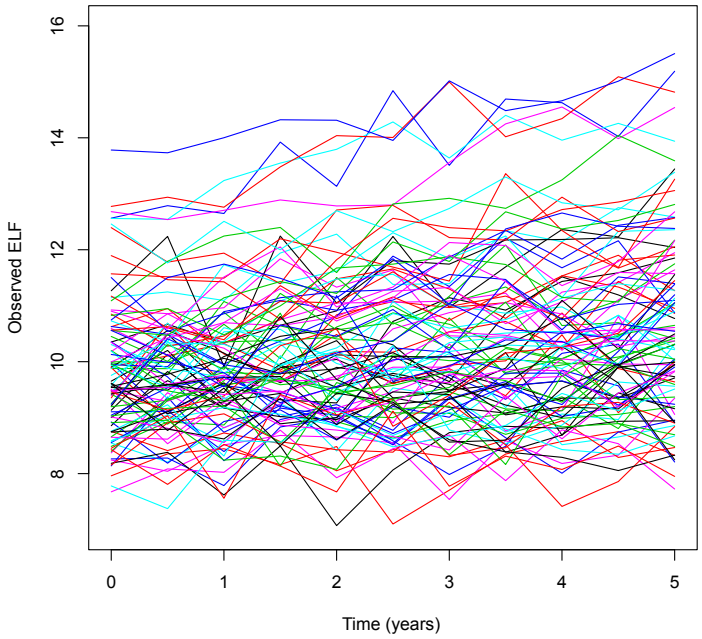


Figure 8.3: Observed ELF measures from monitoring data simulation.

8.3.3.1 Fibrosis progression rate

An estimate of the rate of fibrosis progression, based on data for 1,157 patients[179] with chronic hepatitis C followed up for over 40 years, was obtained from Poynard et al. When consulting clinical experts it was suggested the estimate provided by Poynard et al[179] was identified in a population that was not comparable with that of the ELUCIDATE study (participants in the Poynard study were thought to have less severe disease). The estimate from Poynard et al was used primarily in the simulation model with an adjusted estimate (doubled) used for sensitivity analyses. Estimates of fibrosis progression were measured using the METAVIR system, assumed to approximate Scheuer fibrosis units per year, where Scheuer scores range from 0 to 4 and measure severity of liver disease with stage 0 showing no fibrosis and 4 showing liver cirrhosis.[178]

8.3.3.2 Fibrosis stage at entry

A cross-sectional data set with ELF results and Scheuer fibrosis scores following liver biopsy for 921 patients[180] was used to identify the distribution of fibrosis stages in the cohort. This data set contained fibrosis scores from histology and ELF values for participants.

8.3.3.3 Measurement error

To estimate the error associated with each observed ELF test value, three data sources were considered. Firstly, a longitudinal data set with repeat ELF measurements (baseline and at 3 months) for 220 patients[182] was subjected to analysis of variance to identify variability at the individual level. Without repeated analyses of test observations for individuals, this variability was the combined analytical and within-individual variability (σ^2_{A+I}). The manufacturer also provided information on the measurement error of the ELF test (σ^2_{A+I}).[181] Details of the analysis and design of this study were unclear but it was understood that repeated measurement of a small sample of healthy volunteers were used. Due to discrepancies between the estimates from the two sources, data was obtained directly from the ELUCIDATE trial. Analysis of variance was used to obtain an estimate of the analytical and individual level variability for the registration and randomisation ELF values for the first 112 eligible participants; this estimate was used in the simulation model.

8.3.3.4 ELF value link to fibrosis stage

The cross-sectional data set[180] (including ELF values and fibrosis units from histology for 921 patients) was used to provide an estimate of the observed ELF value for patients at each level of fibrosis with a corresponding measure of variability (σ_{Y_s} is the observed standard deviation at each stage of fibrosis). To estimate the latent and unobserved standard deviation of ELF values at each fibrosis stage (the between individual variation, σ^2_s), the measurement error (σ^2_{A+I}) that would have been included in these observed measures was accounted for. As the latent ELF variability (σ^2_s) and the measurement error (σ^2_{A+I}) are independent in the simulation, $\sigma^2_{Y_s} = \sigma^2_s + \sigma^2_{A+I}$, so $\sigma_s = \sqrt{\sigma^2_{Y_s} - \sigma^2_{A+I}}$.

To estimate the latent and unobserved mean ELF value at each stage of fibrosis the observed estimates were assumed to give the mean value for the midpoint of the corresponding fibrosis stage.

8.3.4 Implementation of a monitoring strategy

The effect of implementing different monitoring strategies was predicted using simulated observed values of ELF. The specified monitoring strategy (decision rule, use of retesting and frequency of testing) changed the simulated observed values that would be measured and how the value or values for each individual would be interpreted.

8.3.4.1 Monitoring strategies

Simple decision rule (strategy A)

The simplest decision rule was based on a single value threshold (snap-shot rule). The threshold value was specified and any observed value over this threshold indicated a positive result for that participant at that time point.

Retesting (strategy B)

Patients with an initial test value within 1 ELF unit of the threshold value were subjected to retesting, an additional test would be carried out one week after the standard test. When a patient required retesting, the mean of the original test and the retest result was calculated and this value was subjected to the decision rules to identify positive participants. Patients with a value above the upper limit of the range were classed as positive on the initial test without further testing and patients below the limit of the retesting range were classed as negative using just the initial test. The retesting component could be used with any of the alternative decision rules.

Frequency of monitoring (strategy C)

The frequency of monitoring was varied by increasing the interval between monitoring tests, from six months to twelve months. Varying the frequency of monitoring could be used in conjunction with any of the alternative decision rules explained.

Alternative decision rules

Decision rules incorporating previous test results as well as the current result (track-shot rules) to identify positive patients were also considered. Absolute and relative increases from randomisation ELF or from the last recorded ELF measure were investigated. A rule using predictions from a linear regression model fitted using all available observed data points was also considered.

Decisions rules based on absolute and relative increases and the linear regression method required at least two observations to declare a participant as positive. A simple threshold rule was used to identify participants at the first monitoring point.

Absolute increase from start value (strategy D)–A result was considered positive when the absolute difference between the test value and the first recorded value for the patient was greater than the threshold.

Absolute increase from last observed value (strategy E)–A result was considered positive when the absolute difference between the test value and the last observed test value for that patient was greater than the threshold.

Relative increase from start value (strategy F)–A result was considered positive when the relative difference between the test value and the first recorded test value for the patient was greater than the threshold.

Relative increase from last observed value (strategy G)–A result was considered positive when the relative difference between the test value and the last observed test value for that patient was greater than the threshold.

Linear regression (strategy H)–The linear regression decision rule involved fitting a linear regression model to the data for each participant at each time point, using all available measures for that participant. The prediction from the model was then used to identify the patient as test positive or negative. A result was considered positive when the prediction (at that monitoring time point) from the linear regression model was greater than the threshold.

8.3.5 Evaluation of a monitoring strategy

To evaluate each strategy, the decision made from implementing the monitoring strategy using the simulated observed values and the corresponding latent values was assessed. With knowledge of the true underlying disease state of each participant, the performance of a variety of monitoring strategies was evaluated. These measures were adapted from the criteria introduced by Li and Gastonis.[31]

8.3.5.1 Comparison of observed results with the underlying disease state

Participants had a positive or negative test result based on the simulated observed data and the decision rule used. The test result was then found to be either true or false depending on the underlying disease state. The purpose of the ELF test was to identify when patients entered compensated cirrhosis, stage 4. As it may be beneficial to identify patients a short period of time prior to entering compensated cirrhosis, participants were classed as 'diseased' three months prior to this time point.

If, at a testing point, a participant was 'diseased', a positive result would be a true positive and a negative result would be a false negative. If, at a testing point, the patient was not diseased, a negative result would be a true negative and a positive result would be a false positive. As a positive result (true or false) caused a change in management and cessation of monitoring, patients with a positive result do not have a test result at subsequent monitoring times. Figure 8.4 illustrates how a strategy with a simple threshold decision rule can be evaluated.

8.3.5.2 Measuring the performance of a monitoring strategy

The performance of a monitoring strategy was assessed at each monitoring point by calculating the number of patients at each monitoring test visit, and the number of true positive, false positive, true negative and false negative test results.

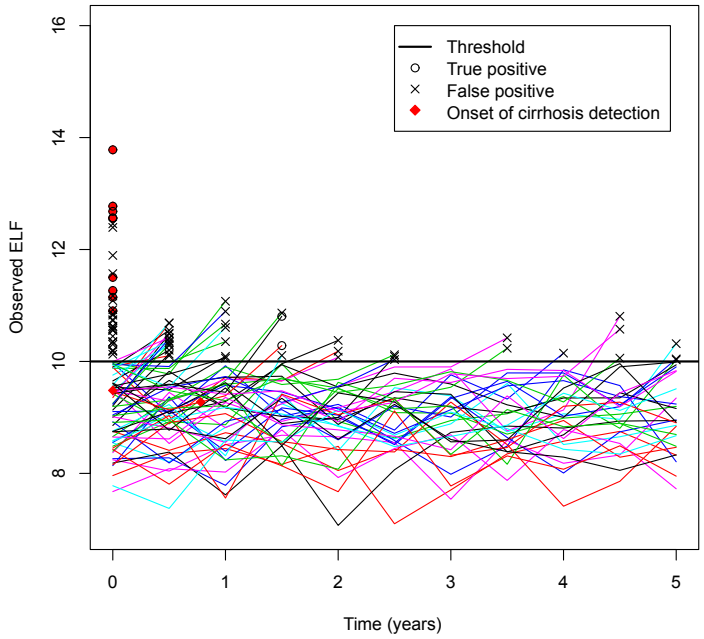


Figure 8.4: Implementing a monitoring strategy using simulated data.

The criteria used to asses strategies across all time points (adapted from the criteria introduced by Li and Gastonis.[31]) were:

- the number of tests carried out across the duration of the strategy–to represent resource use;
- the positive predictive value (PPV)–to investigate how likely it was for an individual with a positive result to be diseased;
- and, the time between the onset of compensated cirrhosis and a patient having a positive test result–to measure the potential patient outcome.

When comparing strategies the number of tests per person for the duration of monitoring, PPV (for all tests over the duration of monitoring) and percentage of patients with 'delayed diagnosis' (delay from onset of disease to diagnosis of over 12 months) were used to measure performance. To allow for comparisons to be made between strategies where only two of the three measures of performance ranged, thresholds used by monitoring strategies were varied to obtain a PPV of 25%. A PPV of 25% was chosen as this would be an acceptable PPV in practice.[183]

With the PPV kept constant, differences in the percentage of patients with delay to diagnosis of 12 months or more and the number of tests can be compared and appropriate strategies selected. The delay to diagnosis estimate indicates the impact of false negatives and the false positives are controlled by fixing the PPV.

8.3.5.3 Evaluation of strategies

The strategy evaluated first was the simple threshold strategy with observations every six months and no retest component (reference strategy). Alternative strategies were evaluated where individual components of the reference strategy were varied: the frequency of monitoring, the decision rule and whether a retest value was used. The same simulated data were used when evaluating strategies A to H, to allow a direct comparison.

Sensitivity analyses

Sensitivity analyses were carried out to estimate the effect of inaccurate information regarding test performance and progression of liver disease. Estimates of between-individual variability, measurement error and fibrosis progression rate were altered (halved and doubled) and the reference strategy (strategy A) was evaluated with all aspects of the strategy kept constant (including the threshold value). Analyses were performed with the threshold varied to give PPV of 25% using data with altered estimates also. Further sensitivity analyses were undertaken in which the fibrosis progression rate was adjusted based on expert opinion, these were analyses of: strategies A to H as for the main analysis (with PPV held at 25%), and the reference strategy using varied estimates to generate monitoring data.

Validation of the ELUCIDATE trial data

To assess the accuracy of the model, the mean and standard deviation of randomisation ELF values were calculated and compared for the ELUCIDATE and simulated data sets. Analysis of variance was used to assess between-individual and within-individual variability of ELF values recorded for patients in the trial and the simulated results. Multilevel models (accounting for the repeated observations for patients) were fitted, both with and without

estimating a random slope, using the simulated observed values and observed values from the ELUCIDATE trial (for participants with two or more ELF measures post registration) and the results from these models were compared. In the ELUCIDATE trial ELF measurements were not taken in the majority of cases after the participant had an ELF result of 9.5 or above. To allow for this the ELUCIDATE and simulated data sets were modified so that each patient with an ELF measure of 9.5 or above did not have any subsequent measures.

Sample size

Simulations were based on a cohort of 20,000 patients to give adequate precision. With 20,000 test results, if one of the performance measures gave an estimate of 15% a corresponding 95% confidence interval would range from 14.5% to 15.5%; for an estimate of 1.5% a 95% confidence interval would range from 1.3% to 1.7%.

8.4 Results

The same simulated data were used when evaluating strategies A to H, to facilitate direct comparisons. For the simulated cohort of 20,000 patients, 5,314 (26.6%) would develop cirrhosis.

8.4.1 Reference monitoring strategy (strategy A)

Table 8.4 shows the performance of the reference monitoring strategy at each testing time point. For the reference monitoring strategy (simple threshold, observations six-monthly and no retest component), the threshold required to maintain the PPV at 25% was an ELF value of 10.715. This was identified using an iterative process of modifying the threshold used in the simulation. The sensitivity and PPV calculated for the strategy were highest at the initial observation point and the percentage of tests with a positive result was also larger, due to cases being identified from a prevalent population at the initial testing point. The percentage of false negative results generally rises as the strategy continues in time. Over the duration

Table 8.4: Monitoring simulation results by observation point for the reference strategy (strategy A).

Observation month	**Tests (n)**	**TP results n (%)**	**FP results n (%)**	**FN results n (%)**	**TN results n (%)**	**Positive results n (%)**	**Diseased**[a] **n (%)**	**PPV (%)**	**Sensitivity (%)**
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	207 (1.2)	920 (5.5)	733 (4.4)	14742 (88.8)	1127 (6.8)	940 (5.7)	18.4	22.0
12	15475	147 (0.9)	661 (4.3)	740 (4.8)	13927 (90.0)	808 (5.2)	887 (5.7)	18.2	16.6
18	14667	102 (0.7)	558 (3.8)	783 (5.3)	13224 (90.2)	660 (4.5)	885 (6.0)	15.5	11.5
24	14007	113 (0.8)	495 (3.5)	786 (5.6)	12613 (90.0)	608 (4.3)	899 (6.4)	18.6	12.6
30	13399	104 (0.8)	446 (3.3)	813 (6.1)	12036 (89.8)	550 (4.1)	917 (6.8)	18.9	11.3
36	12849	127 (1.0)	458 (3.6)	805 (6.3)	11459 (89.2)	585 (4.6)	932 (7.3)	21.7	13.6
42	12264	121 (1.0)	429 (3.5)	828 (6.8)	10886 (88.8)	550 (4.5)	949 (7.7)	22.0	12.8
48	11714	129 (1.1)	449 (3.8)	817 (7.0)	10319 (88.1)	578 (4.9)	946 (8.1)	22.3	13.6
54	11136	135 (1.2)	390 (3.5)	806 (7.2)	9805 (88.0)	525 (4.7)	941 (8.5)	25.7	14.3
60	10611	125 (1.2)	369 (3.5)	814 (7.7)	9303 (87.7)	494 (4.7)	939 (8.8)	25.3	13.3
All	152724	2472 (1.6)	7411 (4.9)	8696 (5.7)	134145 (87.8)	9883 (6.5)	11168 (7.3)	25.0	22.1

[a] Tests performed when the patient was diseased.

of the monitoring strategy 7.64 tests per person (152,724 tests in total) were performed and 6.10% of all patients had delay to diagnosis.

8.4.2 Comparing strategies with changes to individual components to the reference strategy

Figure 8.5 and Table 8.5 show the performance of various monitoring strategies. Results of each strategy by observation point can be found in Appendix F Tables F.1 to F.7.

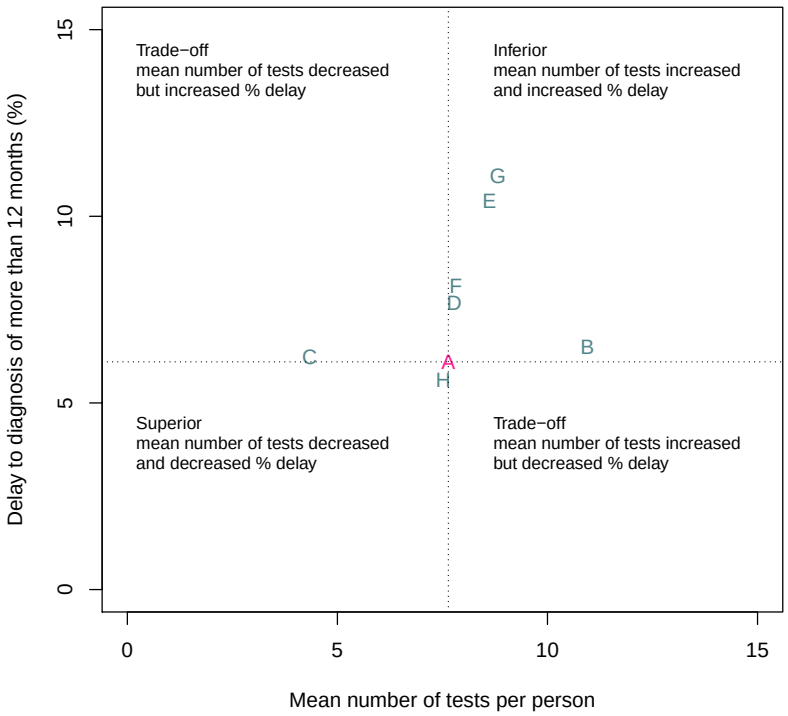


Figure 8.5: Performance of various monitoring strategies on simulated monitoring data with PPV of 25%. A is the simple threshold strategy; B is the retest strategy; C is the decreased monitoring frequency strategy; D is the absolute increase from initial value strategy; E is the absolute increase from last value strategy; F is the relative increase from initial value strategy; G is the relative increase from last value strategy; H is the linear regression strategy.

Table 8.5: Monitoring simulation results of various monitoring strategies.

	Monitoring strategy components					PPV	Tests[a]			Delay to diagnosis[c]			Test performance			
	Decision rule	Threshold value	Interval (months)	Retest	Initial threshold	%	Total	Mean pp[b]	Median (Q1, Q3)	N	% of all[d]	% of stage 4[e]	TP pp[f]	FP pp[g]	Positive n (%)	Sensitivity (%)
A	Simple threshold	10.715	6	FALSE	-	25	152724	7.64	11 (3,11)	1220	6.10	22.96	0.12	0.37	9883 (6.47)	22.13
B	Simple threshold	10.580	6	TRUE	-	25	218974†	10.95	12 (6,15)	1300	6.50	24.46	0.12	0.35	9406 (6.11)	21.15
C	Simple threshold	10.550	12	FALSE	-	25	86787	4.34	6 (2,6)	1249	6.25	23.50	0.13	0.38	10053 (11.58)	35.34
D	Absolute increase from initial value	1.295	6	FALSE	10.715	25	155648	7.78	11 (4,11)	1536	7.68	28.90	0.13	0.40	10598 (6.81)	19.85
E	Absolute increase from last value	1.460	6	FALSE	10.715	25	172363	8.62	11 (7,11)	2085	10.42	39.24	0.08	0.24	6305 (3.66)	8.45
F	Relative increase from initial value	1.144	6	FALSE	10.715	25	156460	7.82	11 (4,11)	1630	8.15	30.67	0.13	0.38	10266 (6.56)	18.12
G	Relative increase from last value	1.1795	6	FALSE	10.715	25	176385	8.82	11 (9,11)	2217	11.09	41.72	0.07	0.20	5338 (3.03)	6.68
H	Linear regression	10.495	6	FALSE	10.715	25	150478	7.52	11 (2,11)	1126	5.63	21.19	0.12	0.35	9342 (6.21)	21.58

[a] Tests over the duration of monitoring.
[b] Mean number of tests per person over the duration of monitoring.
[c] Patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).
[d] % of all patients with delay to diagnosis.
[e] % of patients that would reach cirrhosis within the trial period with delay to diagnosis.
[f] TP pp is the mean number of true positive results per person over the duration of monitoring.
[g] FP pp is the mean number of false positive results per person over the duration of monitoring.
† 218974 tests were carried out to generate 153971 results due to retests being used.

8.4.2.1 Inferior strategies

The retest strategy (strategy B) and the strategies with decision rules based on absolute and relative increases from the first and last recorded value (strategies D, E, F and G) were inferior to the reference strategy, requiring more tests and causing more patients who had progressed to liver cirrhosis to experience a delay to diagnosis, whilst achieving the same PPV.

The main effect of the retest strategy was to increase the number of tests performed (increase of 3.30 tests per person), with also a small increase in the percentage of patients with delay to diagnosis of 12 months or more (absolute increase of 0.40 percentage points). Whereas, the strategies with decision rules based on absolute and relative increases from an initial value showed only small increases in the number of tests required (increases of 0.14 and 0.18 tests per person respectively) but larger increases in the percentage of patients with delay to diagnosis of 12 months or more (absolute increases of 1.58% and 2.05% respectively) compared with the reference strategy. The absolute and relative increase from last recorded value decision rules both increased the number of tests required (by 0.98 and 1.18 per person, respectively) and increased the percentage of patients with delay to diagnosis of 12 months or more (to 10.42% and 11.09%).

8.4.2.2 'Trade-off' strategies

The reduced monitoring frequency strategy (strategy C) showed a 'trade-off' between delay to diagnosis and the number of tests required when compared with the results of the reference strategy. The number of tests required decreased by 3.30 tests per person and the percentage of patients with delay to diagnosis of 12 months or more increased by 0.15 percentage points (absolute increase) compared with the reference strategy.

8.4.2.3 Superior strategies

The reference strategy was found to be inferior to the linear regression strategy. The linear regression strategy used fewer tests (decrease of 0.12 tests per person) and had a lower percentage of patients with delay to diagnosis of 12 months or more (absolute decrease of 0.47%) when compared to the reference strategy.

8.4.3 Sensitivity analyses

8.4.3.1 Comparing results from the reference strategy when varying estimates of test performance and disease progression

Table 8.6 demonstrates the effect on the reference strategy of increasing (doubling) or decreasing (halving) parameter estimates (measurement error, between-individual variability and fibrosis progression rate). Results for each monitoring time point using these alternate estimates can be seen in Appendix F Tables F.8 to F.19. Analyses were performed with the estimates varied to understand the impact of incorrect data on the simulation; these scenarios were also analysed with the thresholds manipulated to give a PPV of 25%.

Improved estimates of test performance (decreased measurement error and decreased between-individual variability) both improved PPV (absolute increases of 4.6% and 8.6% respectively) and increased the number of tests required (increases of 0.73 and 0.91 tests per person) with decreased measurement error also increasing the percentage of patients with delay to diagnosis (absolute increase of 1.30%). Both increased and decreased between-individual variability reduced the percentage of patients with delay to diagnosis (absolute decrease of 0.72% and 2.12% respectively). An increased rate of fibrosis progression led to both increased PPV (absolute increase of 4.2%) and percentage of patients with delay to diagnosis (absolute increase of 1.52%) but decreased the number of tests required (decrease of 0.64 tests per person).

The largest difference in PPV was achieved by increasing the between-individual variability (absolute decrease of 8.8%); the largest difference in number of tests required was achieved by increasing measurement error (decrease of 1.84 tests per person); and the largest difference

Table 8.6: Monitoring simulation results of using the reference strategy when changing estimates required for data simulation *(Difference to reference strategy for original simulation data).*

Change in data simulation	**Threshold**	**PPV (%)**	**Mean number of tests per person**[a]	**Delay**[b] **(%)**	**Develop cirrhosis**[c] **n (%)**
None	10.715	25.0	7.64	6.10	5314 (26.66)
Decreased†measurement error	10.715	29.6 *(+4.6)*	8.37 *(+0.73)*	7.4 *(+1.30)*	5248 (26.24) *(-66 (0.33))*
	10.450	25.0	7.70 *(+0.06)*	5.73 *(-0.37)*	
Increased‡measurement error	10.715	17.60 *(-7.4)*	5.84 *(-1.80)*	3.85 *(-2.25)*	5421 (27.11) *(+107 (0.54))*
	11.365	25.0	7.65 *(+0.01)*	7.05 *(+0.95)*	
Decrease†between-individual variability	10.715	33.6 *(+8.6)*	8.55 *(+0.91)*	3.98 *(-2.12)*	5139 (25.70) *(-175 (0.88))*
	10.463	25.0	7.84 *(+0.20)*	2.28 *(-3.82)*	
Increased‡between-individual variability	10.715	16.5 *(-8.8)*	5.80 *(-1.84)*	5.38 *(-0.72)*	5272 (26.36) *(-42 (0.21))*
	11.905	25.0	8.26 *(+0.62)*	9.95 *(+3.85)*	
Decreased†fibrosis progression rate	10.715	22.7 *(-2.3)*	7.90 *(+0.26)*	4.95 *(-1.15)*	4440 (22.20) *(-874 (4.37))*
	10.860	25.0	8.29 *(+0.65)*	5.68 *(-0.42)*	
Increased‡fibrosis progression rate	10.715	29.2 *(+4.2)*	7.00 *(-0.64)*	7.62 *(+1.52)*	7689 (38.45) *(+2375 (11.88))*
	10.460	25.0	6.21 *(-1.43)*	5.63 *(-0.47)*	

[a] Mean number of tests per person over the duration of monitoring.
[b] % of all patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).
[c] Patients that would go on to develop cirrhosis in the monitoring duration if no intervention were received.
† Decrease is halving the estimate used in the original simulation.
‡ Increase is doubling the estimate used in the original simulation.

in the percentage of patients with delay to diagnosis was achieved by decreasing between-individual variability (absolute decrease of 2.12%).

Analyses were performed with the thresholds calibrated to give PPV of 25%. The biggest difference in delay to diagnosis of 12 months or more was changing between-individual variability. The biggest difference in the number of tests per person was caused by changing the fibrosis progression rate.

8.4.3.2 Adjusted fibrosis progression rate

Results of evaluating strategies based on data with the adjusted estimate of fibrosis progression can be seen in Appendix F Tables F.20 to F.41 and Figure F.1. Results for strategies, comparing to the reference strategy, appeared similar to results when using the unadjusted estimate.

8.4.4 Comparison to ELUCIDATE data

The ELUCIDATE data contained 705 observations taken from 420 participants randomised to the ELF monitoring arm of the trial. After removing measurements following an ELF value of 9.5 or above for each individual (akin to the trial setting), the simulated data set contained 66,320 observations for 20,000 participants and the simulated data set with adjusted fibrosis progression included 59,000 observations for 20,000 participants.

Analysis of the ELF value at the point of randomisation for each of the data sets showed similar results–mean (SD) for the ELUCIDATE data was 9.57 (1.21); for the simulated data 9.71 (1.15); and for the simulated data with adjusted fibrosis progression 9.83 (1.20)–with the mean value being slightly lower for the ELUCIDATE data than the two simulated data sets. The between-individual standard deviation was higher for the ELUCIDATE data than for the simulated data sets (0.93 for the ELUCIDATE data compared with 0.76 for the simulated data and 0.82 for the simulated data with adjusted fibrosis progression). The within-individual standard deviation was similar for the ELUCIDATE data and both simulated data sets (0.53

for the ELUCIDATE data and, 0.51 and 0.52 for the simulated and simulated with adjusted fibrosis progression data sets respectively). Results of analysis of randomisation ELF and analysis of variance on ELF at all recorded time points can be seen in Table 8.7.

Table 8.7: Comparison of monitoring simulation to trial data–results of analysis of randomisation ELF and analysis of variance for ELF measurements at all time points.

	ELUCIDATE data	**Simulated data**	**Simulated data with adjusted fibrosis progression**
Randomisation point ELF			
ELF mean (SD)	9.57 (1.21)	9.71 (1.15)	9.83 (1.20)
Analysis of variance			
Between-individual SD	0.93	0.76	0.82
Within-individual SD	0.53	0.51	0.52

The ELUCIDATE data modelled consisted of 429 observations from 153 participants, with each participant having a minimum of 2 and a maximum of 6 ELF observations and the average number of observations per person was 2.8. The number of observation points used from the simulation model was therefore capped to give a similar mean number of observations per person to the value seen in the ELUCIDATE data. Allowing more observations per person would introduce bias as patients with slower progressing disease will have more ELF measurements prior to having a test result of 9.5 or above. The bias seen here relates to comments made by Bellera et al[135] when analysing monitoring data, with those patients contributing the most monitoring observations generally in a more stable disease state and potentially different to those contributing few monitoring observation points, see Chapter 7.

The model fitted to simulated data used 26,429 observation points for 9,608 simulated participants and the simulated data with adjusted fibrosis progression rate used 23,972 observations for 8,779 simulated participants. For the simulated data sets the mean number of observations was 2.8 for the original data set with unadjusted fibrosis progression and 2.7 for the data set with adjusted fibrosis progression. Results of modelling the ELUCIDATE data, simulated data and adjusted fibrosis progression estimate simulated data can be seen in Tables 8.8 and 8.9.

Modelling of the ELUCIDATE data estimated the increase in ELF per year was 0.31 (95% CI

Table 8.8: Results of multilevel model of repeated ELF measures from ELUCIDATE trial and monitoring simulation.

	Estimate	95% Confidence Interval	P-value
ELUCIDATE ELF			
Time since randomisation (years)	0.31	(0.22, 0.39)	< 0.001
Constant	8.73	(8.63, 8.82)	< 0.001
Between-individual SD	0.43	(0.36, 0.51)	
Within-individual SD	0.48	(0.44, 0.52)	
Simulated ELF			
Time since randomisation (years)	0.24	(0.23, 0.26)	< 0.001
Constant	8.84	(8.83, 8.85)	< 0.001
Between-individual SD	0.42	(0.41, 0.43)	
Within-individual SD	0.47	(0.46, 0.47)	
Simulated ELF with adjusted fibrosis progression			
Time since randomisation (years)	0.28	(0.27, 0.30)	< 0.001
Constant	8.86	(8.84, 8.87)	< 0.001
Between-individual SD	0.42	(0.41, 0.43)	
Within-individual SD	0.46	(0.46, 0.47)	

Table 8.9: Results of multilevel model (with random slope estimated) of repeated ELF measures from ELUCIDATE trial and monitoring simulation.

	Estimate	95% Confidence Interval	P-value
ELUCIDATE ELF			
Time since randomisation (years)	0.38	(0.28, 0.49)	< 0.001
Constant	8.71	(8.62, 8.80)	< 0.001
Between-individual SD	0.41	(0.34, 0.49)	
Within-individual SD	0.43	(0.38, 0.47)	
Random slope	0.34	(0.23, 0.51)	
Simulated ELF			
Time since randomisation (years)	0.32	(0.30, 0.34)	< 0.001
Constant	8.83	(8.82, 8.84)	< 0.001
Between-individual SD	0.39	(0.38, 0.40)	
Within-individual SD	0.42	(0.41, 0.42)	
Random slope	0.46	(0.44, 0.48)	
Simulated ELF with adjusted fibrosis progression			
Time since randomisation (years)	0.36	(0.35, 0.38)	< 0.001
Constant	8.84	(8.83, 8.86)	< 0.001
Between-individual SD	0.38	(0.37, 0.39)	
Within-individual SD	0.41	(0.41, 0.42)	
Random slope	0.46	(0.44, 0.48)	

(0.22, 0.39); p-value < 0.001). Modelling of the simulated data showed the increase in ELF per year to be comparable at 0.24 (95% CI (0.23, 0.26); p-value < 0.001) and for the simulated data with adjusted fibrosis progression the increase in ELF per year was 0.28 (95% CI (0.27, 0.30); p-value < 0.001). When allowing for the random slope the estimate of increase in ELF per year from the ELUCIDATE data was 0.38 (95% CI (0.28, 0.49); p-value< 0.001); from the model this was 0.32 (95% CI (0.30, 0.34); p-value< 0.001) and from the model with increased fibrosis progression this was 0.36 (95% CI (0.35, 0.38); p-value< 0.001).

Findings from the simulated data were broadly comparable with the trial. Comparison of the cirrhosis outcomes for the simulated data and ELUCIDATE data (monitoring arm) showed a higher percentage of positive results (70.7%/74.4% vs 64.2%) in the simulated data compared with the ELF arm of the trial. The percentage of positive results in the standard care arm was lower (4.5%). However, the percentage of diagnoses of cirrhosis after the first testing time point were greater for the trial than the simulated data (18.7%/18.3% vs 29.9%), see Table 8.10.

Table 8.10: Comparison of outcomes for trial and simulated data.

Outcome	**RCT monitoring**	**RCT standard care**	**Model Unadjusted Fibrosis progression**	**Model Adjusted Fibrosis progression**
Diagnosis of cirrhosis during trial	281/438 (64.2%)	20/440 (4.5%)	14132/20000 (70.7%)	14876/20000 (74.4%)
Diagnosis of cirrhosis after 1st measurement	84/438 (29.9%)	20/440 (4.5%)	3740/20000 (18.7%)	3655/20000 (18.3%)

8.5 Discussion

8.5.1 Reference strategy

At the initial testing point a monitoring strategy will be identifying cases from a prevalent population where a large proportion of patients will have high ELF values. At subsequent time points those with a previous positive result will not be tested and the tested population will contain cirrhotic patients that were falsely negative at the previous testing point or have developed cirrhosis since the last testing point (incident cases), hence the difference in

results at the initial monitoring time point compared with others. The percentage of false negative results generally increased with each time point as patients with low ELF trajectory have reached compensated cirrhosis but as they have a low ELF value for their disease stage they are required to progress further to have a positive test result using the simple threshold decision rule. The increasing percentage of false negative results as the testing points advance suggests the simple threshold should be reduced at later time points to account for the patients that have false negative results using the original threshold.

8.5.2 Comparing strategies with changes to individual components to the reference strategy

8.5.2.1 Inferior strategies

It was anticipated that the strategy with retesting would result in an increase in the number of tests per person required compared to the reference strategy (due to the increase in tests at each observation point) but the percentage of patients with delay to diagnosis increased also. Due to the measurement error of both the initial and retest results some patients would have been positive on their initial test (as with the reference strategy) but using the mean of the initial and retest means they have a negative result. The slight increase in time to diagnosis when using a retest strategy will have a small effect on the percentage of participants with delay to diagnosis also.

The strategies using absolute and relative increase from last recorded value decision rules were notably worse than the strategies using absolute and relative changes from the initial recorded value. When using a decision rule based on detecting a magnitude of change between one value and another, the two values used to calculate the change will both have measurement error. Comparisons with the initial value will consider increases in ELF across the entire monitoring period rather than increases since the previous monitoring point only. Differences from the initial value rather than the last value were better for detecting true change over measurement error (signal from noise).

The simple threshold strategy outperformed the strategies comparing current to previous values. This is linked to the index of individuality (II), the ratio of within-individual and between-individual variation. If a test has a high II value, where an individual can have results spanning a wide range of the possible results for a group of people, comparison to constant thresholds will be more meaningful than for tests with low II values, where an individual will have tests results spanning only part of the possible range of results and comparisons to previous results will be more beneficial.[35] Modelling results showed the measurement error and between-individual variability to be similar which would lead to a large II, see Table 8.8.

8.5.2.2 'Trade-off' strategies

The reduced test frequency strategy showed a large decrease in the number of tests per person used for a small increase in the percentage of people with delay to diagnosis. It may be that for a substantial decrease in the number of tests required, and therefore the resource used, the slight potential for increased harm to patients (through later diagnosis) is acceptable.

8.5.2.3 Superior strategies

The linear regression strategy was the only strategy tested that showed a reduction in both the number of tests required and the percentage of patients with delay to diagnosis. By fitting a regression model using all previous observations for an individual and obtaining a prediction from this, the linear regression method utilised all available data and some allowance was made for the fluctuation in results due to measurement error. The linear regression strategy, however, only resulted in small benefits compared with the reference strategy. This modest improvement in monitoring strategy performance may not merit the extra complexity involved when using the linear regression method.

8.5.3 Estimates of test performance and disease progression (sensitivity analyses)

The results obtained when varying estimates in the simulation model and evaluating the reference strategy highlight the importance of accurate data. The increases and decreases in estimates of test performance (measurement error and between-individual variability) and fibrosis progression rate affected the three measures of performance in different ways.

8.5.3.1 Measurement error and between-individual variability

The measurement error of a test affects the number of false positive test results, with larger measurement error resulting in more false positive results and smaller measurement error resulting in fewer false positive results. Between-individual variability will affect the underlying ELF values possible at each fibrosis stage. Providing ELF is related to fibrosis stage, if the between-individual variability is smaller it will be easier to correctly identify fibrosis stage from ELF resulting in fewer false positive results and more true positive results.

With fewer false positives and more true positives, PPV will increase, the number of tests required will increase as the reduction in false positives means the number of patients correctly staying in the monitoring programme will increase. With reduced measurement error the observed values reflect more closely the underlying disease state of each patient, if the threshold does not adequately account for this patients will need to progress for longer to have a test value over the threshold indicating a positive result. When the between-individual variability is reduced, due to the increase in true positive results the percentage of patients with delay to diagnosis will decrease.

8.5.3.2 Comparison of simulated and trial data

Analysis of the simulated and trial data showed similar results; indicating the model may reflect patient progression well. The monitoring arm of the trial detected cirrhosis in 64.2% of patients compared with just 4.5% in the standard care arm and the model predicted this

would be 70.7%. Such a difference in detection suggests the false positive rate is high with the strategy used in the trial. Had there been sufficient time this modelling exercise could have been used to modify the strategy.

8.5.3.3 Fibrosis progression rate

Fibrosis progression rate will affect the number of diseased patients. With an increased fibrosis progression rate more patients will have compensated cirrhosis, which will lead to an increase in PPV. With increased fibrosis progression rate patients have positive results earlier in the strategy and the strategy will require fewer tests to be performed. If patients have increased fibrosis progression rate more patients will have been in cirrhosis for more than 12 months meaning more patients can be undetected for over 12 months.

8.5.4 Comparison of modelled data to ELUCIDATE

The fibrosis progression rate, adjusted (increased) due to clinical opinion, seemed reasonable compared to the ELUCIDATE data. The estimates of between and within-individual standard deviations for the simulated data were similar to the ELUCIDATE data. The comparison of individuals detected as positive was higher in the simulated data, but the proportion of patients identified after the first testing point was underestimated in the simulated data, suggesting a decreased severity of disease at entry in the trial.

Overall, the simulated data compared well with the ELUCIDATE data, showing the possibilities of modelling to design a monitoring strategy prior to evaluating in a trial.

8.5.5 Limitations

8.5.5.1 Data sources

The estimates from data sources used to inform the simulation model will have a large impact on the results of the simulation model. The suitability of data was assessed, by consultation

with clinical colleagues, and where necessary estimates were adjusted for sensitivity analyses. However, as the model was dependent on the information it used, the quality and suitability of data used will always be a limitation. Just one cross-sectional study provided information on both the link between ELF values and fibrosis stage and the distribution of fibrosis stages at entry to the trial. When looking to identify an estimate of measurement error several sources were identified with the estimates from each found to be vastly different. The data used to obtain the estimate of measurement error did not allow within-individual and analytical variation to be estimated separately meaning an estimate of total imprecision was used and applied at each time point. The data linking ELF to fibrosis stage defined fibrosis stage by biopsy. Even though biopsy is the reference standard for staging fibrosis, biopsy is known to not be accurate in some cases.

The ELUCIDATE trial data used to assess the simulation model was not completely appropriate as the dataset contained repeated observations from 153 participants with many participants having only two observations; more observations per person would allow the model to better estimate the error terms and the changes over time. ELF measurements only being taken until the point of a measurement being classed as positive also hinders the ability of the data to estimate the true progression of ELF over time as those with higher ELF values (and possibly more developed cirrhosis) cease to have ELF recorded and so progression beyond this ELF value cannot be assessed. Patients with lower ELF measures (below 9.5) continued monitoring meaning they had more measures, and therefore contributed more data to the model; however they were potentially very different to those with fewer measures, who were likely in a worse health state.

8.5.5.2 Assumptions

A limitation of the simulation model is the number of assumptions required. Some of the estimates used to generate the monitoring data, such as fibrosis progression rate and measurement error, can be varied in magnitude and the results assessed to identify the impact of using data of insufficient quality or suitability in the model. However, there were assumptions in the development of the model that were not explored.

The model assumes fibrosis progression is constant and requires patients to have positive fibrosis progression. The model assumed linear increases in ELF between fibrosis stages, normally distributed ELF within fibrosis stage and constant fibrosis progression rate. The error associated with each observation was assumed to be normally distributed and a simple error term was used. The measurement error used in the simulation may be simplistic as it is randomly chosen from a distribution that is constant across individuals and time, and not linked to the magnitude of the ELF value. As no alternative data or substantiated opinion was available to enable modelling of these factors in any other way, these assumptions were necessary for the development of the model. Longitudinal data sets with ELF values and biopsy recorded in addition to data from a biological variability study of ELF would be required to test these assumptions.

8.5.5.3 Trial considerations

Several criteria were required to allow the simulation model to generate data for a trial (described in Table 8.2). Whilst these criteria were included to avoid anomalies and were based on clinical advice, there is no data to support them.

8.5.6 Further work

A greater variety of strategies could be evaluated with multiple components assessed simultaneously. More complex decision rules and frequencies could be explored, for example a simple threshold decision rule where the threshold remains the same across patients but varies by time point within a monitoring strategy or changing the frequency of testing to be non-constant.

It may be possible for the simulation model to be adapted to account for usual care (and variation in usual care). If usual care could be modelled, it may be possible to compare the use of monitoring strategies (in addition to usual care) to usual care alone and with further simulation work estimate differences in patient outcomes between the approaches.

The model can be used to show lifetime progression for a time-matched cohort of patients with fibrosis (if the data is simulated with all patients starting at the onset of liver fibrosis). These data may be beneficial to the assessment of how a strategy would perform in practice rather than specifically in the trial setting as this would provide information on how newly diagnosed patients would benefit from monitoring.

Well-designed studies of biological variability would mean accurate estimates of variability (see Chapters 3 and 5) could be used in the model enabling more accurate monitoring data to be generated and analysed.

8.6 Conclusions

Simulation can be used to obtain monitoring data for candidate monitoring strategies and to enable an appropriate strategy to be selected for full scale evaluation.

To generate monitoring data there has to be available evidence on the natural history of the disease and the performance of the monitoring test (measurement error and test accuracy)– this evidence can be from existing data sets, reviewing the literature or potentially expert opinion. If the data informing the simulation model is inaccurate the results obtained from evaluation of strategies will not reflect the truth. Inaccurate estimates will affect results in a complex way. The results of sensitivity analyses highlighted the importance of accurate estimates of test performance and progression. See Chapters 2 to 5 for reviews of methods and studies, analysis and studies of sample size for biological variability studies.

Comparison of the trial data and the simulated data provided similar results. Bias in monitoring data, particularly concerning the number of recorded results, should be considered when analysing as those contributing more monitoring data points are generally different to those contributing few (see Chapter 7).

Chapter 9

Discussion and Conclusions

Monitoring of disease progression and recurrence is frequently used by healthcare providers in the management of patients with little or no evidence of effectiveness.[2] The evidence required to know how and when a test can be used to monitor disease progression and recurrence is often poorly understood and neglected.[1]

The overarching questions addressed in this thesis were:

- How can optimal monitoring strategies be designed?
- What are the appropriate study designs and methods for estimating variability of tests?

To investigate estimating test variability the design, analysis and quality of reporting of studies was assessed to understand test behaviour in a monitoring setting. Results of these studies help researchers understand how tests can be best used to monitor patients, informing decision rules for identifying test positives. To investigate how to optimise monitoring strategies, work focussed on how monitoring strategies can be designed and evaluated prior to full scale

evaluation in a randomised controlled trial. The purpose of this work was to optimise all aspects of the monitoring strategy (the decision rule, threshold for a positive test, frequency of monitoring and duration of monitoring) considering both the performance of monitoring strategies and the impact on patient outcomes. The aim was to combine knowledge of disease progression and the ability of the monitoring test to optimise a monitoring strategy that could then be used in a randomised controlled trial.

9.1 Overview of thesis

9.1.1 Research questions

This thesis looked broadly at two areas: the design, analysis and reporting of biological variability studies, which provide estimates of measurement error, and the use of modelling techniques to combine evidence and allow comparison of monitoring strategies, so that optimal strategies can be used in further investigations.

How can optimal monitoring strategies be designed? What are the appropriate study designs and methods for estimating variability of tests? In order to deliver on these main questions, this thesis aimed to answer the following questions:

- What are the current methods for assessing biological variability?
- How well are biological variability studies designed, analysed and reported?
- Can the design and analysis of biological variability studies be improved, specifically sample size planning and outlier detection methods? Are the current methods for analysis of biological variability studies valid, considering sample size and outlier detection?
- What are the current methods for the design and analysis of monitoring strategies?
- Can modelling methods be used to predict the performance of monitoring strategies, to identify optimal strategies to be evaluated in an RCT?

The first Chapter provided an overview of the issues with non-evidence based monitoring strategies being commonly used in practice. This Chapter provided background detail on biological variability, the standard study design and associated study designs. The broad concepts for developing a monitoring strategy and full evaluation of a monitoring strategy were also introduced. The eGFR-C and the ELUCIDATE studies were introduced in Chapter 1 as they were case studies used in subsequent Chapters of the thesis. This Chapter defined the scope of the thesis and specified the main aims.

The second Chapter focussed on biological variability and introduced the key ideas to assessing biological variability. This Chapter considered the design, analysis and reporting of studies to evaluate biological variability using literature from a variety of areas, not just laboratory sciences.

The third Chapter was a review of the current state of biological variability studies. This review identified studies of biological variability across test areas (laboratory, physiological and imaging) and assessed the design, analysis and reporting of these studies.

The fourth Chapter showed analysis of a biological variability case study conducted as part of the eGFR-C study. The analysis of these data was performed in several ways to understand the impact of the various methods identified in Chapters 2 and 3 (transformation of data and outlier detection).

The fifth Chapter focussed on sample size for biological variability studies. Using simulation, biological variability data were generated and the standard methods were used to analyse these data. This simulation process allows researchers to understand the validity of results and the potential results they may generate from studies given their chosen sample size.

The sixth Chapter investigated the impact of outlier detection methods when analysing biological variability data. Again, simulated data were used to show the effect of using different methods to identify outliers and removing these prior to analysis. The effect of these methods, and the validity of results, was investigated and guidance provided.

The seventh Chapter reviewed the literature in the area and related areas of developing and

evaluating monitoring strategies. This review discussed modelling methods, signal and noise, screening, and health economic approaches.

The eighth Chapter developed a model allowing monitoring data to be simulated and monitoring strategies to be compared, with optimal strategies selected. This Chapter used the ELUCIDATE study as a case study. The model simulated trial data and this was compared with the actual trial data. This analysis allowed the performance of monitoring strategies to be compared.

9.1.2 What are the current methods for assessment of biological variability?

To understand the current methods used to assess the biological variability of tests (design, analysis and reporting) a review of the literature was undertook in Chapter 2, a review of studies of biological variability in Chapter 3 and a case study analysis using these methods in Chapter 4.

There was little identified literature regarding the design of biological variability studies in the review in Chapter 2. For laboratory based studies there appeared to be one main source of design advice, the Fraser-Harris framework published in 1989. This design guidance was evident in the studies of biological variability identified and reviewed in Chapter 3. The Fraser-Harris framework offered limited information regarding sample size but there have been recent publications commenting on this issue.

The review of the literature (Chapter 2) identified the methods for the analysis of biological variability studies. The terminology of results differed across disciplines and the area of clinical chemistry had the most well defined methods. The primary method of evaluating data was ANOVA or a random effects model. In the laboratory setting there was additional guidance regarding transforming data and outlier removal, again stemming from the Fraser-Harris framework. When biological variability data are log-transformed prior to analysis, methods can be used to provide 'exact' results (see Chapter 2). The identified literature demonstrated the estimates of variability were different across areas, with laboratories favouring the calcu-

lation of coefficients of variation and reference change values; whereas, in the medical setting intra-cluster correlation coefficients were commonly referred to. Methods for similar settings to biological variability include inter-intra reader analysis and Bland-Altman analysis (comparison of methods for measuring the same thing). Chapter 3 highlighted how biological variability analysis methods were used in the identified studies and Chapter 4 provided examples of the use of these analyses.

Chapter 2 identified literature calling for the standardisation of terminology and notation and a checklist for reporting. The infrequent use of confidence intervals to display uncertainty of estimates in laboratory studies of biological variability was also identified as a concern, as the meaning and interpretation of estimates may be affected by lack of clear reporting of the range of likely values for estimates of variability.

9.1.3 How well are biological variability studies designed, analysed and reported?

The main finding from the review of biological variability studies (Chapter 3) was the lack of these studies, particularly considering their importance in identifying the optimal use of a test in monitoring. The majority of the studies identified were for laboratory tests, rather than physiological and imaging tests, although this may be due to the search criteria used. There is a need to educate researchers to understand the key importance of these studies and the vital information they provide to tailor the use of tests.

The review (Chapter 3) identified the design of biological variability studies as suboptimal, often recruiting very few participants (particularly studies of laboratory tests) and rarely providing any justification for the sample size chosen. Participants in biological variability studies are often healthy participants and do not reflect the population that would receive the test to monitor progressive or recurrent disease; estimates of variability from studies of healthy populations may be very different to disease populations. Some studies of biological variability did not assess analytical, within-individual and between-individual variability in the same study and instead relied on external estimates; this practice is not appropriate and

has the potential to result in incorrect estimates of variability.

The review of biological variability studies (Chapter 3) showed the analysis was in most cases appropriately performed by ANOVA or random effects modelling. In the studies of biological variability of laboratory tests additional methods were used prior to this analysis to assess the normality of data and identify outliers, leading to transformation and deletion of outliers. The process of assessing normality and transforming data may be performed to ensure the methods of analysis are appropriate but may also be due to the beneficial impact on analysis when log-transformed biological variability data are used (coefficients of variation can easily be derived), with exact estimates produced for log-normal data.

The purpose of outlier detection and removal methods is less obvious and more concerning. The purpose of these studies is to evaluate variability, with the removal of variability prior to analysis potentially introducing bias (removing outliers generally removes variability). Chapter 4 further investigated the impact of transformation and outlier detection using a case study, showing the results vary when these methods are employed. The use of log-transformation needs to be complemented with the calculation of 'exact' results and be clearly specified when reporting methods and results. The analysis with and without outlier exclusion showed the impact of outlier removal often led to a reduction in the estimated variability.

The review of studies of biological variability identified reporting was poor often due to lack of detail regarding: sample size rationale, number of measurements, timing of repeats and the method of analysis. Confidence intervals were rarely used to indicate the uncertainty of estimates. It was also likely that outlier detection and deletion and normality assessment and transformation may have been performed in some studies but the methods were not explicitly stated.

9.1.4 Can the design and analysis of biological variability studies be improved, specifically sample size planning and outlier detection methods? Are the current methods for analysis of biological variability studies valid, considering sample size planning and outlier detection methods?

The review of biological variability studies (Chapter 3) highlighted the need for improvements in areas of design, analysis and reporting. Subsequent work in Chapter 5 was developed to provide guidance to researchers when planning biological variability studies by investigating the impact of sample size (number of participants, observations of participants, and assessments of observations of participants) on the precision of results. Chapter 6 focussed on understanding the impact of different methods for outlier detection on the standard analysis of biological variability data and the results obtained. The empirical analysis of a biological variability study conducted in Chapter 4 further uncovered the issues with analysis and how this should be reported.

Chapter 3 identified issues with the design of biological variability studies. One of the main issues identified was sample size and this was further investigated in Chapter 5. A simulation model was developed allowing biological variability data to be simulated and evaluated, and using multiple simulations the bias and precision of estimates of biological variability were assessed. The results showed estimates may not be valid in some test scenarios with small sample sizes, due to the bias of variability estimates. The results identified where to focus resource to gain precision. Increasing the number of participants appeared to be most beneficial as this increased the precision of analytical, within-individual and between-individual variability which impacts on the precision of all measures of coefficient of variation, the reference change value and the index of individuality. An application was developed to carry out the simulation and guide researchers, `https://alicesitch.shinyapps.io/bvs_simulation/`.

Chapter 6 investigated one of the concerns regarding analysis of biological variability data, which was the use of outlier detection methods and deletion of the measures identified. The results of this simulation study showed, with log-normally distributed data, outlier detection

methods generally led to underestimation of the variability. Outlier detection methods involving the Cochran C test and Tukey IQR rule were particularly poor. With outlying data present the best performing outlier detection strategies were identified.

The review work in Chapter 3 and empirical analysis in Chapter 4 concluded and recommended reporting of biological variability studies needs to be transparent. The Bartlett Checklist has been developed which should improve reporting, and particularly enabling studies of biological variability in laboratory medicine to be identified. This checklist could be further enhanced to include the issues identified the methods for: checking normality of data, transformation leading to the methods used to produce results, as well as whether any methods have been used to identify and remove outliers, and any potential bias resulting from using these methods. When outliers are detected and removed this should be fully reported and it would be good practice for this analysis to be presented as a sensitivity analysis alongside the analysis of the full data set.

The review identified confidence intervals were rarely given for estimates of biological variability; Chapter 2 provides the formulas for these confidence intervals and an application was developed alongside the simulation work in Chapter 5 allowing researchers to easily calculate and visualise these for a range of sample sizes (`https://alicesitch.shinyapps.io/bvs_cis`). Reporting of confidence intervals indicates the uncertainty of estimates and allows results from separate studies to be compared.

9.1.5 What are the current methods for the design and analysis of monitoring strategies?

The review of monitoring and monitoring related literature (Chapter 7) identified a paucity of research, with the few relevant studies reporting an analysis or simulation of a particular monitoring situation; most studies simply offered a standard method for designing and analysing a monitoring strategy. The review identified a general model of monitoring data, with the observed values comprising of the true (unobserved) value and the measurement error.

In the published literature on screening tests, methods focussed on identifying the optimal frequency of screening. The review identified thresholds are often developed using variability data, especially in the area of treatment titration. Whilst variability information is necessary when developing a monitoring strategy, it is important to include considerations such as impact on patient health for monitoring progressive and recurrent disease. The literature review also highlighted the potential biases in monitoring data.

9.1.6 Can modelling methods be used to predict the performance of monitoring strategies, to identify optimal strategies to be evaluated in an RCT?

Chapter 8 introduced a model for monitoring data allowing observed test data and 'true' underlying and unobserved data to be generated. Monitoring data were generated using knowledge of disease progression, accuracy of the monitoring test and variability of the monitoring test. Monitoring strategies were then evaluated using the observed data to give a test outcome and the underlying 'true' value to indicate the true disease state. Monitoring strategies were defined specifying the decision rule, test threshold, frequency and duration of monitoring. Decision rules using a simple threshold for all participants, relative and absolute increase from last and first measure for each individual, including a retest component and using predictions from a linear regression model were compared. For the ELUCIDATE case study, the best performing strategy was the linear regression strategy with similar results for the simple threshold strategy. Comparing the simulated data with the ELUCIDATE data showed the results were similar and the simulation model may have been able to identify a preferable monitoring strategy for the trial, with a higher threshold used to reduce false positive results.

9.2 Strengths and Limitations

9.2.1 Strengths

The work presented in this thesis looks at an under researched area where non evidence based practice is a common occurrence. The importance of understanding biological variability and the impact of this information on how tests are used for monitoring is under-acknowledged, thus under-utilised.

Biological variability studies are often small niche projects in laboratories. This thesis offers a coherent statistical review of methods which will help researchers in this field understand why they are using these methods and challenge the specific ways these methods are employed. Guidance and understanding will improve the design, analysis and reporting of biological variability studies.

The sample size simulation work allows researchers to plan biological variability studies according to the precision of all estimates of biological variability, adding to the previous knowledge in this area. Also the application for calculating confidence intervals may help with the reporting of estimates. Guidance on the issues associated with outlier detection will help understanding of the potential problems when using these methods and enable researchers to use these approaches when necessary and with appropriate caution.

The further simulation work focusses on the whole monitoring strategy and evaluates monitoring test performance. This approach allows all aspects of a monitoring strategy to be evaluated rather than each component in isolation. Optimal strategies can then be selected for further evaluation.

9.2.2 Limitations

Chapters 2, 3 and 7 provide reviews of the literature for assessing biological variability, biological variability studies and monitoring and monitoring-related methodology respectively. These reviews were not systematic reviews. The purpose of the literature reviews (Chapters 2

and 7) was to understand the main methods available and a systematic review would not have been an efficient way to locate this information. The purpose of Chapter 3 was to understand the current state of the field of biological variability studies. These studies were difficult to identify, especially studies of physiological and imaging tests. It may be laboratory studies are over represented in the review of biological variability studies as these were the studies easily located due to the recorded and updated Westgard QC database. A systematic review may have identified more biological variability studies to enhance this review.

Chapter 4 used a single case study for empirical analysis. This case study had multiple outcomes but all of these were related to kidney function. This case study also focussed on laboratory tests and analyses of physiological and imaging tests were not performed. The sample size for the eGFR-C study was small with only 20 participants, four observations of each participant and analysis in duplicate; however, this is reflective of these studies in practice.

Chapters 5 and 6 presented simulation studies to understand the precision of estimates of results using simulated data which are normally or log-normally distributed. These distributions were used so the estimated result from the simulation could be compared with the expected result. However, it is unlikely real biological variability data will perfectly follow these distributions. Use of the sample size application requires researchers to input estimates for each level of variability, which may not be known.

Chapter 8 produced a model which was dependent on many data sources and assumptions. Data required for this model included natural progression of disease (case mix at study entry and an estimate of the rate of progression), the performance of the test (accuracy, specifically the link between biomarker and disease stage) and the measurement error of the test (biological variability estimates). Limited data sources were available to contribute these data. It is unlikely such an array of data would be available in many clinical areas for many tests. The model assumed linear progression between stages of disease, which may not be fully appropriate.

9.3 Implications for practice

The work presented in this thesis emphasises the importance of estimating biological variability and the potential studies of biological variability have to impact on development of evidence based approaches to monitoring disease progression and recurrence. Highlighting the importance of these studies will not just increase the quality of the design, analysis and reporting but also the number of these studies conducted and the variety of settings these studies are conducted in. Researchers need to understand the importance of having good estimates of biological variability particularly for monitoring purposes, as do funders, as this is an area that can be vastly improved for limited resource. The results of biological variability studies may also show a test is not fit for monitoring prior to full scale evaluation limiting further cost.

This thesis allows better planning of biological variability studies. The issues identified in the review of biological variability studies and the empirical analyses need to be communicated to researchers in this area. The application developed to help researchers plan the sample size of biological variability studies considering the precision of estimates gives usable guidance which has the potential to greatly improve these studies. The application for confidence intervals for some biological variability estimates provides a tool for improving the reporting of uncertainty for these studies.

This thesis also provides recommendations to improve the analysis of biological variability studies with researchers empowered by further understanding of the methods and the issues related to outlier detection. The work on outlier detection also highlights the need for caution when interpreting results from studies that have used outlier detection methods.

The methods for planning and evaluating monitoring strategies prior to a trial, in terms of performance, need refining and tailoring to the specific situations they are applied to. However, this work shows the data required to undertake such an evaluation prior to a trial, the small preliminary studies that should be performed when developing a strategy and the possibilities if this information is available.

The work presented in this thesis has provided a cohesive and complete pathway for developing an evidence-based monitoring strategy to be fully evaluated in a monitoring RCT. Figure 9.1 demonstrates the necessary evidence required to design a monitoring strategy; from proof of principle, biological variability and accuracy studies, and using results from these studies with natural history data to model strategies allowing selection of optimal strategies and further evaluation using a monitoring RCT.

As monitoring RCTs are typically very expensive, following this pathway of studies can identify monitoring tests that are not fit for purpose (due to lack of proof of principle, high variability, poor accuracy or substandard monitoring strategy performance) early in the process using small-scale studies and modelling. Using this pathway also ensures only optimal tests are evaluated using monitoring RCTs.

9.4 Future research

Searches for biological variability studies indicated the need for clear labelling of these studies and the development of search terms.

A checklist for planning studies of biological variability would be beneficial. This would highlight the importance of sample size, choosing an appropriate population, specifying the measures of variability to be assessed and pre-specifying the methods for analysis. The Bartlett Checklist requires updating to reflect the additional points identified for transparent reporting; namely the methods for outlier detection and removal. Reporting of exactly how this has been conducted, the number of observations removed and the recommendation the analysis of the complete data set be reported also need to be included. A risk of bias tool could be developed to assess these studies also.

The impact of using of healthy populations to measure variability and using external sources to obtain some measures of variability require further investigation. These issues stem from a lack of understanding and indicated researchers in this area require further guidance when planning biological variability studies.

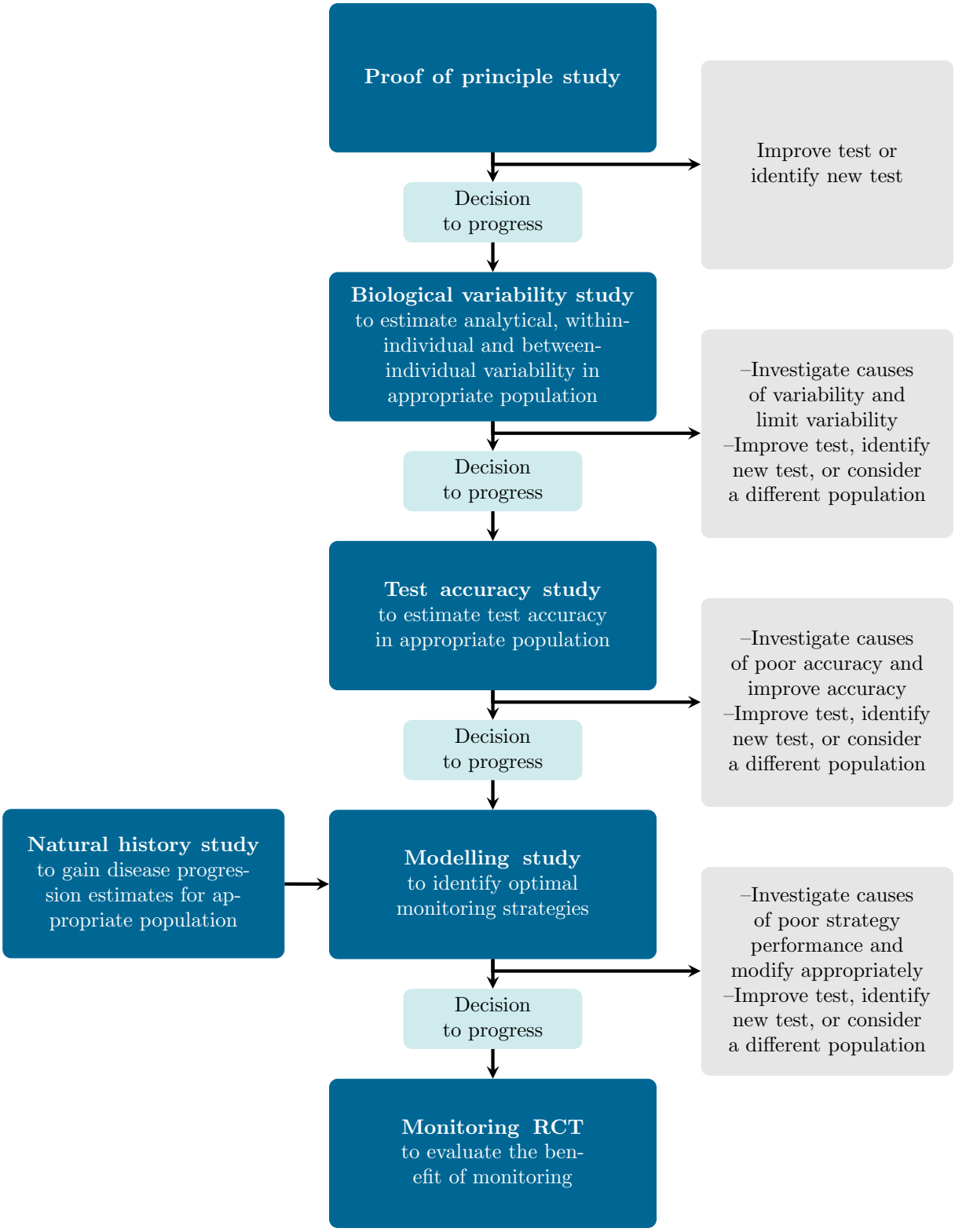


Figure 9.1: Pathway of designing monitoring studies.

The work concerning biological variability studies could be developed to further understand the impact of outlier detection and removal. This may be through additional empirical analyses to understand the impact of this practice or via a simulation study, using data that do not perfectly follow a distribution and is closer to real data.

The monitoring simulation work needs to be developed further to provide a generic case which can be modified by researchers to develop an idea of the performance of a strategy. The simulations performed can then have a greater degree of flexibility of strategies (changing more than one component at a time).

When the final results of the ELUCIDATE study (long-term follow up) are obtained analysis can be undertaken to further compare the simulation model with the ELUCIDATE data.

The monitoring simulation models can be further developed to consider the potential patient outcomes and cost-effectiveness of monitoring strategies. Also, given the need for data from a variety of sources, value of information analysis may be useful to understand where to focus resource.

9.5 Conclusions

This thesis has identified key information and methods that can be used to develop monitoring strategies allowing monitoring of disease progression and recurrence to be evidence-based.

In addition to the biological variability of the monitoring test, other key information for evaluating the use of monitoring of progressive or recurrence disease is the natural history of the disease and the accuracy of the monitoring test. Incorporating all this knowledge and using modelling techniques allows observed monitoring data and the underlying disease state to be simulated. These data can then be used to compare monitoring strategies in terms of strategy performance. This approach was validated by comparing the modelled data to observed study data.

This thesis has highlighted the importance of biological variability studies and the estimates obtained from them. Work needs to be done to ensure these studies are performed more often and in a variety of clinical areas, and researchers need to understand how to design these studies using the correct population, using the correct methods for analysis, and reporting the methods and results in a transparent manner. Tools have been developed allowing researchers to easily calculate confidence intervals for variance estimates and to plan studies with appropriate sample sizes to understand the likely precision of key estimates. Recommendations have also been made regarding the use of outlier detection methods when analysing biological variability studies and the interpretation from studies that have used these methods.

Using the pathway identified in this thesis, poor monitoring tests (for example, the ELF test to monitor progression of liver fibrosis and prostate specific antigen for monitoring recurrence of prostate cancer) can be identified early using small-scale studies (proof of principle, biological variability studies, test accuracy and modelling). Use of this pathway also ensures monitoring strategies are optimised prior to full evaluation in a monitoring RCT.

Appendix A

Biological variability studies: review of design, analysis, and reporting

A.1 Studies identified for review of biological variability studies

Table A.1: Studies identified for review of biological variability studies.

ID	Search	Authors	Year	Title	Journal
1	A	Curtis et al[184]	2014	Evaluation of dried blood spots with a multiplex assay for measuring recent HIV-1 infection	PLoS One
2	A	Gabriele et al[185]	2014	Reproducibility of the Carpet View system: a novel technical solution for display and off line analysis of OCT images	Int J Cardiovasc Imaging
3	A	Jimmerson et al[87]	2014	Development and validation of a dried blood spot assay for the quantification of ribavirin using liquid chromatography coupled to mass spectrometry	J Chromatogr B Analyt Technol Biomed Life Sci
4	A	Manley et al[186]	2014	Comparison of IFCC-calibrated HbA(1c) from laboratory and point of care testing systems	Diabetes Res Clin Pract
5	A & B1	Saez-Benito Godino et al[187]	2014	Multicentre evaluation of glycated haemoglobin (HbA1c) of Roche Diagnostics in Andalusia	Clin Biochem
6	B1	Wu et al[188]	2014	Biological variation of the osmolality and the osmolal gap	Clin Biochem
7	B2	Alizai et al[189]	2014	Cartilage lesion score: comparison of a quantitative assessment score with established semiquantitative MR scoring systems	Radiology
8	B2	Donati et al[86]	2014	Diffusion-weighted MR imaging of upper abdominal organs: field strength and intervendor variability of apparent diffusion coefficients	Radiology
9	B2	Frings et al[190]	2014	Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study	Radiology
10	B2	Giles et al[191]	2014	Whole-body diffusion-weighted MR imaging for assessment of treatment response in myeloma	Radiology
11	B2	Knobloch et al[192]	2014	Arterial, venous, and cerebrospinal fluid flow: simultaneous assessment with Bayesian multipoint velocity-encoded MR imaging	Radiology

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
12	B2	Roujol et al[193]	2014	Accuracy, precision, and reproducibility of four T1 mapping sequences: a head-to-head comparison of MOLLI, ShMOLLI, SASHA, and SAPPHIRE	Radiology
13	B2	Suh et al[194]	2014	Atypical imaging features of primary central nervous system lymphoma that mimics glioblastoma: utility of intravoxel incoherent motion MR imaging	Radiology
14	B2	Thevenot et al[195]	2014	Assessment of risk of femoral neck fracture with radiographic texture parameters: a retrospective study	Radiology
15	B3	Aakre et al[85]	2014	Weekly and 90-minute biological variations in cardiac troponin T and cardiac troponin I in hemodialysis patients and healthy controls	Clin Chem
16	B3	Bailey et al[82]	2014	Pediatric within-day biological variation and quality specifications for 38 biochemical markers in the CALIPER cohort	Clin Chem
17	B3	Karon et al[196]	2014	Precision and reliability of 5 platelet function tests in healthy volunteers and donors on daily antiplatelet agent therapy	Clin Chem
18	B3	Noceti et al[197]	2014	Tacrolimus pharmacodynamics and pharmacogenetics along the calcineurin pathway in human lymphocytes	Clin Chem
19	B3	Simpson et al[90]	2014	Use of observed within-person variation of cardiac troponin in emergency department patients for determination of biological variation and percentage and absolute reference change values	Clin Chem
20	C1	Beco et al[198]	1998	Study of the female urethra's submucous vascular plexus by color Doppler	World Journal of Urology
21	C1	Chen et al[83]	2011	The assessment of voluntary pelvic floor muscle contraction by three-dimensional transperineal ultrasonography	Archives of Gynecology & Obstetrics
22	C1	Heit[199]	2002	Intraurethral sonography and the test-retest reliability of urethral sphincter measurements in women	Journal of Clinical Ultrasound
23	C1	Oelke et al[200]	2009	Manual versus automatic bladder wall thickness measurements: a method comparison study	World Journal of Urology

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
24	C1	Oliveira et al[84]	2007	Ultrasonographic and Doppler velocimetric evaluation of the levator ani muscle in premenopausal women with and without urinary stress incontinence	European Journal of Obstetrics, Gynecology, & Reproductive Biology
25	C1	Otcenasek et al[201]	2002	New approach to the urogynecological ultrasound examination	European Journal of Obstetrics, Gynecology, & Reproductive Biology
26	C2	Naresh et al[202]	2013	Day-to-day variability in spot urine albumin-creatinine ratio	American Journal of Kidney Diseases
27	C2	Ristiniemi et al[203]	2012	Evaluation of a new immunoassay for cystatin C, based on a double monoclonal principle, in men with normal and impaired renal function	Nephrology Dialysis Transplantation
28	C2	Rule et al[204]	2013	Estimating the glomerular filtration rate from serum creatinine is better than from cystatin C for evaluating risk factors associated with chronic kidney disease.	Kidney International
29	C2	Sjostrom et al[205]	2009	Cystatin C as a filtration marker–haemodialysis patients expose its strengths and limitations	Scandinavian Journal of Clinical & Laboratory Investigation
30	C2	Walser et al[206]	1993	Prediction of glomerular filtration rate from serum creatinine concentration in advanced chronic renal failure	Kidney International
31	C3	Beeh et al[207]	2003	Long-term repeatability of induced sputum cells and inflammatory markers in stable, moderately severe COPD	Chest
32	C3	Herpel et al[89]	2006	Variability of spirometry in chronic obstructive pulmonary disease: results from two clinical trials	American Journal of Respiratory & Critical Care Medicine

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
33	C3	Liistro et al[208]	2006	Technical and functional assessment of 10 office spirometers: A multicenter comparative study	Chest
34	C3	Madsen et al[209]	1996	Patient-administered sequential spirometry in healthy volunteers and patients with alpha 1-antitrypsin deficiency	Respiratory Medicine
35	C3	McCarley et al[210]	2007	A pilot home study of temporal variations of symptoms in chronic obstructive lung disease	Biological Research for Nursing
36	C3	Timmins et al[211]	2013	Day-to-day variability of oscillatory impedance and spirometry in asthma and COPD	Respiratory Physiology & Neurobiology
37	D	Alexander et al[58]	2013	Prognostic utility of biochemical markers of cardiovascular risk: impact of biological variability	Clin Chem Lab Med
38	D	Alvarez et al[212]	2000	Components of biological variation of biochemical markers of bone turnover in Paget's bone disease	Bone
39	D	Alvarez et al[91]	2003	Biological variation of seminal parameters in healthy subjects	Human Reproduction
40	D	Andersen et al[213]	2010	Comparison of within- and between-subject variation of serum cystatin C and serum creatinine in children aged 2-13 years	Scand J Clin Lab Invest
41	D	Andersson et al[214]	2003	Variation in levels of serum inhibin B, testosterone, estradiol, luteinizing hormone, follicle-stimulating hormone, and sex hormone-binding globulin in monthly samples from healthy men during a 17-month period: possible effects of seasons	J Clin Endocrinol Metab
42	D	Ankrah-Tetteh et al[215]	2008	Intraindividual variation in serum thyroid hormones, parathyroid hormone and insulin-like growth factor-1	Ann Clin Biochem
43	D	Braga et al[216]	2011	Revaluation of biological variation of glycated hemoglobin (HbA(1c)) using an accurately designed protocol and an assay traceable to the IFCC reference system	Clin Chim Acta

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
44	D	Brown et al[217]	2008	Assay validation and biological variation of serum receptor for advanced glycation end-products	Ann Clin Biochem
45	D	Browne et al[218]	2007	Accuracy and biological variation of human serum paraoxonase 1 activity and polymorphism (Q192R) by kinetic enzyme assay	Clin Chem
46	D	Carlsen et al[95]	2011	Within-subject biological variation of glucose and HbA(1c) in healthy persons and in type 1 diabetes patients	Clin Chem Lab Med
47	D	Cembrowski et al[94]	2010	The use of serial patient blood gas, electrolyte and glucose results to derive biologic variation: a new tool to assess the acceptability of intensive care unit testing	Clin Chem Lab Med
48	D	Cheuvront et al[219]	2010	Biological variation and diagnostic accuracy of dehydration assessment markers	Am J Clin Nutr
49	D	Cho et al[220]	2005	The biological variation of C-reactive protein in polycystic ovarian syndrome	Clin Chem
50	D	Corte et al[221]	2010	Biological variation of free plasma amino acids in healthy individuals	Clin Chem Lab Med
51	D	Dednam et al[93]	2008	Biological variation of myeloperoxidase	Clin Chem
52	D	Desmeules et al[222]	2010	Biological variation of glycated haemoglobin in a paediatric population and its application to calculation of significant change between results	Ann Clin Biochem
53	D	Dittadi et al[223]	2004	Biological variation of plasma chromogranin A	Clin Chemistry Lab Med
54	D	Dittadi et al[224]	2008	Biological variability evaluation and comparison of three different methods for C-peptide measurement	Clin Chem Lab Med
55	D	Dittadi et al[225]	2008	Within-subject biological variation in disease: the case of tumour markers	Ann Clin Biochem
56	D	Frankenstein et al[62]	2011	Biological variation and reference change value of high-sensitivity troponin T in healthy individuals during short and intermediate follow-up periods	Clin Chem
57	D	Garde et al[226]	2000	Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A(1c), IgA, prolactin, and free testosterone in healthy women	Clin Chem

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
58	D	González et al[227]	2001	Biological Variation of Interleukin-1β, Interleukin-8 and Tumor Necrosis Factor-α in Serum of Healthy Individuals	Clin Chemistry Lab Med
59	D	Jensen et al[228]	2007	Biological variation of thyroid autoantibodies and thyroglobulin	Clin Chem Lab Med
60	D	Kristoffersen et al[229]	2012	A model for calculating the within-subject biological variation and likelihood ratios for analytes with a time-dependent change in concentrations; exemplified with the use of D-dimer in suspected venous thromboembolism in healthy pregnant women	Ann Clin Biochem
61	D	Lara-Riegos et al[230]	2013	Short-term estimation and application of biological variation of small dense low-density lipoproteins in healthy individuals	Clin Chem Lab Med
62	D	Martinez-Morillo et al[88]	2012	Reference intervals and biological variation for kallikrein 6: influence of age and renal failure	Clin Chem Lab Med
63	D	McKinley et al[231]	2001	Plasma homocysteine is not subject to seasonal variation	Clin Chem
64	D	Melzi d'Eril Aet al[232]	2001	Biological variation of serum amyloid A in healthy subjects	Clin Chem
65	D	Melzi d'Eril et al[233]	2003	Biological variation of N-terminal pro-brain natriuretic peptide in healthy individuals	Clin Chem
66	D	Meo et al[234]	2005	Biological variation of vascular endothelial growth factor	Clin Chem Lab Med
67	D	Moller et al[235]	2003	Biological variation of soluble CD163	Scand J Clin Lab Invest
68	D	Mosca et al[236]	2013	Analytical goals for the determination of HbA(2)	Clin Chem Lab Med
69	D	Nguyen et al[237]	2008	Within-subject variability and analytic imprecision of insulinlike growth factor axis and collagen markers: implications for clinical diagnosis and doping tests	Clin Chem
70	D	Pagani et al[238]	2001	Biological variation in serum activities of three hepatic enzymes	Clin Chem
71	D	Pineda-Tenor et al[239]	2013	Biological variation and reference change values of common clinical chemistry and haematologic laboratory analytes in the elderly population	Clin Chem Lab Med

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
72	D	Reclos et al[240]	2006	Estimation of the biological variation of glucose-6-phosphate dehydrogenase in dried blood spots	Accreditation and Quality Assurance
73	D	Reinhard et al[241]	2009	Biological variation of cystatin C and creatinine	Scand J Clin Lab Invest
74	D	Rohlfing et al[242]	2002	Biological variation of glycohemoglobin	Clin Chem
75	D	Rossi et al[243]	2013	High biological variation of serum hyaluronic acid and Hepascore, a biochemical marker model for the prediction of liver fibrosis	Clin Chem Lab Med
76	D	Serteser et al[244]	2012	Biological variation in pregnancy-associated plasma protein-A in healthy men and non-pregnant healthy women	Clin Chem Lab Med
77	D	Shand et al[245]	2006	Biovariability of plasma adiponectin	Clin Chem Lab Med
78	D	Talwar et al[246]	2005	Biological variation of vitamins in blood of healthy individuals	Clin Chem
79	D	Trapé et al[247]	2000	Reference Change Value for HbA1c in Patients with Type 2 Diabetes Mellitus	Clin Chemistry Lab Med
80	D	Trapé et al[248]	2003	Reference change value for alpha-fetoprotein and its application in early detection of hepatocellular carcinoma in patients with hepatic disease	Clin Chem
81	D	Trapé et al[249]	2005	Biological variation of tumor markers and its application in the detection of disease progression in patients with non-small cell lung cancer	Clin Chem
82	D	Trapé et al[250]	2010	Determination of biological variation of alpha-fetoprotein and choriogonadotropin (beta chain) in disease-free patients with testicular cancer	Clin Chem Lab Med
83	D	Valero-Politi et al[251]	2001	Annual Rhythmic and Non-Rhythmic Biological Variation of Magnesium and Ionized Calcium Concentrations	Clin Chemistry Lab Med
84	D	van der Merwe et al[252]	2002	Biological variation in sweat sodium chloride conductivity	Ann Clin Biochem
85	D	van Hoydonck et al[253]	2003	Reproducibility of blood markers of oxidative status and endothelial function in healthy individuals	Clin Chem

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
86	D	Vasile et al[254]	2010	Biological and analytical variability of a novel high-sensitivity cardiac troponin T assay	Clin Chem
87	D	Vasile et al[255]	2011	Biologic variation of a novel cardiac troponin I assay	Clin Chem
88	D	Viljoen et al[256]	2008	Analytical quality goals for parathyroid hormone based on biological variation	Clin Chem Lab Med
89	D	Wu et al[257]	2009	Short- and long-term biological variation in cardiac troponin I measured with a high-sensitivity assay: implications for clinical practice	Clin Chem
90	D	Wu et al[258]	2012	Long-term biological variation in cardiac troponin I	Clin Biochem
91	D & E	Bandaranayake et al[92]	2007	Intra-individual variation in creatinine and cystatin C	Clin Chem Lab Med
92	D & E	Delanaye et al[259]	2008	New data on the intraindividual variation of cystatin C	Nephron Clin Pract
93	D & E	Toffaletti et al[260]	2008	Variation of serum creatinine, cystatin C, and creatinine clearance tests in persons with normal renal function	Clin Chim Acta
94	E	Gaspari et al[261]	1998	Precision of plasma clearance of iohexol for estimation of GFR in patients with renal disease	J Am Soc Nephrol
95	E	Gowans et al[262]	1988	Biological Variation of Serum and Urine Creatinine and Creatinine Clearance: Ramifications for Interpretation of Results and Patient Care	Annals of Clinical Biochemistry: An international journal of biochemistry and laboratory medicine
96	E	Keevil et al[263]	1998	Biological variation of cystatin C: implications for the assessment of glomerular filtration rate	Clin Chem
97	E	Khullar et al[24]	1994	A novel technique for measuring bladder wall thickness in women using transvaginal ultrasound	Ultrasound Obstet Gynecol
98	E	Kuo[264]	2009	Measurement of detrusor wall thickness in women with overactive bladder by transvaginal and transabdominal sonography	Int Urogynecol J Pelvic Floor Dysfunct

See §3.3.1 for detail of searches.

ID	Search	Authors	Year	Title	Journal
99	E	Lekskulchai et al[265]	2008	Detrusor wall thickness as a test for detrusor overactivity in women	Ultrasound Obstet Gynecol
100	E	Panayi et al[266]	2010	Ultrasound measurement of vaginal wall thickness: a novel and reliable technique	Int Urogynecol J
101	E	Tubaro et al[267]	2013	Intra- and inter-reader variability of transvaginal ultrasound bladder wall thickness measurements: results from the shrink study	Neurology and Urodynamics

See §3.3.1 for detail of searches.

Table A.2: Details of studies identified for review of biological variability studies.

Study details				Study design			Analysis	Reporting					
ID	**Test type**	**Measure**	**Tests (n)**	**Situations (n)**	**Participants**	**Participants (n)**	**Analysis**	CV_A	CV_I	CV_G	**RCV**	**II**	**ICC**
1	Laboratory	Antibody reactivity (blood spot assay HIV)	6	6	Non-healthy participants	51	ANOVA/RE (assumed)	Yes	No	No	No	No	No
2	Imaging	Stents (angina)	2	6	Non-healthy participants	21	Other	No	No	No	No	No	Yes
3	Laboratory	Ribavirin (hepatitis C)	1	1	Unknown	4	ANOVA/RE (assumed)	Yes	No	No	No	No	No
4	Laboratory	HbA1c	3	3	Non-healthy participants	23	Other	No	No	No	No	No	No
5	Laboratory	HbA1c	1	4	Non-healthy participants	35	ANOVA/RE (assumed)	Yes	No	No	No	No	No
6	Laboratory	Osmolal gap, sodium, glucose	6	6	Healthy participants	20	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	Yes	No
7	Imaging	Cartilage legion score	3	3	Mixed participants	77	ANOVA/RE (assumed)	No	No	No	No	No	Yes
8	Imaging	Abdominal diffusion	6	6	Healthy participants	10	ANOVA/RE	Yes	No	No	No	No	Yes
9	Imaging	Tumour size	19	19	Non-healthy participants	34	ANOVA/RE	No	No	No	No	No	Yes

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
10	Imaging	Myeloma treatment response	1	1	Mixed participants	15	ANOVA/RE (assumed)	No	Yes	No	No	No	No
11	Imaging	Blood and CSF flow	1	1	Healthy participants	10	ANOVA/RE (assumed)	No	No	No	No	No	Yes
12	Imaging	Extra cellular volume fraction	4	8	Healthy participants	7	ANOVA/RE	No	No	No	No	No	No
13	Imaging	Tumour parameters and blood volume	4	8	Non-healthy participants	60	ANOVA/RE (assumed)	No	No	No	No	No	Yes
14	Imaging	Bone texture	4	4	Unknown	53	ANOVA/RE (assumed)	No	No	No	No	No	Yes
15	Laboratory	Cardiac troponin	2	8	Mixed participants	39	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
16	Laboratory	Biochemical markers	38	53	Healthy participants	29	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
17	Laboratory	Platelet function	5	10	Mixed participants	53	ANOVA/RE	Yes	Yes	No	No	No	Yes
18	Laboratory	Lymphocytes	4	4	Healthy participants	5	ANOVA/RE (assumed)	Yes	Yes	No	No	Yes	No
19	Laboratory	Cardiac troponin	1	1	Non-healthy participants	283	ANOVA/RE	No	Yes	Yes	Yes	Yes	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
20	Imaging	Length and thicknesss of sheath and distance	3	3	Mixed participants	27	ANOVA/RE (assumed)	No	No	No	No	No	No
21	Imaging	US pelvic floor	8	8	Mixed participants	20	ANOVA/RE	No	No	No	No	No	Yes
22	Imaging	US urethral spincter measurements	8	8	Healthy participants	29	ANOVA/RE	No	No	No	No	No	Yes
23	Imaging	US bladder wall thickness	4	4	Non-healthy participants	50	Other	No	No	No	No	No	No
24	Imaging	US and Doppler area measures	8	8	Mixed participants	63	Other	No	No	No	No	No	No
25	Imaging	US	6	6	Non-healthy participants	10	Other	No	No	No	No	No	No
26	Laboratory	Albumin creatinine ratio	1	1	Non-healthy participants	157	Other	No	No	No	No	No	No
27	Laboratory	Cystatin C	1	1	Healthy participants	170	ANOVA/RE (assumed)	No	Yes	No	No	No	No
28	Laboratory	eGFR	5	5	Unknown	40	ANOVA/RE (assumed)	No	Yes	No	No	No	No
29	Laboratory	Serum Cystatin C	1	1	Non-healthy participants	134	ANOVA/RE (assumed)	No	Yes	Yes	No	No	No
30	Laboratory	GFR and creatinine	2	2	Non-healthy participants	85	ANOVA/RE (assumed)	No	Yes	No	No	No	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
31	Physiological	Sputum and FEV	14	14	Non-healthy participants	12	ANOVA/RE (assumed)	No	No	No	No	No	Yes
32	Physiological	Spirometry	4	4	Non-healthy participants	7101	Other	No	No	No	No	No	No
33	Physiological	Pulmonary function tests	11	11	Healthy participants	9	Other	No	No	No	No	No	No
34	Physiological	Spirometry	1	2	Mixed participants	20	Other	No	No	No	No	No	No
35	Physiological	Peak flow	1	1	Non-healthy participants	10	ANOVA/RE	No	No	No	No	No	No
36	Physiological	Respiratory measures	5	30	Mixed participants	30	ANOVA/RE	No	Yes	No	No	No	Yes
37	Laboratory	Cholesterol	9	18	Healthy participants	15	ANOVA/RE	No	Yes	Yes	No	Yes	Yes
38	Laboratory	Bone turnover	7	14	Mixed participants	29	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
39	Laboratory	Seminal parameters	6	6	Healthy participants	20	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
40	Laboratory	Cystatin C	2	2	Unknown	30	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	Yes	No
41	Laboratory	Male hormone	7	7	Healthy participants	27	ANOVA/RE	No	Yes	Yes	No	No	No
42	Laboratory	Hormone	5	5	Healthy participants	10	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
43	Laboratory	HbA1c	1	3	Healthy participants	18	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
44	Laboratory	Blood (RAGE)	1	1	Healthy participants	21	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
45	Laboratory	CHD risk markers	9	9	Unknown	17	ANOVA/RE	Yes	Yes	Yes	No	No	No
46	Laboratory	HbA1c	3	6	Mixed participants	30	ANOVA/RE	Yes	Yes	Yes	Yes	No	No
47	Laboratory	Electrolytes (ICU testing)	9	9	Non-healthy participants		ANOVA/RE (assumed)	Yes	Yes	No	No	No	No
48	Laboratory	Dehydration markers	6	6	Healthy participants	18	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	Yes	No
49	Laboratory	CRP	1	1	Mixed participants	23	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	No	Yes
50	Laboratory	Plasma blood	25	25	Healthy participants	11	ANOVA/RE	No	Yes	Yes	Yes	No	No
51	Laboratory	MPO	1	1	Healthy participants	12	ANOVA/RE	Yes	Yes	Yes	No	Yes	No
52	Laboratory	HbA1c	1	1	Non-healthy participants	38	ANOVA/RE	No	Yes	Yes	Yes	No	No
53	Laboratory	Plasma	1	1	Healthy participants	22	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
54	Laboratory	C-peptide	1	1	Healthy participants	15	ANOVA/RE	No	Yes	Yes	Yes	No	No
55	Laboratory	Tumour markers	2	4	Mixed participants	43	ANOVA/RE	No	Yes	No	No	No	No
56	Laboratory	Troponin T	2	4	Healthy participants	37	ANOVA/RE	Yes	Yes	No	Yes	No	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
57	Laboratory	Blood markers	6	18	Healthy participants	21	ANOVA/RE	No	Yes	Yes	No	Yes	No
58	Laboratory	Cytokines	3	9	Healthy participants	15	ANOVA/RE	No	Yes	Yes	Yes	Yes	Yes
59	Laboratory	Thyroid markers	3	4	Healthy participants	24	ANOVA/RE	No	Yes	Yes	No	No	No
60	Laboratory	D-dimer	1	3	Mixed participants	36	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
61	Laboratory	LDL	1	3	Healthy participants	24	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
62	Laboratory	klk6 CKD	1	1	Healthy participants	4	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
63	Laboratory	Vitamin B intake	6	6	Healthy participants	22	ANOVA/RE	Yes	Yes	Yes	No	No	Yes
64	Laboratory	Serum amyloid A (SSA) and CRP	2	6	Healthy participants	24	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
65	Laboratory	Brain prohormone	1	3	Healthy participants	16	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
66	Laboratory	Growth factor	3	12	Healthy participants	28	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
67	Laboratory	CD163	1	2	Healthy participants	12	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
68	Laboratory	HbA(2)	1	1	Healthy participants	17	ANOVA/RE	No	Yes	Yes	No	No	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
69	Laboratory	Growth factor	6	6	Healthy participants	1103	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
70	Laboratory	Hepatic enzymes	3	9	Healthy participants	10	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
71	Laboratory	Biochemical and haematological analytes (elderly population)	26	52	Healthy participants	253	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
72	Laboratory	glucose-6-phosphate (dried blood spots newborn screening)	1	1	Healthy participants	20	ANOVA/RE	Yes	Yes	Yes	No	Yes	No
73	Laboratory	Cystatin C	2	4	Mixed participants	39	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
74	Laboratory	Glycohemoglobin	2	2	Unknown	48	ANOVA/RE	No	Yes	Yes	No	No	No
75	Laboratory	Liver fibrosis markers	3	12	Mixed participants	80	ANOVA/RE (assumed)	No	Yes	Yes	Yes	No	No
76	Laboratory	Plasma protein	1	1	Healthy participants	11	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
77	Laboratory	Plasma adiponectin	3	3	Mixed participants	20	ANOVA/RE	No	Yes	Yes	Yes	Yes	No
78	Laboratory	Vitamins	15	15	Healthy participants	14	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
79	Laboratory	HbA1c	1	4	Non-healthy participants	47	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
80	Laboratory	AFP	1	7	Mixed participants	115	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
81	Laboratory	Tumour markers	3	6	Mixed participants	40	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
82	Laboratory	AFP	2	2	Non-healthy participants	28	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
83	Laboratory	Magnesium and calcium	2	2	Healthy participants	51	ANOVA/RE (assumed)	No	Yes	Yes	Yes	No	No
84	Laboratory	Sweat sodium chloride conductivity	1	3	Mixed participants	55	ANOVA/RE (assumed)	No	Yes	Yes	No	Yes	No
85	Laboratory	Oxidative status	7	7	Healthy participants	25	ANOVA/RE (assumed)	Yes	Yes	Yes	No	No	No
86	Laboratory	Cardiac troponin	1	2	Healthy participants	20	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
87	Laboratory	Cardiac troponin	1	2	Healthy participants		ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
88	Laboratory	Parathyroid hormone	1	1	Healthy participants	20	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	Yes	No
89	Laboratory	Cardiac troponin	1	2	Mixed participants	29	ANOVA/RE	Yes	Yes	Yes	Yes	Yes	No
90	Laboratory	Cardiac troponin	1	1	Unknown	19	ANOVA/RE	No	Yes	Yes	Yes	Yes	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
91	Laboratory	Creatinine and cystatin C	2	2	Healthy participants	10	ANOVA/RE (assumed)	No	Yes	Yes	Yes	Yes	No
92	Laboratory	Cystatin C	3	3	Healthy participants	13	ANOVA/RE (assumed)	Yes	Yes	No	Yes	No	No
93	Laboratory	Serum creatinine, cystatin C, and creatinine clearance	4	4	Healthy participants	31	ANOVA/RE (assumed)	No	Yes	Yes	No	No	No
94	Laboratory	Plasma clearance of iohexol (GFR)	1	6	Non-healthy participants	24	ANOVA/RE (assumed)	No	Yes	No	No	No	No
95	Laboratory	Creatinine	5	11	Healthy participants	15	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	Yes	No
96	Laboratory	Cystatin c	2	2	Healthy participants	12	ANOVA/RE (assumed)	Yes	Yes	Yes	Yes	No	Yes
97	Imaging	US bladder wall thickness	1	1	Non-healthy participants	10	Other	No	No	No	No	No	No
98	Imaging	Sonography detrusor wall thickness	1	1	Mixed participants	10	Other	No	No	No	No	No	No
99	Imaging	US detrusor wall thickness	1	1	Unknown	67	ANOVA/RE	No	No	No	No	No	Yes
100	Imaging	US vaginal wall thickness	6	6	Unknown	25	Other	No	No	No	No	No	No

See §3.3.1 for detail of searches.

ID	Test type	Measure	Tests	Situations	Participants	Sample size	Analysis	CV_A	CV_I	CV_G	RCV	II	ICC
101	Imaging	US bladder wall thickness	1	1	Unknown	40	Other	No	No	No	No	No	No

See §3.3.1 for detail of searches.

Appendix B

Analysis of biological variability: a case study evaluating glomerular filtration rate (GFR)

B.1 Detailed results of analyses

Table B.1: Analysis of the eGFR-C biological variability study–results of iohexol outlier tests.

	Not transformed			Transformed		
Outlier Detection	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**
Observations	150	142	138	150[a]	142[a]	138[a]
Participants	20	20	19	20	20	19
Mean (SD)	47.6 (8.4)	47.7 (8.4)	47.8 (8.4)	3.8 (0.2)	3.9 (0.2)	3.9 (0.2)
Median (IQR)	47.8 (40.8, 52.3)	47.8 (40.8, 52.2)	47.8 (41.1, 52.2)	3.9 (3.7, 4.0)	3.9 (3.7, 4.0)	3.9 (3.7, 4.0)
Cochran C duplicates I						
Observations	-	6	6	-	6	6
Duplicates	-	3	3	-	3	3
Participants	-	3	3	-	3	3
Cochran C duplicates II						
Observations	-	-	0	-	-	0
Duplicates	-	-	0	-	-	0
Participants	-	-	0	-	-	0
Cochran C individuals I						
Observations	-	2	6	-	2	6
Duplicates	-	1	3	-	1	3
Participants	-	1	1	-	1	1
Cochran C individuals II						
Observations	-	-	0	-	-	0
Duplicates	-	-	0	-	-	0
Participants	-	-	0	-	-	0
Reed's test						
Observations	-	0	0	-	0	0
Duplicates	-	0	0	-	0	0
Participants	-	0	0	-	0	0
Total outliers[a]	0	8	12	0	8	12

[a] Eight observations were removed prior to analysis due to issues with test procedure.

Table B.2: Analysis of the eGFR-C biological variability study–creatinine outlier tests.

	Not transformed		Transformed	
Outlier Detection	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**
Observations	158	154	158	154
Participants	20	20	20	20
Mean (SD)	127.6 (25.5)	127.1 (25.1)	4.8 (0.2)	4.8 (0.2)
Median (IQR)	124 (109, 144)	124 (109, 144)	4.8 (4.7, 5.0)	4.8 (4.7, 5.0)
Cochran C duplicates I				
Observations	-	4	-	4
Duplicates	-	2	-	2
Participants	-	2	-	2
Cochran C duplicates II				
Observations	-	-/0	-	-/0
Duplicates	-	-/0	-	-/0
Participants	-	-/0	-	-/0
Cochran C individuals I				
Observations	-	0	-	0
Duplicates	-	0	-	0
Participants	-	0	-	0
Cochran C individuals II				
Observations	-	-/0	-	-/0
Duplicates	-	-/0	-	-/0
Participants	-	-/0	-	-/0
Reed's test				
Observations	-	0	-	0
Duplicates	-	0	-	0
Participants	-	0	-	0
Total Outliers	0	4	0	4

Table B.3: Analysis of the eGFR-C biological variability study–Cystatin C outlier tests.

	Not transformed			Transformed	
Outlier Detection	**None**	**Fraser-Harris method**	**Cochran C test & Reed's Criterion**	**None**	**Fraser-Harris method/Cochran C test & Reed's Criterion**
Observations	158	150	136	158	152
Participants	20	20	18	20	20
Mean (SD)	1.7 (0.3)	1.7 (0.3)	1.6 (0.3)	0.5 (0.2)	0.5 (0.2)
Median (IQR)	1.7 (1.5, 1.9)	1.7 (1.5, 1.9)	1.6 (1.5, 1.9)	0.5 (0.4, 0.6)	0.5 (0.4, 0.6)
Cochran C duplicates I					
Observations	-	6	6	-	6
Duplicates	-	3	3	-	3
Participants	-	3	3	-	3
Cochran C duplicates II					
Observations	-	0	0	-	-/0
Duplicates	-	0	0	-	-/0
Participants	-	0	0	-	-/0
Cochran C individuals I					
Observations	-	2	8	-	0
Duplicates	-	1	4	-	0
Participants	-	1	2	-	0
Cochran C individuals II					
Observations	-	0	8	-	-/0
Duplicates	-	0	4	-	-/0
Participants	-	0	2	-	-/0
Cochran C individuals III					
Observations	-	-	0	-	-
Duplicates	-	-	0	-	-
Participants	-	-	0	-	-
Reed's test					
Observations	-	0	0	-	-/0
Duplicates	-	0	0	-	-/0
Participants	-	0	0	-	-/0
Total outliers	0	8	22	0	6

Appendix C

Sample size guidance and justification for studies of biological variation

C.1 Results of the normally distributed data simulation; varying sample size

Table C.1: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	0.5	1	2	0.50 (0.44, 0.55)[0.25, 0.76]	0.98 (0.84, 1.13)[0.35, 1.54]	1.81 (1.37, 2.31)[0.00, 4.11]
10	4	2	0.5	1	2	0.50 (0.46, 0.53)[0.31, 0.71]	1.00 (0.89, 1.09)[0.57, 1.54]	1.91 (1.57, 2.26)[0.51, 3.70]
20	4	2	0.5	1	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
30	4	2	0.5	1	2	0.50 (0.48, 0.52)[0.40, 0.62]	0.99 (0.94, 1.05)[0.67, 1.32]	1.98 (1.79, 2.17)[1.03, 3.19]
40	4	2	0.5	1	2	0.50 (0.48, 0.52)[0.42, 0.58]	1.00 (0.95, 1.05)[0.73, 1.22]	1.99 (1.82, 2.15)[1.31, 2.81]
60	4	2	0.5	1	2	0.50 (0.48, 0.51)[0.43, 0.57]	1.00 (0.96, 1.04)[0.77, 1.19]	2.00 (1.87, 2.12)[1.40, 2.73]
100	4	2	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.56]	1.00 (0.97, 1.03)[0.86, 1.16]	2.00 (1.89, 2.11)[1.58, 2.47]
20	**2**	2	0.5	1	2	0.50 (0.46, 0.53)[0.31, 0.68]	0.98 (0.87, 1.10)[0.48, 1.55]	1.97 (1.70, 2.22)[0.74, 3.23]
20	**4**	2	0.5	1	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	**6**	2	0.5	1	2	0.50 (0.48, 0.52)[0.40, 0.59]	1.00 (0.95, 1.05)[0.70, 1.28]	1.98 (1.74, 2.20)[0.96, 3.25]
20	**8**	2	0.5	1	2	0.50 (0.48, 0.52)[0.40, 0.60]	1.00 (0.95, 1.05)[0.81, 1.20]	1.94 (1.72, 2.18)[0.98, 3.05]
20	**12**	2	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.57]	1.00 (0.96, 1.03)[0.82, 1.16]	1.97 (1.76, 2.21)[1.02, 3.13]
20	**20**	2	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.56]	1.00 (0.97, 1.03)[0.87, 1.12]	1.94 (1.74, 2.17)[0.99, 2.92]
20	4	**2**	0.5	1	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	4	**3**	0.5	1	2	0.50 (0.48, 0.52)[0.42, 0.58]	0.99 (0.93, 1.06)[0.68, 1.28]	1.97 (1.73, 2.22)[0.77, 3.27]
20	4	**4**	0.5	1	2	0.50 (0.48, 0.51)[0.43, 0.59]	0.99 (0.93, 1.06)[0.74, 1.32]	1.98 (1.74, 2.24)[1.03, 3.15]
20	4	**6**	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.57]	1.00 (0.94, 1.06)[0.67, 1.31]	1.98 (1.72, 2.20)[1.06, 3.27]
20	4	**10**	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	0.99 (0.93, 1.06)[0.74, 1.30]	1.97 (1.75, 2.21)[0.96, 3.10]

Table C.2: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
5	4	2	5	10	20	0.56	30.99	4.96 (4.35, 5.52)[2.51, 8.29]	9.85 (8.34, 11.34)[3.28, 17.19]	17.98 (13.46, 23.44)[0.00, 52.91]
10	4	2	5	10	20	0.56	30.99	4.97 (4.55, 5.42)[3.04, 7.45]	10.00 (8.93, 11.07)[5.41, 17.11]	19.06 (15.69, 23.11)[5.49, 36.44]
20	4	2	5	10	20	0.56	30.99	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
30	4	2	5	10	20	0.56	30.99	4.97 (4.75, 5.25)[3.96, 6.56]	9.91 (9.34, 10.61)[6.87, 14.02]	19.85 (17.82, 21.81)[10.37, 31.96]
40	4	2	5	10	20	0.56	30.99	5.00 (4.77, 5.23)[4.11, 5.96]	9.98 (9.41, 10.59)[7.40, 12.98]	19.85 (18.09, 21.52)[12.96, 28.08]
60	4	2	5	10	20	0.56	30.99	4.99 (4.82, 5.17)[4.24, 5.98]	10.01 (9.60, 10.45)[8.08, 12.14]	19.94 (18.64, 21.30)[14.01, 26.74]
100	4	2	5	10	20	0.56	30.99	5.00 (4.87, 5.14)[4.44, 5.69]	10.01 (9.64, 10.33)[8.49, 11.68]	19.92 (18.89, 21.17)[16.02, 24.71]
20	**2**	2	5	10	20	0.56	30.99	5.00 (4.59, 5.40)[2.96, 7.39]	9.84 (8.74, 11.09)[5.03, 16.05]	19.75 (16.82, 22.33)[7.15, 34.30]
20	**4**	2	5	10	20	0.56	30.99	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
20	**6**	2	5	10	20	0.56	30.99	4.98 (4.74, 5.26)[3.71, 6.19]	10.00 (9.40, 10.62)[6.55, 13.40]	19.83 (17.53, 22.14)[9.60, 33.09]
20	**8**	2	5	10	20	0.56	30.99	5.00 (4.76, 5.23)[3.88, 6.03]	9.95 (9.43, 10.61)[7.59, 13.52]	19.43 (17.14, 21.96)[9.79, 33.20]
20	**12**	2	5	10	20	0.56	30.99	4.99 (4.79, 5.21)[4.02, 5.93]	9.98 (9.48, 10.44)[7.77, 12.57]	19.78 (17.47, 22.14)[10.37, 32.09]
20	**20**	2	5	10	20	0.56	30.99	5.01 (4.82, 5.21)[4.10, 6.14]	10.00 (9.61, 10.44)[8.18, 11.91]	19.57 (17.47, 21.66)[9.26, 30.21]
20	4	**2**	5	10	20	0.56	30.99	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
20	4	**3**	5	10	20	0.56	30.99	5.00 (4.77, 5.27)[3.83, 6.25]	9.95 (9.24, 10.68)[6.76, 14.28]	19.77 (17.36, 22.31)[7.85, 32.12]
20	4	**4**	5	10	20	0.56	30.99	4.99 (4.79, 5.21)[4.04, 6.16]	9.94 (9.17, 10.72)[7.14, 14.03]	19.90 (17.25, 22.51)[9.87, 31.56]
20	4	**6**	5	10	20	0.56	30.99	5.00 (4.82, 5.21)[4.13, 6.04]	9.95 (9.28, 10.68)[6.48, 14.04]	19.82 (17.14, 22.17)[10.23, 32.17]
20	4	**10**	5	10	20	0.56	30.99	5.00 (4.82, 5.19)[4.24, 5.77]	9.94 (9.27, 10.65)[6.95, 13.23]	19.65 (17.36, 21.99)[9.45, 30.58]

Table C.3: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
5	4	2	5	10	20	0.56	30.99	0.61 (0.46, 0.83)[0.20, 28,030.82]	30.58 (26.76, 34.65)[14.97, 50.75]	9.99 (9.37, 10.68)[6.60, 12.51]
10	4	2	5	10	20	0.56	30.99	0.58 (0.48, 0.72)[0.27, 2.14]	31.14 (28.28, 33.72)[18.97, 50.19]	9.99 (9.55, 10.41)[7.63, 12.11]
20	4	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
30	4	2	5	10	20	0.56	30.99	0.56 (0.51, 0.63)[0.36, 1.13]	30.81 (29.31, 32.48)[23.76, 41.52]	10.01 (9.73, 10.25)[8.85, 11.39]
40	4	2	5	10	20	0.56	30.99	0.56 (0.51, 0.62)[0.35, 0.93]	30.93 (29.52, 32.49)[24.95, 38.92]	9.99 (9.77, 10.22)[9.13, 11.00]
60	4	2	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.39, 0.78]	31.02 (29.94, 32.20)[25.96, 36.45]	9.99 (9.81, 10.17)[9.30, 10.94]
100	4	2	5	10	20	0.56	30.99	0.56 (0.53, 0.59)[0.43, 0.74]	30.99 (30.07, 31.87)[26.69, 35.65]	9.99 (9.85, 10.15)[9.33, 10.72]
20	**2**	2	5	10	20	0.56	30.99	0.56 (0.47, 0.68)[0.28, 1.97]	30.68 (27.86, 33.80)[18.46, 47.06]	10.01 (9.67, 10.32)[8.27, 11.44]
20	**4**	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
20	**6**	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.35, 1.18]	30.96 (29.47, 32.63)[22.79, 40.01]	9.99 (9.67, 10.31)[8.52, 11.32]
20	**8**	2	5	10	20	0.56	30.99	0.57 (0.51, 0.65)[0.36, 1.21]	30.88 (29.42, 32.61)[23.93, 40.32]	10.00 (9.69, 10.31)[8.29, 11.52]
20	**12**	2	5	10	20	0.56	30.99	0.56 (0.50, 0.64)[0.35, 1.10]	30.94 (29.62, 32.17)[24.77, 38.26]	10.00 (9.71, 10.30)[8.75, 11.41]
20	**20**	2	5	10	20	0.56	30.99	0.57 (0.51, 0.64)[0.37, 1.10]	31.04 (29.91, 32.23)[25.88, 36.52]	9.98 (9.68, 10.28)[8.58, 11.43]
20	4	**2**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
20	4	**3**	5	10	20	0.56	30.99	0.56 (0.50, 0.65)[0.30, 1.59]	30.89 (29.02, 32.81)[22.62, 42.76]	9.99 (9.67, 10.31)[8.63, 11.40]
20	4	**4**	5	10	20	0.56	30.99	0.56 (0.49, 0.65)[0.32, 1.13]	30.78 (28.83, 32.89)[22.98, 41.46]	10.01 (9.69, 10.31)[8.47, 11.36]
20	4	**6**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.05]	30.95 (29.21, 32.79)[22.35, 41.96]	9.99 (9.68, 10.27)[8.49, 11.30]
20	4	**10**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.34, 1.22]	30.81 (29.16, 32.64)[23.19, 39.63]	9.99 (9.70, 10.33)[8.67, 11.50]

C.2 Results for log-normal data simulation; varying sample size

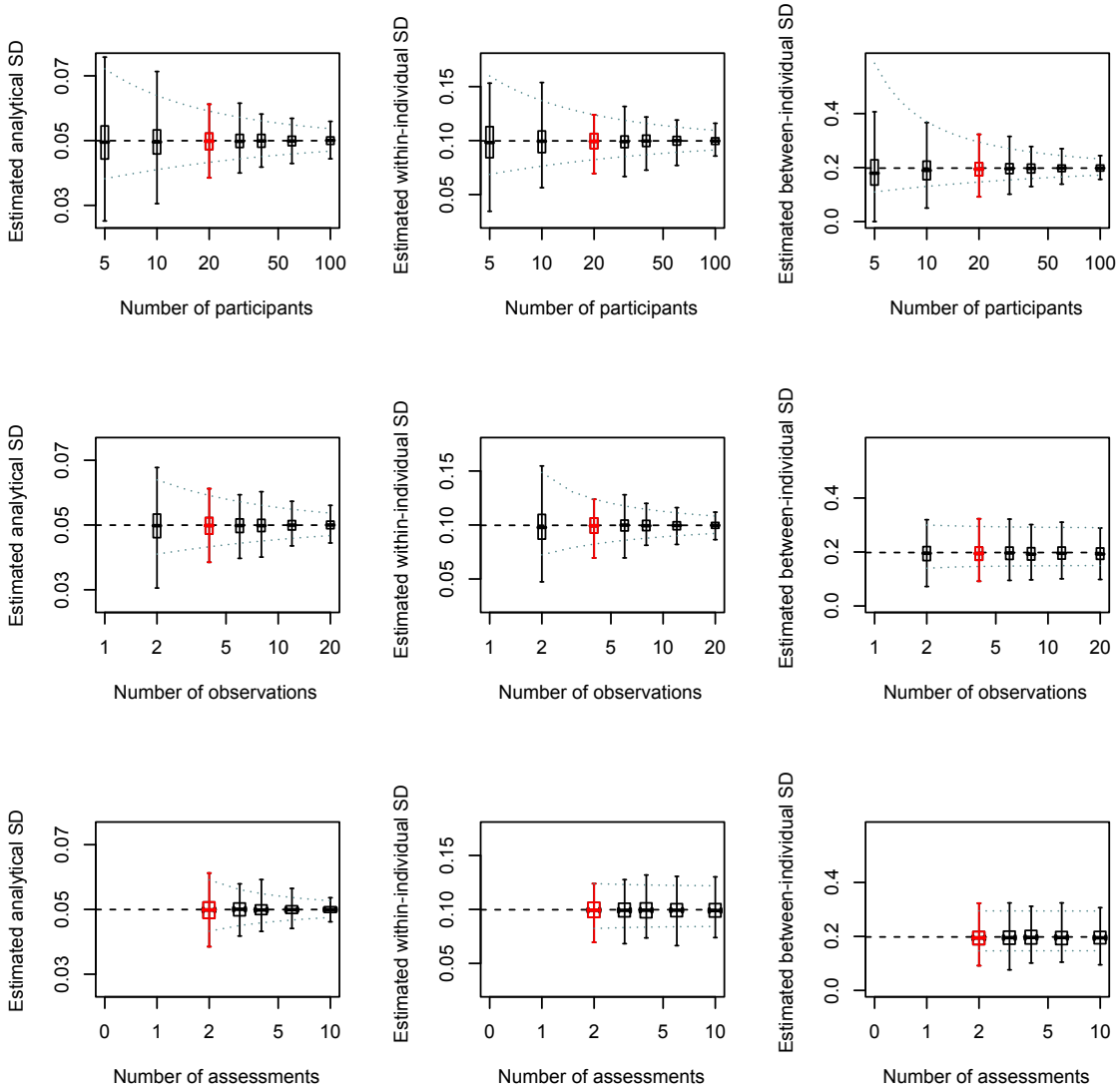


Figure C.1: Log-normal biological variability sample size simulation: SD estimates from biological variability data simulations varying sample size: SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying number of participants (n_1, top row), number of observations per participant (n_2, middle row), and number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true SD and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

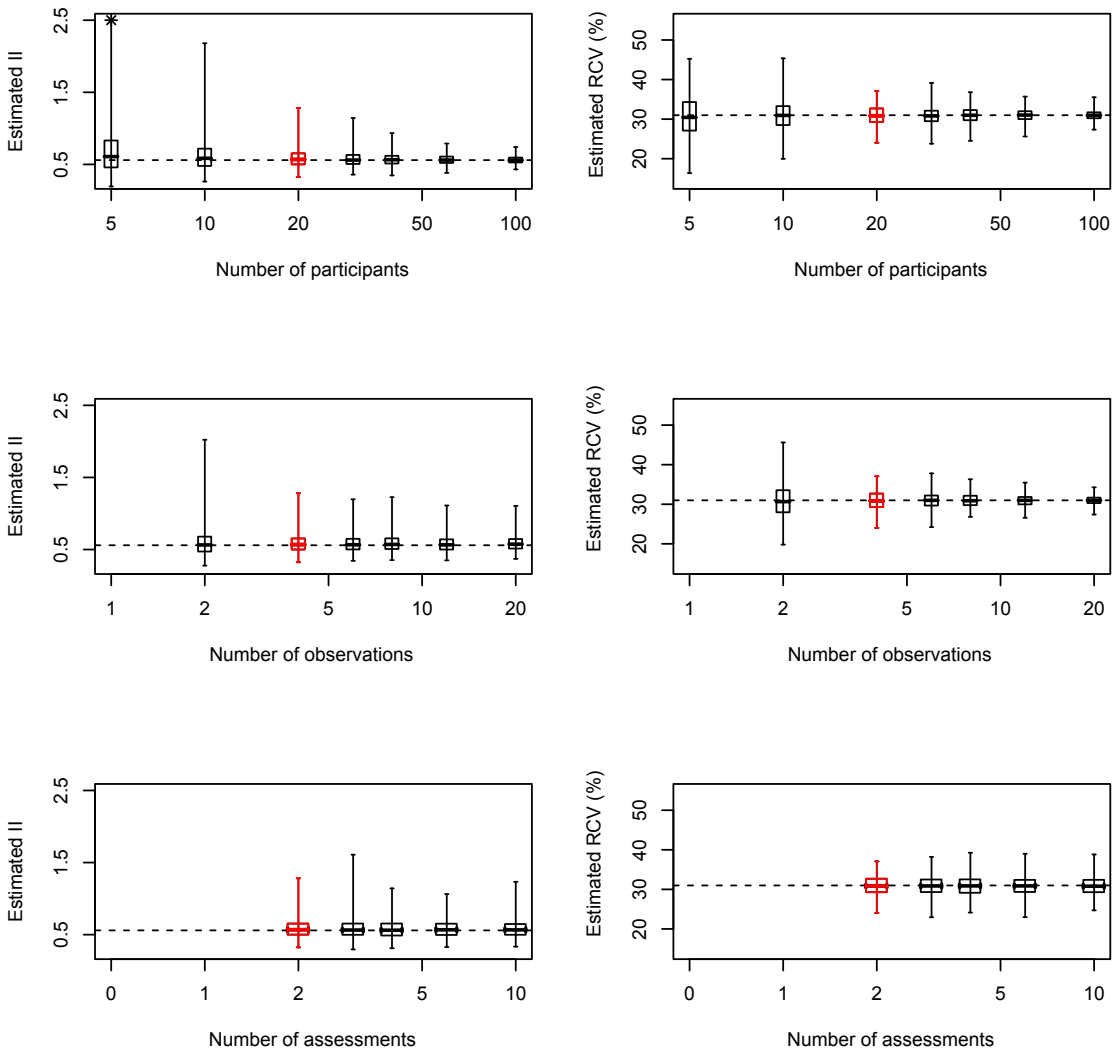


Figure C.2: Log-normal biological variability sample size simulation: II and RCV estimates from biological variability data simulations varying sample size: II (left column) and RCV (right column) estimates when varying number of participants (n_1, top row), number of observations per participant (n_2, middle row), and number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true II or RCV.

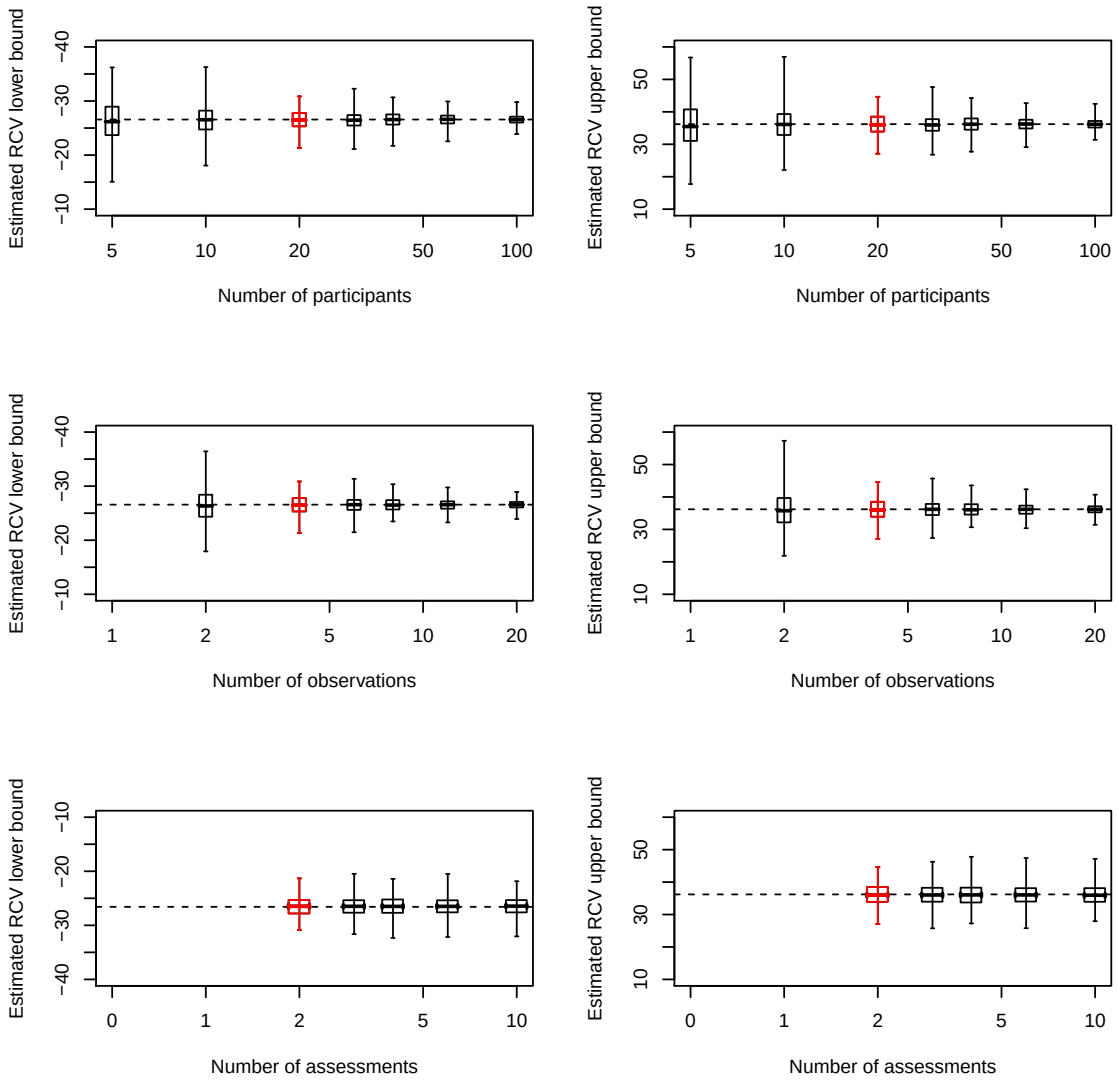


Figure C.3: Log-normal biological variability sample size simulation: asymmetric RCV estimates from biological variability data simulations varying sample size: RCV lower bound (left column) and RCV upper bound (right column) estimates when varying number of participants (n_1, top row), number of observations per participant (n_2, middle row), and number of replicate assessments per observation per participant (n_3, bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true RCV bound.

Table C.4: Log normal simulation: bias performance measures varying number of participants, observations and assessments.

Inputs						Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	0.05	0.1	0.2	-2.816	-18.327	-143.968	-0.564	-1.837	-7.270	-3.697	-8.796	-20.233
10	4	2	0.05	0.1	0.2	-3.111	-3.279	-64.312	-0.623	-0.329	-3.247	-5.697	-2.306	-12.778
20	4	2	0.05	0.1	0.2	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
30	4	2	0.05	0.1	0.2	-0.639	-6.156	-16.875	-0.128	-0.617	-0.852	-2.067	-7.236	-5.988
40	4	2	0.05	0.1	0.2	-0.540	-1.560	-13.908	-0.108	-0.156	-0.702	-1.968	-2.079	-5.645
60	4	2	0.05	0.1	0.2	-0.858	0.761	-6.208	-0.172	0.076	-0.313	-3.861	1.270	-3.282
100	4	2	0.05	0.1	0.2	0.361	-1.270	-0.256	0.072	-0.127	-0.013	2.069	-2.794	-0.169
20	**2**	2	0.05	0.1	0.2	-1.521	-13.547	-26.719	-0.304	-1.358	-1.349	-2.689	-7.899	-7.202
20	**4**	2	0.05	0.1	0.2	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
20	**6**	2	0.05	0.1	0.2	-1.434	-0.857	-17.434	-0.287	-0.086	-0.880	-4.517	-1.071	-5.091
20	**8**	2	0.05	0.1	0.2	-1.151	-0.830	-41.846	-0.230	-0.083	-2.113	-4.159	-1.241	-12.109
20	**12**	2	0.05	0.1	0.2	-0.111	-3.544	-11.459	-0.022	-0.355	-0.579	-0.496	-6.434	-3.485
20	**20**	2	0.05	0.1	0.2	0.106	-0.152	-38.421	0.021	-0.015	-1.940	0.617	-0.378	-12.481
20	4	**2**	0.05	0.1	0.2	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
20	4	**3**	0.05	0.1	0.2	0.570	-5.380	-19.387	0.114	-0.539	-0.979	1.996	-5.384	-5.596
20	4	**4**	0.05	0.1	0.2	-0.182	-5.645	-8.260	-0.036	-0.566	-0.417	-0.795	-5.776	-2.307
20	4	**6**	0.05	0.1	0.2	-0.236	-3.945	-21.483	-0.047	-0.395	-1.085	-1.359	-4.261	-6.163
20	4	**10**	0.05	0.1	0.2	0.033	-4.955	-26.301	0.007	-0.497	-1.328	0.253	-5.474	-7.681

Table C.5: Log normal simulation: accuracy and coverage performance measures varying number of participants, observations and assessments.

Inputs						**Accuracy and coverage**								
						Mean squared error ($\times 10^{-4}$)			**Coverage**			**Mean 95% CI width**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	0.05	0.1	0.2	0.581	4.375	52.701	0.956	0.962[a]	0.972[b]	0.034	0.092[a]	0.472[b]
10	4	2	0.05	0.1	0.2	0.299	2.023	25.745	0.958	0.957	0.951[c]	0.023	0.061	0.236[c]
20	4	2	0.05	0.1	0.2	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
30	4	2	0.05	0.1	0.2	0.096	0.728	7.969	0.961	0.945	0.957	0.013	0.033	0.115
40	4	2	0.05	0.1	0.2	0.075	0.563	6.090	0.955	0.945	0.939	0.011	0.029	0.098
60	4	2	0.05	0.1	0.2	0.049	0.359	3.582	0.957	0.948	0.953	0.009	0.024	0.079
100	4	2	0.05	0.1	0.2	0.030	0.207	2.302	0.952	0.950	0.954	0.007	0.018	0.060
20	**2**	2	0.05	0.1	0.2	0.320	2.959	13.834	0.939	0.953	0.957[d]	0.023	0.076	0.161[d]
20	**4**	2	0.05	0.1	0.2	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
20	**6**	2	0.05	0.1	0.2	0.101	0.640	11.755	0.951	0.945	0.945	0.013	0.032	0.144
20	**8**	2	0.05	0.1	0.2	0.077	0.447	12.118	0.954	0.950	0.934	0.011	0.027	0.141
20	**12**	2	0.05	0.1	0.2	0.050	0.305	10.825	0.950	0.947	0.947	0.009	0.021	0.141
20	**20**	2	0.05	0.1	0.2	0.030	0.161	9.624	0.955	0.953	0.960	0.007	0.016	0.138
20	4	**2**	0.05	0.1	0.2	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
20	4	**3**	0.05	0.1	0.2	0.082	1.001	12.039	0.948	0.941	0.954	0.011	0.040	0.147
20	4	**4**	0.05	0.1	0.2	0.052	0.958	12.823	0.958	0.949	0.935	0.009	0.039	0.148
20	4	**6**	0.05	0.1	0.2	0.030	0.859	12.197	0.960	0.958	0.942	0.007	0.038	0.146
20	4	**10**	0.05	0.1	0.2	0.017	0.822	11.793	0.949	0.956	0.944	0.005	0.038	0.146

[a]12 CIs could not be calculated;[b]79 CIs could not be calculated, [c]4 CIs could not be calculated, [d]2 CIs could not be calculated.

Table C.6: Log normal simulation: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments.

Inputs						Median (Q1, Q3)[minimum, maximum]			
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	
5	4	2	0.05	0.1	0.2	0.05 (0.04, 0.05)[0.03, 0.08]	0.10 (0.08, 0.11)[0.03, 0.15]	0.18 (0.14, 0.23)[0.00, 0.41]	
10	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.15]	0.19 (0.16, 0.22)[0.05, 0.37]	
20	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]	
30	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.10, 0.32]	
40	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.20 (0.18, 0.21)[0.13, 0.28]	
60	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.10, 0.10)[0.08, 0.12]	0.20 (0.19, 0.21)[0.14, 0.27]	
100	4	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.10, 0.10)[0.09, 0.12]	0.20 (0.19, 0.21)[0.16, 0.24]	
20	**2**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.05, 0.15]	0.20 (0.17, 0.22)[0.07, 0.32]	
20	**4**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]	
20	**6**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.17, 0.22)[0.09, 0.32]	
20	**8**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.08, 0.12]	0.19 (0.17, 0.22)[0.10, 0.30]	
20	**12**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.10, 0.10)[0.08, 0.12]	0.20 (0.17, 0.22)[0.10, 0.31]	
20	**20**	2	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.10, 0.10)[0.09, 0.11]	0.19 (0.17, 0.21)[0.10, 0.29]	
20	4	**2**	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]	
20	4	**3**	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.13]	0.20 (0.17, 0.22)[0.08, 0.32]	
20	4	**4**	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.13]	0.20 (0.17, 0.22)[0.10, 0.31]	
20	4	**6**	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.13]	0.20 (0.17, 0.22)[0.10, 0.32]	
20	4	**10**	0.05	0.1	0.2	0.05 (0.05, 0.05)[0.05, 0.05]	0.10 (0.09, 0.11)[0.07, 0.13]	0.19 (0.17, 0.22)[0.09, 0.31]	

Table C.7: Log normal simulation: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
5	4	2	5	10	20	0.56	30.99	4.96 (4.43, 5.46)[2.52, 7.59]	9.84 (8.43, 11.33)[3.45, 15.41]	18.06 (13.59, 23.18)[0.00, 42.41]
10	4	2	5	10	20	0.56	30.99	4.96 (4.59, 5.33)[3.06, 7.14]	9.97 (8.90, 10.93)[5.64, 15.47]	19.12 (15.62, 22.67)[5.00, 37.93]
20	4	2	5	10	20	0.56	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
30	4	2	5	10	20	0.56	30.99	4.98 (4.79, 5.19)[4.00, 6.16]	9.93 (9.37, 10.45)[6.68, 13.22]	19.77 (17.91, 21.69)[10.16, 32.35]
40	4	2	5	10	20	0.56	30.99	4.99 (4.80, 5.19)[4.18, 5.82]	9.98 (9.48, 10.53)[7.27, 12.23]	19.86 (18.16, 21.52)[13.03, 28.35]
60	4	2	5	10	20	0.56	30.99	4.99 (4.84, 5.14)[4.30, 5.69]	10.02 (9.61, 10.41)[7.71, 11.95]	20.00 (18.68, 21.18)[13.93, 27.50]
100	4	2	5	10	20	0.56	30.99	5.01 (4.89, 5.12)[4.44, 5.60]	10.00 (9.67, 10.29)[8.59, 11.64]	19.97 (18.88, 21.17)[15.72, 24.81]
20	**2**	2	5	10	20	0.56	30.99	4.98 (4.61, 5.34)[3.06, 6.78]	9.80 (8.71, 11.02)[4.74, 15.56]	19.72 (16.95, 22.27)[7.22, 32.80]
20	**4**	2	5	10	20	0.56	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
20	**6**	2	5	10	20	0.56	30.99	4.99 (4.77, 5.19)[3.97, 5.94]	10.02 (9.47, 10.49)[6.96, 12.86]	19.83 (17.38, 22.07)[9.51, 33.04]
20	**8**	2	5	10	20	0.56	30.99	4.98 (4.80, 5.18)[4.01, 6.03]	9.96 (9.51, 10.46)[8.14, 12.05]	19.37 (17.16, 21.84)[9.72, 30.87]
20	**12**	2	5	10	20	0.56	30.99	5.00 (4.85, 5.14)[4.36, 5.74]	9.98 (9.59, 10.33)[8.22, 11.65]	19.69 (17.52, 22.10)[10.09, 31.82]
20	**20**	2	5	10	20	0.56	30.99	5.00 (4.89, 5.12)[4.45, 5.61]	9.99 (9.72, 10.26)[8.66, 11.23]	19.42 (17.41, 21.75)[9.85, 29.52]
20	4	**2**	5	10	20	0.56	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
20	4	**3**	5	10	20	0.56	30.99	5.01 (4.81, 5.20)[4.18, 5.79]	9.94 (9.33, 10.60)[6.84, 12.81]	19.71 (17.24, 22.22)[7.62, 33.28]
20	4	**4**	5	10	20	0.56	30.99	4.99 (4.85, 5.14)[4.32, 5.93]	9.95 (9.25, 10.64)[7.37, 13.25]	19.81 (17.33, 22.50)[10.17, 31.94]
20	4	**6**	5	10	20	0.56	30.99	4.99 (4.88, 5.12)[4.42, 5.65]	9.95 (9.35, 10.56)[6.66, 13.12]	19.78 (17.11, 22.07)[10.51, 33.29]
20	4	**10**	5	10	20	0.56	30.99	5.00 (4.91, 5.09)[4.62, 5.37]	9.92 (9.32, 10.57)[7.40, 13.07]	19.66 (17.41, 22.11)[9.51, 31.39]

Table C.8: Log normal simulation: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
5	4	2	5	10	20	0.56	30.99	0.61 (0.46, 0.83)[0.19, 46076.64]	30.44 (27.10, 34.31)[16.35, 45.24]	9.97 (9.91, 10.04)[9.64, 10.22]
10	4	2	5	10	20	0.56	30.99	0.58 (0.48, 0.72)[0.26, 2.18]	30.93 (28.50, 33.28)[19.96, 45.36]	9.97 (9.93, 10.02)[9.74, 10.18]
20	4	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
30	4	2	5	10	20	0.56	30.99	0.56 (0.51, 0.63)[0.36, 1.14]	30.81 (29.48, 32.13)[23.78, 39.15]	9.97 (9.95, 10.00)[9.86, 10.11]
40	4	2	5	10	20	0.56	30.99	0.56 (0.51, 0.62)[0.35, 0.93]	30.97 (29.72, 32.29)[24.52, 36.79]	9.97 (9.95, 10.00)[9.89, 10.07]
60	4	2	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.38, 0.79]	31.03 (30.00, 32.00)[25.61, 35.69]	9.97 (9.96, 9.99)[9.90, 10.07]
100	4	2	5	10	20	0.56	30.99	0.56 (0.53, 0.59)[0.43, 0.74]	30.97 (30.23, 31.72)[27.36, 35.53]	9.97 (9.96, 9.99)[9.91, 10.05]
20	**2**	2	5	10	20	0.56	30.99	0.56 (0.47, 0.68)[0.28, 2.02]	30.66 (27.96, 33.54)[19.79, 45.61]	9.97 (9.94, 10.01)[9.80, 10.12]
20	**4**	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
20	**6**	2	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.34, 1.20]	30.98 (29.67, 32.21)[24.22, 37.80]	9.97 (9.94, 10.00)[9.83, 10.10]
20	**8**	2	5	10	20	0.56	30.99	0.57 (0.51, 0.65)[0.35, 1.23]	30.88 (29.77, 32.14)[26.81, 36.31]	9.97 (9.94, 10.01)[9.80, 10.12]
20	**12**	2	5	10	20	0.56	30.99	0.56 (0.50, 0.64)[0.35, 1.11]	30.98 (29.97, 31.83)[26.58, 35.46]	9.97 (9.95, 10.00)[9.85, 10.11]
20	**20**	2	5	10	20	0.56	30.99	0.57 (0.51, 0.65)[0.37, 1.10]	30.97 (30.30, 31.67)[27.40, 34.27]	9.97 (9.94, 10.00)[9.83, 10.12]
20	4	**2**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	**3**	5	10	20	0.56	30.99	0.56 (0.50, 0.65)[0.30, 1.61]	30.86 (29.32, 32.49)[22.96, 38.20]	9.97 (9.94, 10.01)[9.84, 10.11]
20	4	**4**	5	10	20	0.56	30.99	0.56 (0.49, 0.65)[0.31, 1.14]	30.88 (29.14, 32.53)[24.14, 39.25]	9.98 (9.94, 10.00)[9.82, 10.11]
20	4	**6**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.06]	30.87 (29.35, 32.40)[22.98, 38.99]	9.97 (9.94, 10.00)[9.82, 10.10]
20	4	**10**	5	10	20	0.56	30.99	0.57 (0.50, 0.65)[0.33, 1.23]	30.80 (29.29, 32.41)[24.70, 38.82]	9.97 (9.94, 10.01)[9.84, 10.12]

Table C.9: Log normal simulation: asymmetric RCV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs							**Median (Q1, Q3)[minimum, maximum]**	
n_1	n_2	n_3	CV_A	CV_I	CV_G	$RCV-$	$RCV+$	*RCV* lower bound	*RCV* upper bound
5	4	2	5	10	20	-26.58	36.20	-26.18 (-28.95, -23.69)[-36.20, -15.07]	35.46 (31.04, 40.75)[17.74, 56.74]
10	4	2	5	10	20	-26.58	36.20	-26.54 (-28.22, -24.74)[-36.28, -18.07]	36.12 (32.87, 39.32)[22.06, 56.93]
20	4	2	5	10	20	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
30	4	2	5	10	20	-26.58	36.20	-26.44 (-27.40, -25.47)[-32.26, -21.13]	35.95 (34.17, 37.74)[26.79, 47.63]
40	4	2	5	10	20	-26.58	36.20	-26.57 (-27.52, -25.65)[-30.67, -21.71]	36.18 (34.49, 37.97)[27.72, 44.24]
60	4	2	5	10	20	-26.58	36.20	-26.61 (-27.31, -25.85)[-29.91, -22.55]	36.25 (34.87, 37.56)[29.12, 42.68]
100	4	2	5	10	20	-26.58	36.20	-26.57 (-27.10, -26.02)[-29.80, -23.88]	36.18 (35.17, 37.18)[31.38, 42.45]
20	**2**	2	5	10	20	-26.58	36.20	-26.34 (-28.41, -24.34)[-36.43, -17.94]	35.75 (32.17, 39.68)[21.86, 57.30]
20	**4**	2	5	10	20	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	**6**	2	5	10	20	-26.58	36.20	-26.57 (-27.46, -25.61)[-31.36, -21.47]	36.19 (34.43, 37.85)[27.35, 45.68]
20	**8**	2	5	10	20	-26.58	36.20	-26.50 (-27.41, -25.69)[-30.34, -23.47]	36.05 (34.56, 37.76)[30.67, 43.56]
20	**12**	2	5	10	20	-26.58	36.20	-26.57 (-27.19, -25.83)[-29.75, -23.30]	36.18 (34.83, 37.33)[30.37, 42.36]
20	**20**	2	5	10	20	-26.58	36.20	-26.57 (-27.07, -26.07)[-28.93, -23.92]	36.18 (35.26, 37.12)[31.43, 40.70]
20	4	**2**	5	10	20	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	**3**	5	10	20	-26.58	36.20	-26.48 (-27.66, -25.35)[-31.63, -20.48]	36.02 (33.96, 38.24)[25.75, 46.26]
20	4	**4**	5	10	20	-26.58	36.20	-26.50 (-27.69, -25.22)[-32.33, -21.41]	36.05 (33.72, 38.28)[27.25, 47.79]
20	4	**6**	5	10	20	-26.58	36.20	-26.49 (-27.60, -25.37)[-32.16, -20.50]	36.04 (34.00, 38.12)[25.79, 47.40]
20	4	**10**	5	10	20	-26.58	36.20	-26.44 (-27.60, -25.33)[-32.04, -21.84]	35.94 (33.92, 38.13)[27.95, 47.15]

C.3 Results of the normally distributed data simulation; varying variability

Table C.10: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	1	2	0.12 (0.12, 0.13)[0.10, 0.15]	1.00 (0.93, 1.06)[0.71, 1.25]	1.96 (1.72, 2.21)[0.93, 3.25]
20	4	2	**0.25**	1	2	0.25 (0.24, 0.26)[0.19, 0.31]	1.00 (0.93, 1.06)[0.71, 1.25]	1.96 (1.72, 2.20)[0.93, 3.25]
20	4	2	**0.375**	1	2	0.37 (0.35, 0.39)[0.29, 0.46]	1.00 (0.93, 1.06)[0.71, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	4	2	**0.5**	1	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	4	2	**0.625**	1	2	0.62 (0.59, 0.65)[0.48, 0.77]	0.99 (0.92, 1.07)[0.67, 1.25]	1.96 (1.71, 2.21)[0.93, 3.27]
20	4	2	**0.75**	1	2	0.75 (0.71, 0.78)[0.58, 0.92]	0.99 (0.92, 1.08)[0.62, 1.28]	1.96 (1.71, 2.20)[0.93, 3.27]
20	4	2	**0.875**	1	2	0.87 (0.83, 0.92)[0.67, 1.07]	0.99 (0.91, 1.08)[0.56, 1.32]	1.96 (1.71, 2.20)[0.92, 3.27]
20	4	2	**1**	1	2	1.00 (0.95, 1.05)[0.77, 1.23]	0.99 (0.90, 1.09)[0.48, 1.37]	1.96 (1.72, 2.20)[0.90, 3.28]
20	4	2	0.5	**0.5**	2	0.50 (0.47, 0.52)[0.39, 0.61]	0.50 (0.45, 0.54)[0.24, 0.68]	1.96 (1.74, 2.20)[0.99, 3.22]
20	4	2	0.5	**0.75**	2	0.50 (0.47, 0.52)[0.39, 0.61]	0.75 (0.69, 0.80)[0.49, 0.95]	1.96 (1.73, 2.20)[0.96, 3.24]
20	4	2	0.5	**1**	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	4	2	0.5	**1.25**	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.25 (1.17, 1.33)[0.89, 1.55]	1.96 (1.71, 2.21)[0.88, 3.28]
20	4	2	0.5	**1.5**	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.49 (1.40, 1.60)[1.06, 1.87]	1.96 (1.69, 2.22)[0.83, 3.30]
20	4	2	0.5	**1.75**	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.74 (1.63, 1.86)[1.24, 2.18]	1.95 (1.68, 2.22)[0.76, 3.32]
20	4	2	0.5	**2**	2	0.50 (0.47, 0.52)[0.39, 0.61]	1.99 (1.87, 2.12)[1.42, 2.50]	1.95 (1.67, 2.23)[0.68, 3.33]
20	4	2	0.5	1	**1**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	0.98 (0.83, 1.12)[0.33, 1.67]
20	4	2	0.5	1	**1.5**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.47 (1.27, 1.65)[0.65, 2.47]
20	4	2	0.5	1	**2**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	1.96 (1.72, 2.20)[0.93, 3.26]
20	4	2	0.5	1	**2.5**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	2.46 (2.16, 2.75)[1.19, 4.05]
20	4	2	0.5	1	**3**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	2.95 (2.59, 3.29)[1.46, 4.84]
20	4	2	0.5	1	**3.5**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	3.43 (3.03, 3.84)[1.72, 5.63]
20	4	2	0.5	1	**4**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	3.92 (3.47, 4.39)[1.98, 6.42]
20	4	2	0.5	1	**4.5**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	4.41 (3.91, 4.94)[2.23, 7.20]
20	4	2	0.5	1	**5**	0.50 (0.47, 0.52)[0.39, 0.61]	1.00 (0.93, 1.07)[0.70, 1.24]	4.90 (4.34, 5.49)[2.49, 7.99]

Table C.11: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
20	4	2	**1.25**	10	20	**0.50**	**27.93**	1.25 (1.18, 1.33)[0.90, 1.71]	9.98 (9.30, 10.70)[6.60, 13.38]	19.59 (17.26, 22.22)[9.10, 33.33]
20	4	2	**2.5**	10	20	**0.52**	**28.57**	2.50 (2.36, 2.65)[1.79, 3.41]	9.98 (9.30, 10.72)[6.61, 13.29]	19.60 (17.23, 22.18)[9.10, 33.12]
20	4	2	**3.75**	10	20	**0.53**	**29.60**	3.75 (3.53, 3.98)[2.69, 5.12]	9.97 (9.29, 10.75)[6.63, 13.22]	19.64 (17.23, 22.20)[9.10, 32.91]
20	4	2	**5**	10	20	**0.56**	**30.99**	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
20	4	2	**6.25**	10	20	**0.59**	**32.69**	6.25 (5.88, 6.63)[4.48, 8.55]	9.96 (9.22, 10.81)[6.32, 13.69]	19.65 (17.24, 22.16)[9.10, 32.51]
20	4	2	**7.5**	10	20	**0.63**	**34.65**	7.50 (7.06, 7.96)[5.38, 10.27]	9.99 (9.15, 10.83)[5.88, 13.96]	19.64 (17.25, 22.17)[9.10, 32.31]
20	4	2	**8.75**	10	20	**0.66**	**36.83**	8.75 (8.24, 9.28)[6.28, 11.99]	10.00 (9.07, 10.90)[5.29, 14.24]	19.63 (17.21, 22.20)[9.11, 32.12]
20	4	2	**10**	10	20	**0.71**	**39.20**	10.00 (9.42, 10.61)[7.18, 13.71]	9.97 (8.96, 10.97)[4.54, 14.54]	19.61 (17.20, 22.20)[8.89, 31.94]
20	4	2	5	**5**	20	**0.35**	**19.60**	5.02 (4.72, 5.29)[3.60, 6.87]	4.99 (4.47, 5.48)[2.28, 7.33]	19.69 (17.38, 22.23)[9.67, 32.42]
20	4	2	5	**7.5**	20	**0.45**	**24.99**	5.01 (4.71, 5.29)[3.59, 6.85]	7.48 (6.90, 8.10)[4.68, 10.37]	19.68 (17.30, 22.20)[9.43, 32.55]
20	4	2	5	**10**	20	**0.56**	**30.99**	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
20	4	2	5	**12.5**	20	**0.67**	**37.32**	5.00 (4.71, 5.31)[3.58, 6.82]	12.47 (11.61, 13.43)[8.28, 16.59]	19.57 (17.08, 22.20)[8.68, 32.89]
20	4	2	5	**15**	20	**0.79**	**43.83**	5.01 (4.71, 5.31)[3.57, 6.80]	14.95 (13.93, 16.13)[9.89, 19.89]	19.53 (16.87, 22.26)[8.15, 33.09]
20	4	2	5	**17.5**	20	**0.91**	**50.45**	5.01 (4.72, 5.31)[3.56, 6.78]	17.44 (16.26, 18.81)[11.50, 23.28]	19.50 (16.76, 22.31)[7.50, 33.32]
20	4	2	5	**20**	20	**1.03**	**57.14**	5.00 (4.72, 5.32)[3.55, 6.79]	19.92 (18.59, 21.48)[13.10, 26.68]	19.47 (16.71, 22.34)[6.67, 33.57]
20	4	2	5	10	**10**	**1.12**	30.99	5.00 (4.73, 5.26)[3.78, 6.42]	9.96 (9.31, 10.73)[6.76, 12.72]	9.75 (8.33, 11.20)[3.30, 16.41]
20	4	2	5	10	**15**	**0.75**	30.99	5.00 (4.73, 5.27)[3.68, 6.52]	9.95 (9.31, 10.74)[6.66, 13.00]	14.69 (12.73, 16.67)[6.42, 24.03]
20	4	2	5	10	**20**	**0.56**	30.99	5.00 (4.71, 5.30)[3.59, 6.84]	9.97 (9.28, 10.77)[6.55, 13.44]	19.63 (17.24, 22.16)[9.10, 32.71]
20	4	2	5	10	**25**	**0.45**	30.99	5.00 (4.69, 5.33)[3.50, 7.20]	9.99 (9.23, 10.81)[6.45, 14.07]	24.62 (21.67, 27.95)[11.66, 42.80]
20	4	2	5	10	**30**	**0.37**	30.99	5.01 (4.68, 5.36)[3.41, 7.60]	10.02 (9.21, 10.88)[6.36, 14.80]	29.51 (26.06, 33.72)[14.14, 54.01]
20	4	2	5	10	**35**	**0.32**	30.99	5.01 (4.66, 5.39)[3.33, 8.05]	10.03 (9.17, 10.93)[6.26, 15.61]	34.49 (30.25, 39.48)[16.58, 66.54]
20	4	2	5	10	**40**	**0.28**	30.99	5.01 (4.64, 5.42)[3.26, 8.55]	10.03 (9.12, 10.97)[6.17, 16.52]	39.48 (34.53, 45.42)[18.97, 80.62]
20	4	2	5	10	**45**	**0.25**	30.99	5.01 (4.61, 5.45)[3.18, 9.12]	10.04 (9.08, 11.03)[6.08, 17.54]	44.67 (38.86, 51.22)[21.33, 96.56]
20	4	2	5	10	**50**	**0.22**	30.99	5.00 (4.59, 5.48)[3.11, 9.78]	10.07 (9.03, 11.07)[6.00, 18.68]	49.44 (43.08, 57.25)[23.66, 114.76]

Table C.12: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
20	4	2	**1.25**	10	20	**0.50**	**27.93**	0.51 (0.45, 0.59)[0.29, 1.16]	27.89 (26.00, 29.90)[18.57, 37.23]	9.98 (9.66, 10.29)[8.37, 11.55]
20	4	2	**2.5**	10	20	**0.52**	**28.57**	0.53 (0.46, 0.61)[0.30, 1.18]	28.54 (26.69, 30.60)[19.46, 37.47]	9.98 (9.66, 10.28)[8.38, 11.56]
20	4	2	**3.75**	10	20	**0.53**	**29.60**	0.55 (0.48, 0.62)[0.31, 1.22]	29.54 (27.78, 31.58)[20.87, 39.21]	9.97 (9.67, 10.28)[8.38, 11.57]
20	4	2	**5**	10	20	**0.56**	**30.99**	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
20	4	2	**6.25**	10	20	**0.59**	**32.69**	0.60 (0.53, 0.69)[0.35, 1.33]	32.69 (30.76, 34.74)[24.34, 44.28]	9.97 (9.68, 10.28)[8.38, 11.59]
20	4	2	**7.5**	10	20	**0.63**	**34.65**	0.63 (0.56, 0.73)[0.37, 1.41]	34.68 (32.64, 36.77)[25.89, 47.25]	9.97 (9.68, 10.28)[8.38, 11.60]
20	4	2	**8.75**	10	20	**0.66**	**36.83**	0.67 (0.60, 0.78)[0.40, 1.52]	36.86 (34.81, 39.03)[27.70, 50.43]	9.98 (9.68, 10.28)[8.37, 11.61]
20	4	2	**10**	10	20	**0.71**	**39.20**	0.72 (0.63, 0.83)[0.43, 1.70]	39.25 (37.13, 41.58)[29.73, 53.79]	9.98 (9.67, 10.28)[8.37, 11.62]
20	4	2	5	**5**	20	**0.35**	**19.60**	0.36 (0.32, 0.41)[0.22, 0.74]	19.61 (18.59, 20.76)[14.80, 26.88]	9.98 (9.68, 10.27)[8.34, 11.48]
20	4	2	5	**7.5**	20	**0.45**	**24.99**	0.46 (0.41, 0.52)[0.27, 0.98]	25.01 (23.55, 26.53)[18.56, 33.91]	9.98 (9.68, 10.27)[8.36, 11.53]
20	4	2	5	**10**	20	**0.56**	**30.99**	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
20	4	2	5	**12.5**	20	**0.67**	**37.32**	0.69 (0.60, 0.79)[0.38, 1.61]	37.28 (34.97, 39.85)[26.47, 49.62]	9.98 (9.66, 10.28)[8.38, 11.62]
20	4	2	5	**15**	20	**0.79**	**43.83**	0.81 (0.70, 0.94)[0.44, 2.02]	43.69 (41.05, 46.93)[30.41, 57.89]	9.97 (9.66, 10.29)[8.36, 11.67]
20	4	2	5	**17.5**	20	**0.91**	**50.45**	0.93 (0.80, 1.09)[0.50, 2.53]	50.31 (47.18, 54.01)[34.47, 66.31]	9.98 (9.66, 10.30)[8.35, 11.72]
20	4	2	5	**20**	20	**1.03**	**57.14**	1.06 (0.90, 1.25)[0.55, 3.23]	56.99 (53.40, 61.24)[38.60, 75.22]	9.97 (9.65, 10.33)[8.34, 11.77]
20	4	2	5	10	**10**	**1.12**	30.99	1.15 (0.99, 1.35)[0.61, 3.79]	30.92 (29.30, 32.81)[23.41, 38.12]	9.98 (9.82, 10.16)[9.18, 10.90]
20	4	2	5	10	**15**	**0.75**	30.99	0.76 (0.66, 0.88)[0.43, 1.81]	30.93 (29.16, 32.89)[23.06, 39.72]	9.98 (9.75, 10.22)[8.78, 11.24]
20	4	2	5	10	**20**	**0.56**	30.99	0.57 (0.50, 0.65)[0.33, 1.27]	30.99 (29.11, 32.98)[22.72, 41.58]	9.98 (9.67, 10.28)[8.39, 11.58]
20	4	2	5	10	**25**	**0.45**	30.99	0.45 (0.40, 0.52)[0.27, 0.99]	31.06 (29.05, 33.19)[22.39, 43.64]	9.97 (9.59, 10.35)[7.97, 11.91]
20	4	2	5	10	**30**	**0.37**	30.99	0.38 (0.33, 0.43)[0.23, 0.81]	31.08 (28.89, 33.39)[22.07, 45.90]	9.97 (9.51, 10.41)[7.55, 12.25]
20	4	2	5	10	**35**	**0.32**	30.99	0.32 (0.29, 0.37)[0.19, 0.69]	31.11 (28.77, 33.53)[21.75, 48.42]	9.97 (9.43, 10.48)[7.13, 12.59]
20	4	2	5	10	**40**	**0.28**	30.99	0.28 (0.25, 0.32)[0.17, 0.60]	31.14 (28.62, 33.76)[21.45, 51.22]	9.95 (9.35, 10.55)[6.71, 12.92]
20	4	2	5	10	**45**	**0.25**	30.99	0.25 (0.22, 0.29)[0.15, 0.53]	31.15 (28.46, 33.99)[21.16, 54.37]	9.95 (9.27, 10.61)[6.29, 13.26]
20	4	2	5	10	**50**	**0.22**	30.99	0.23 (0.20, 0.26)[0.14, 0.47]	31.16 (28.31, 34.21)[20.87, 57.94]	9.95 (9.19, 10.69)[5.86, 13.60]

C.4 Results of the log-normal data simulation; varying variability

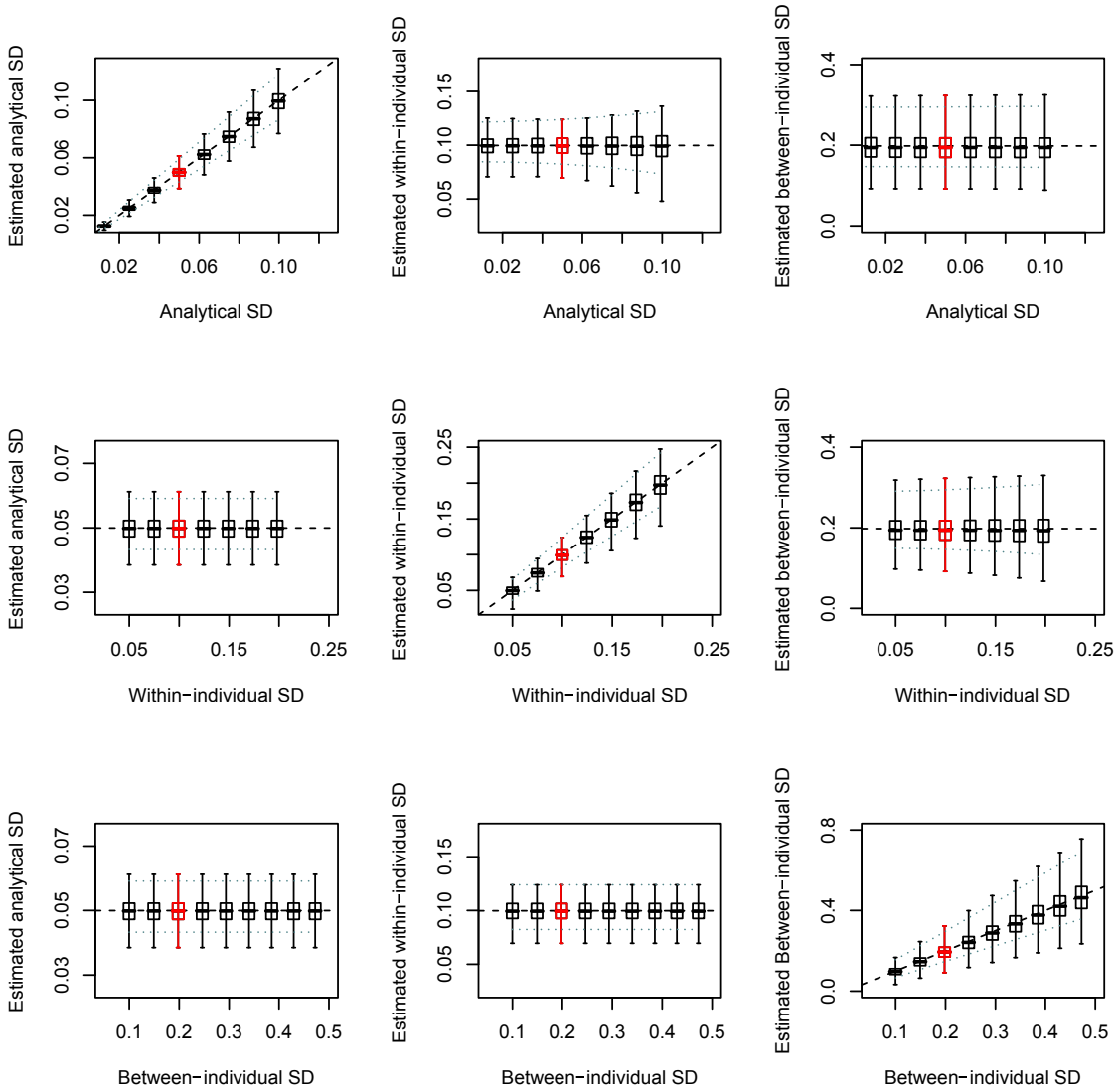


Figure C.4: Log-normal biological variability sample size simulation: SD estimates from biological variability data simulations varying test variability: SD_A (left column), SD_I (middle column) and SD_G (right column) estimates when varying value of SD_A (top row), value of SD_I (middle row), and value of SD_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true SD and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

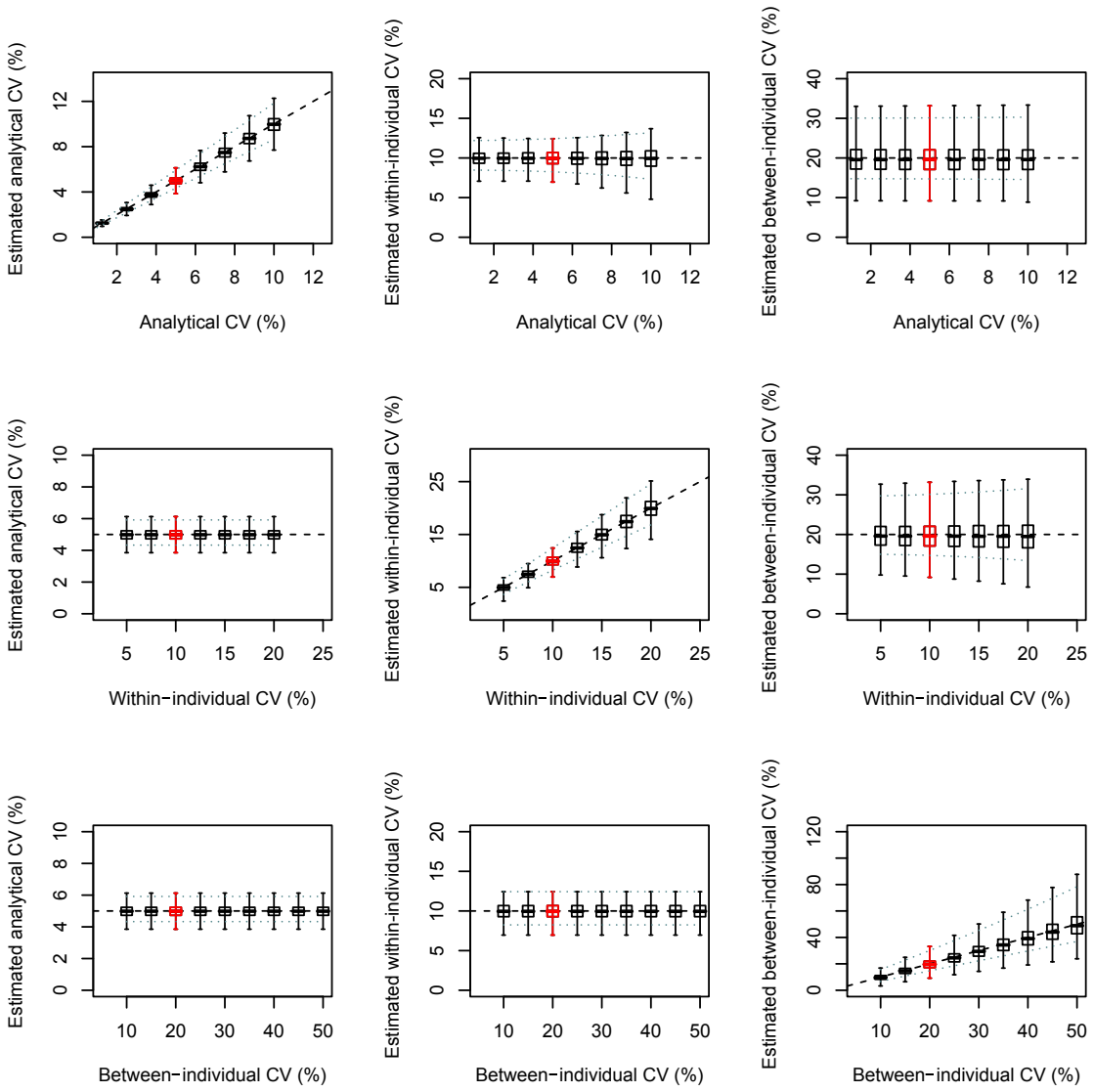


Figure C.5: Log-normal biological variability sample size simulation: CV estimates from biological variability data simulations varying test variability: CV_A (left column), CV_I (middle column) and CV_G (right column) estimates when varying value of CV_A (top row), value of CV_I (middle row), and value of CV_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true CV and the dotted lines are the 95% confidence intervals around the true value of the estimate for the given sample size, using the methods of Burdick and Graybill.[34]

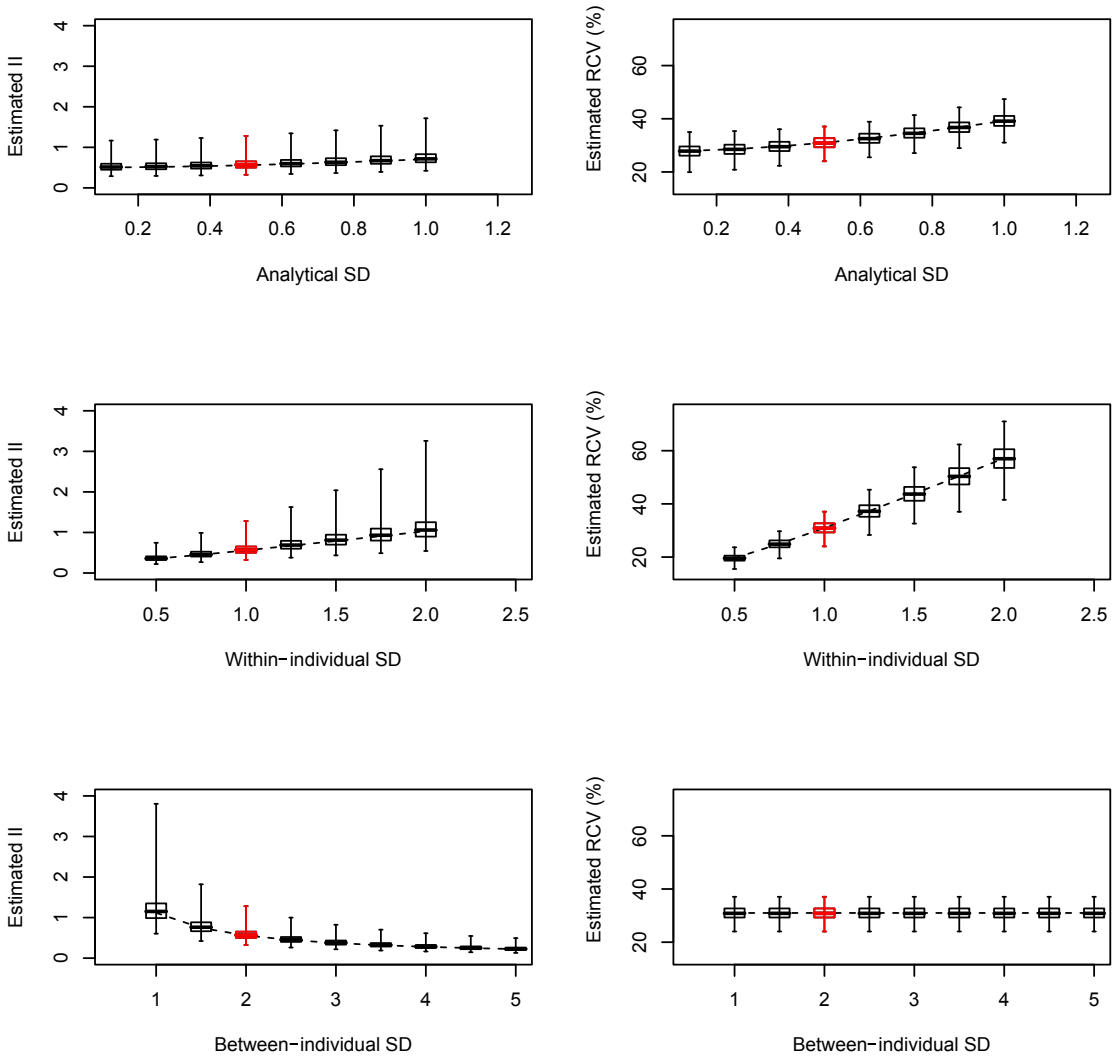


Figure C.6: Log-normal biological variability sample size simulation: II and RCV estimates from biological variability data simulations varying test variability: II (left column) and RCV (right column) estimates when varying value of CV_A (top row), value of CV_I (middle row), and value of CV_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true II or RCV.

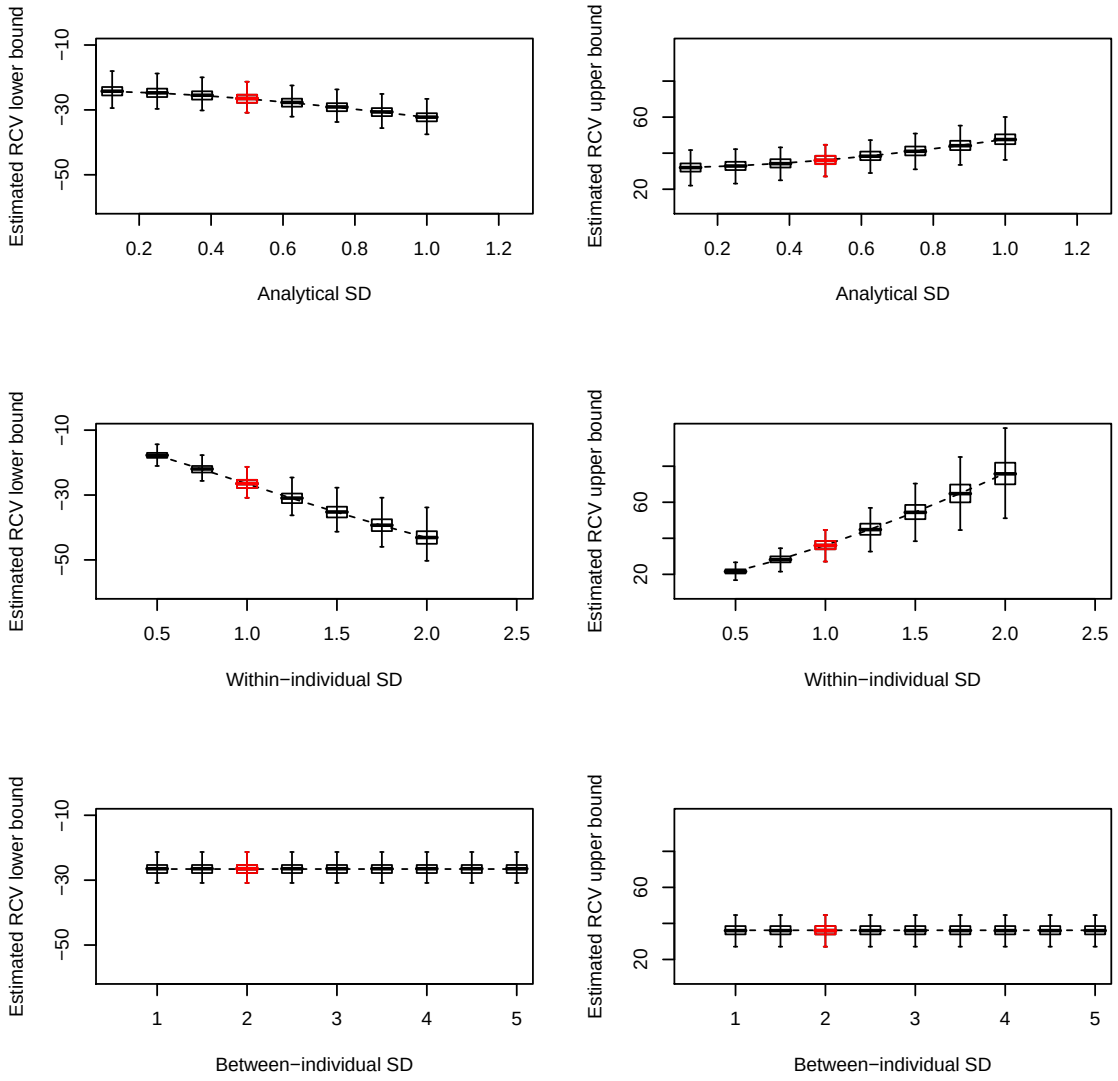


Figure C.7: Log-normal biological variability sample size simulation: asymmetric RCV estimates from biological variability data simulations varying test variability: RCV lower bound (left column) and RCV upper bound (right column) estimates when varying value of CV_A (top row), value of CV_I (middle row), and value of CV_G (bottom row). Median is shown by horizontal line, Q1 and Q3 shown by extremes of box; and minimum and maximum values shown by arrows. Estimates shown in red are for the baseline strategy. The dashed line reflects the true RCV bound.

Table C.13: Log normal simulation: bias performance measures varying CV_A, CV_I and CV_G.

						Bias								
			Inputs			**Bias ($\times 10^{-4}$)**			**Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.01**	0.10	0.20	-0.219	-2.789	-30.686	-0.175	-0.280	-1.549	-2.274	-3.092	-9.025
20	4	2	**0.02**	0.10	0.20	-0.437	-2.824	-30.506	-0.175	-0.283	-1.540	-2.271	-3.073	-8.961
20	4	2	**0.04**	0.10	0.20	-0.655	-2.989	-30.399	-0.175	-0.300	-1.535	-2.269	-3.139	-8.909
20	4	2	**0.05**	0.10	0.20	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
20	4	2	**0.06**	0.10	0.20	-1.092	-3.846	-30.401	-0.175	-0.386	-1.535	-2.270	-3.580	-8.838
20	4	2	**0.07**	0.10	0.20	-1.310	-4.638	-30.513	-0.175	-0.465	-1.541	-2.271	-3.979	-8.821
20	4	2	**0.09**	0.10	0.20	-1.527	-5.766	-30.701	-0.175	-0.578	-1.550	-2.270	-4.510	-8.817
20	4	2	**0.10**	0.10	0.20	-1.744	-7.329	-30.967	-0.175	-0.735	-1.564	-2.270	-5.179	-8.824
20	4	2	0.05	**0.05**	0.20	-0.874	-3.671	-25.223	-0.175	-0.735	-1.274	-2.270	-5.179	-7.667
20	4	2	0.05	**0.07**	0.20	-0.874	-3.060	-27.485	-0.175	-0.409	-1.388	-2.270	-3.698	-8.221
20	4	2	0.05	**0.10**	0.20	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
20	4	2	0.05	**0.12**	0.20	-0.873	-3.797	-33.912	-0.175	-0.305	-1.712	-2.269	-3.164	-9.598
20	4	2	0.05	**0.15**	0.20	-0.874	-4.365	-38.202	-0.175	-0.293	-1.929	-2.270	-3.107	-10.403
20	4	2	0.05	**0.17**	0.20	-0.874	-4.977	-43.335	-0.175	-0.287	-2.188	-2.270	-3.084	-11.278
20	4	2	0.05	**0.20**	0.20	-0.874	-5.612	-49.451	-0.175	-0.283	-2.497	-2.270	-3.075	-12.219
20	4	2	0.05	0.10	**0.10**	-0.874	-3.315	-25.238	-0.175	-0.332	-2.530	-2.270	-3.303	-12.133
20	4	2	0.05	0.10	**0.15**	-0.874	-3.315	-26.552	-0.175	-0.332	-1.780	-2.270	-3.303	-9.806
20	4	2	0.05	0.10	**0.20**	-0.874	-3.316	-30.364	-0.175	-0.332	-1.533	-2.269	-3.304	-8.868
20	4	2	0.05	0.10	**0.25**	-0.873	-3.316	-34.889	-0.175	-0.332	-1.417	-2.269	-3.304	-8.381
20	4	2	0.05	0.10	**0.29**	-0.873	-3.316	-39.677	-0.175	-0.332	-1.352	-2.269	-3.305	-8.090
20	4	2	0.05	0.10	**0.34**	-0.873	-3.317	-44.549	-0.175	-0.333	-1.311	-2.268	-3.305	-7.900
20	4	2	0.05	0.10	**0.39**	-0.873	-3.317	-49.417	-0.175	-0.333	-1.283	-2.268	-3.305	-7.767
20	4	2	0.05	0.10	**0.43**	-0.873	-3.317	-54.229	-0.175	-0.333	-1.263	-2.268	-3.305	-7.669
20	4	2	0.05	0.10	**0.47**	-0.873	-3.317	-58.956	-0.175	-0.333	-1.248	-2.269	-3.305	-7.596

Table C.14: Log normal simulation: accuracy and coverage performance measures varying CV_A, CV_I and CV_G.

						Accuracy and coverage								
			Inputs			**Mean squared error ($\times 10^{-4}$)**			**Coverage**			**Mean 95% CI width**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.01**	0.10	0.20	0.009	0.815	11.655	0.952	0.957	0.951	0.004	0.037	0.146
20	4	2	**0.02**	0.10	0.20	0.037	0.845	11.682	0.952	0.959	0.952	0.008	0.038	0.146
20	4	2	**0.04**	0.10	0.20	0.083	0.907	11.736	0.952	0.959	0.954	0.012	0.039	0.146
20	4	2	**0.05**	0.10	0.20	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
20	4	2	**0.06**	0.10	0.20	0.231	1.156	11.924	0.952	0.954	0.955	0.020	0.045	0.147
20	4	2	**0.07**	0.10	0.20	0.333	1.361	12.058	0.952	0.956	0.954	0.024	0.048	0.148
20	4	2	**0.09**	0.10	0.20	0.453	1.638	12.220	0.952	0.953	0.954	0.028	0.053	0.149
20	4	2	**0.10**	0.10	0.20	0.591	2.008	12.411	0.952	0.961[a]	0.954	0.032	0.059[a]	0.150
20	4	2	0.05	**0.05**	0.20	0.148	0.504	10.885	0.952	0.961[a]	0.949	0.016	0.030[a]	0.140
20	4	2	0.05	**0.07**	0.20	0.148	0.686	11.253	0.952	0.955	0.951	0.016	0.034	0.143
20	4	2	0.05	**0.10**	0.20	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
20	4	2	0.05	**0.12**	0.20	0.148	1.441	12.599	0.952	0.959	0.959	0.016	0.050	0.152
20	4	2	0.05	**0.15**	0.20	0.148	1.975	13.631	0.952	0.962	0.960	0.016	0.058	0.158
20	4	2	0.05	**0.17**	0.20	0.148	2.607	14.953	0.952	0.957	0.963[b]	0.016	0.067	0.164[b]
20	4	2	0.05	**0.20**	0.20	0.148	3.333	16.623	0.952	0.959	0.971[c]	0.016	0.075	0.176[c]
20	4	2	0.05	0.10	**0.10**	0.148	1.008	4.391	0.952	0.958	0.971[d]	0.016	0.042	0.091[d]
20	4	2	0.05	0.10	**0.15**	0.148	1.008	7.403	0.952	0.957	0.958	0.016	0.042	0.116
20	4	2	0.05	0.10	**0.20**	0.148	1.009	11.816	0.952	0.958	0.953	0.016	0.042	0.147
20	4	2	0.05	0.10	**0.25**	0.148	1.008	17.452	0.952	0.958	0.954	0.016	0.042	0.178
20	4	2	0.05	0.10	**0.29**	0.148	1.008	24.212	0.952	0.958	0.952	0.016	0.042	0.209
20	4	2	0.05	0.10	**0.34**	0.148	1.008	32.002	0.952	0.958	0.950	0.016	0.042	0.241
20	4	2	0.05	0.10	**0.39**	0.148	1.008	40.728	0.952	0.958	0.952	0.016	0.042	0.271
20	4	2	0.05	0.10	**0.43**	0.148	1.008	50.291	0.952	0.958	0.952	0.016	0.042	0.301
20	4	2	0.05	0.10	**0.47**	0.148	1.008	60.596	0.952	0.958	0.952	0.016	0.042	0.331

[a]8 CIs could not be calculated; [b]2 CIs could not be calculated; [c]10 CIs could not be calculated; [d]18 CIs could not be calculated.

Table C.15: Log normal simulation: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.01**	0.10	0.20	0.01 (0.01, 0.01)[0.01, 0.02]	0.10 (0.09, 0.11)[0.07, 0.13]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.02**	0.10	0.20	0.02 (0.02, 0.03)[0.02, 0.03]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.04**	0.10	0.20	0.04 (0.04, 0.04)[0.03, 0.05]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.05**	0.10	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.06**	0.10	0.20	0.06 (0.06, 0.07)[0.05, 0.08]	0.10 (0.09, 0.11)[0.07, 0.13]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.07**	0.10	0.20	0.07 (0.07, 0.08)[0.06, 0.09]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.09**	0.10	0.20	0.09 (0.08, 0.09)[0.07, 0.11]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	**0.10**	0.10	0.20	0.10 (0.09, 0.10)[0.08, 0.12]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	0.05	**0.05**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.05 (0.04, 0.05)[0.02, 0.07]	0.19 (0.17, 0.22)[0.10, 0.32]
20	4	2	0.05	**0.07**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.07 (0.07, 0.08)[0.05, 0.09]	0.19 (0.17, 0.22)[0.10, 0.32]
20	4	2	0.05	**0.10**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	0.05	**0.12**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.12 (0.12, 0.13)[0.09, 0.15]	0.19 (0.17, 0.22)[0.09, 0.33]
20	4	2	0.05	**0.15**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.15 (0.14, 0.16)[0.11, 0.19]	0.19 (0.17, 0.22)[0.08, 0.33]
20	4	2	0.05	**0.17**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.17 (0.16, 0.18)[0.12, 0.22]	0.19 (0.17, 0.22)[0.08, 0.33]
20	4	2	0.05	**0.20**	0.20	0.05 (0.05, 0.05)[0.04, 0.06]	0.20 (0.18, 0.21)[0.14, 0.25]	0.19 (0.17, 0.22)[0.07, 0.33]
20	4	2	0.05	0.10	**0.10**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.10 (0.08, 0.11)[0.03, 0.17]
20	4	2	0.05	0.10	**0.15**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.15 (0.13, 0.16)[0.06, 0.25]
20	4	2	0.05	0.10	**0.20**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.19 (0.17, 0.22)[0.09, 0.32]
20	4	2	0.05	0.10	**0.25**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.24 (0.21, 0.27)[0.12, 0.40]
20	4	2	0.05	0.10	**0.29**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.29 (0.25, 0.32)[0.14, 0.47]
20	4	2	0.05	0.10	**0.34**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.33 (0.29, 0.37)[0.17, 0.55]
20	4	2	0.05	0.10	**0.39**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.38 (0.33, 0.42)[0.19, 0.62]
20	4	2	0.05	0.10	**0.43**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.42 (0.37, 0.47)[0.21, 0.69]
20	4	2	0.05	0.10	**0.47**	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.07, 0.12]	0.46 (0.41, 0.52)[0.24, 0.76]

Table C.16: Log normal simulation: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
20	4	2	**1.25**	10	20	**0.50**	**27.93**	1.25 (1.18, 1.31)[0.96, 1.53]	9.96 (9.34, 10.59)[7.07, 12.56]	19.58 (17.21, 22.12)[9.22, 33.02]
20	4	2	**2.5**	10	20	**0.52**	**28.57**	2.49 (2.36, 2.62)[1.93, 3.06]	9.96 (9.33, 10.61)[7.07, 12.51]	19.59 (17.15, 22.09)[9.21, 33.08]
20	4	2	**3.75**	10	20	**0.53**	**29.60**	3.74 (3.55, 3.92)[2.89, 4.60]	9.95 (9.33, 10.65)[7.08, 12.46]	19.61 (17.11, 22.07)[9.20, 33.13]
20	4	2	**5**	10	20	**0.56**	**30.99**	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
20	4	2	**6.25**	10	20	**0.59**	**32.69**	6.23 (5.91, 6.54)[4.81, 7.66]	9.94 (9.22, 10.71)[6.72, 12.56]	19.59 (17.09, 22.11)[9.18, 33.22]
20	4	2	**7.5**	10	20	**0.63**	**34.65**	7.48 (7.09, 7.85)[5.78, 9.20]	9.94 (9.17, 10.76)[6.22, 12.84]	19.62 (17.09, 22.07)[9.17, 33.26]
20	4	2	**8.75**	10	20	**0.66**	**36.83**	8.73 (8.27, 9.16)[6.74, 10.73]	9.92 (9.08, 10.81)[5.59, 13.22]	19.59 (17.07, 22.04)[9.17, 33.31]
20	4	2	**10**	10	20	**0.71**	**39.20**	9.97 (9.45, 10.47)[7.70, 12.27]	9.94 (8.97, 10.90)[4.80, 13.69]	19.60 (17.10, 22.04)[8.86, 33.35]
20	4	2	5	**5**	20	**0.35**	**19.60**	4.99 (4.73, 5.23)[3.85, 6.13]	4.97 (4.48, 5.45)[2.40, 6.83]	19.60 (17.32, 22.02)[9.77, 32.69]
20	4	2	5	**7.5**	20	**0.45**	**24.99**	4.99 (4.73, 5.23)[3.85, 6.13]	7.46 (6.92, 8.04)[4.93, 9.48]	19.59 (17.21, 22.02)[9.52, 32.94]
20	4	2	5	**10**	20	**0.56**	**30.99**	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
20	4	2	5	**12.5**	20	**0.67**	**37.32**	4.99 (4.73, 5.23)[3.85, 6.13]	12.45 (11.66, 13.32)[8.85, 15.57]	19.59 (17.00, 22.15)[8.76, 33.39]
20	4	2	5	**15**	20	**0.79**	**43.83**	4.99 (4.73, 5.23)[3.85, 6.13]	14.94 (13.98, 15.97)[10.60, 18.74]	19.54 (16.87, 22.25)[8.23, 33.59]
20	4	2	5	**17.5**	20	**0.91**	**50.45**	4.99 (4.73, 5.23)[3.85, 6.13]	17.43 (16.33, 18.60)[12.34, 21.93]	19.54 (16.78, 22.26)[7.56, 33.77]
20	4	2	5	**20**	20	**1.03**	**57.14**	4.99 (4.73, 5.23)[3.85, 6.13]	19.92 (18.65, 21.25)[14.09, 25.12]	19.50 (16.64, 22.33)[6.74, 33.94]
20	4	2	5	10	**10**	**1.12**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	9.76 (8.28, 11.18)[3.30, 16.82]
20	4	2	5	10	**15**	**0.75**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	14.73 (12.72, 16.57)[6.47, 24.95]
20	4	2	5	10	**20**	**0.56**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	19.63 (17.10, 22.06)[9.19, 33.17]
20	4	2	5	10	**25**	**0.45**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	24.54 (21.50, 27.56)[11.79, 41.56]
20	4	2	5	10	**30**	**0.37**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	29.44 (25.82, 33.09)[14.31, 50.17]
20	4	2	5	10	**35**	**0.32**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	34.30 (30.10, 38.67)[16.77, 59.05]
20	4	2	5	10	**40**	**0.28**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	39.09 (34.38, 44.31)[19.18, 68.24]
20	4	2	5	10	**45**	**0.25**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	44.01 (38.64, 49.92)[21.53, 77.78]
20	4	2	5	10	**50**	**0.22**	30.99	4.99 (4.73, 5.23)[3.85, 6.13]	9.95 (9.28, 10.68)[6.96, 12.43]	48.88 (42.79, 55.60)[23.83, 87.73]

Table C.17: Log normal simulation: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
20	4	2	**1.25**	10	20	**0.50**	**27.93**	0.52 (0.45, 0.60)[0.29, 1.17]	27.83 (26.12, 29.57)[19.91, 35.00]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**2.5**	10	20	**0.52**	**28.57**	0.53 (0.46, 0.61)[0.30, 1.19]	28.49 (26.77, 30.25)[20.83, 35.36]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**3.75**	10	20	**0.53**	**29.60**	0.55 (0.48, 0.63)[0.31, 1.23]	29.49 (27.86, 31.32)[22.31, 36.06]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**5**	10	20	**0.56**	**30.99**	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**6.25**	10	20	**0.59**	**32.69**	0.60 (0.53, 0.69)[0.34, 1.34]	32.53 (30.94, 34.39)[25.50, 38.88]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**7.5**	10	20	**0.63**	**34.65**	0.63 (0.56, 0.73)[0.37, 1.42]	34.51 (32.87, 36.39)[27.10, 41.37]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**8.75**	10	20	**0.66**	**36.83**	0.67 (0.59, 0.78)[0.39, 1.53]	36.73 (35.00, 38.62)[28.95, 44.29]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	**10**	10	20	**0.71**	**39.20**	0.72 (0.63, 0.83)[0.42, 1.72]	39.10 (37.31, 41.06)[31.02, 47.38]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	5	**5**	20	**0.35**	**19.60**	0.36 (0.32, 0.41)[0.22, 0.75]	19.55 (18.66, 20.52)[15.52, 23.67]	9.98 (9.95, 10.00)[9.81, 10.12]
20	4	2	5	**7.5**	20	**0.45**	**24.99**	0.46 (0.41, 0.53)[0.27, 0.99]	24.86 (23.67, 26.26)[19.51, 29.71]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	5	**10**	20	**0.56**	**30.99**	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	5	**12.5**	20	**0.67**	**37.32**	0.69 (0.60, 0.79)[0.38, 1.63]	37.20 (35.15, 39.48)[28.31, 45.33]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	5	**15**	20	**0.79**	**43.83**	0.81 (0.70, 0.94)[0.43, 2.04]	43.69 (41.17, 46.38)[32.60, 53.77]	9.97 (9.93, 10.00)[9.81, 10.13]
20	4	2	5	**17.5**	20	**0.91**	**50.45**	0.93 (0.80, 1.10)[0.49, 2.56]	50.34 (47.30, 53.44)[37.02, 62.34]	9.96 (9.93, 9.99)[9.80, 10.13]
20	4	2	5	**20**	20	**1.03**	**57.14**	1.06 (0.90, 1.25)[0.54, 3.26]	56.98 (53.52, 60.55)[41.52, 71.00]	9.96 (9.92, 9.99)[9.79, 10.13]
20	4	2	5	10	**10**	**1.12**	30.99	1.15 (0.99, 1.35)[0.60, 3.81]	30.85 (29.28, 32.69)[24.02, 37.07]	9.99 (9.97, 10.00)[9.91, 10.08]
20	4	2	5	10	**15**	**0.75**	30.99	0.76 (0.66, 0.88)[0.42, 1.82]	30.85 (29.28, 32.69)[24.02, 37.07]	9.98 (9.96, 10.00)[9.86, 10.11]
20	4	2	5	10	**20**	**0.56**	30.99	0.57 (0.50, 0.65)[0.32, 1.28]	30.85 (29.28, 32.69)[24.02, 37.07]	9.97 (9.94, 10.00)[9.81, 10.13]
20	4	2	5	10	**25**	**0.45**	30.99	0.45 (0.40, 0.52)[0.26, 1.00]	30.85 (29.28, 32.69)[24.02, 37.07]	9.96 (9.92, 10.00)[9.76, 10.15]
20	4	2	5	10	**30**	**0.37**	30.99	0.38 (0.33, 0.43)[0.22, 0.82]	30.85 (29.28, 32.69)[24.02, 37.07]	9.95 (9.90, 9.99)[9.71, 10.17]
20	4	2	5	10	**35**	**0.32**	30.99	0.32 (0.29, 0.37)[0.19, 0.70]	30.85 (29.28, 32.69)[24.02, 37.07]	9.93 (9.88, 9.98)[9.66, 10.19]
20	4	2	5	10	**40**	**0.28**	30.99	0.28 (0.25, 0.33)[0.16, 0.61]	30.85 (29.28, 32.69)[24.02, 37.07]	9.92 (9.86, 9.97)[9.60, 10.20]
20	4	2	5	10	**45**	**0.25**	30.99	0.25 (0.22, 0.29)[0.15, 0.55]	30.85 (29.28, 32.69)[24.02, 37.07]	9.90 (9.83, 9.96)[9.55, 10.21]
20	4	2	5	10	**50**	**0.22**	30.99	0.23 (0.20, 0.26)[0.13, 0.49]	30.85 (29.28, 32.69)[24.02, 37.07]	9.88 (9.81, 9.95)[9.49, 10.22]

Table C.18: Log normal simulation: asymmetric RCV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**	
n_1	n_2	n_3	CV_A	CV_I	CV_G	$RCV-$	$RCV+$	RCV lower bound	RCV upper bound
20	4	2	**1.25**	10	20	**-24.32**	**32.13**	-24.24 (-25.54, -22.94)[-29.43, -18.03]	32.00 (29.77, 34.30)[22.00, 41.71]
20	4	2	**2.5**	10	20	**-24.80**	**32.97**	-24.74 (-26.03, -23.44)[-29.69, -18.78]	32.87 (30.62, 35.20)[23.12, 42.22]
20	4	2	**3.75**	10	20	**-25.56**	**34.34**	-25.48 (-26.82, -24.26)[-30.17, -19.97]	34.19 (32.03, 36.65)[24.95, 43.20]
20	4	2	**5**	10	20	**-26.58**	**36.20**	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	**6.25**	10	20	**-27.80**	**38.51**	-27.69 (-29.01, -26.54)[-32.08, -22.47]	38.30 (36.13, 40.86)[28.98, 47.23]
20	4	2	**7.5**	10	20	**-29.19**	**41.22**	-29.09 (-30.40, -27.93)[-33.73, -23.69]	41.03 (38.76, 43.67)[31.04, 50.89]
20	4	2	**8.75**	10	20	**-30.70**	**44.30**	-30.63 (-31.91, -29.43)[-35.60, -25.08]	44.15 (41.71, 46.86)[33.47, 55.28]
20	4	2	**10**	10	20	**-32.30**	**47.71**	-32.23 (-33.53, -31.03)[-37.52, -26.60]	47.56 (44.99, 50.44)[36.24, 60.06]
20	4	2	5	**5**	20	**-17.78**	**21.62**	-17.74 (-18.53, -17.00)[-21.04, -14.37]	21.56 (20.49, 22.75)[16.78, 26.65]
20	4	2	5	**7.5**	20	**-22.07**	**28.32**	-21.98 (-23.05, -21.04)[-25.64, -17.70]	28.17 (26.65, 29.96)[21.51, 34.48]
20	4	2	5	**10**	20	**-26.58**	**36.20**	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	**12.5**	20	**-31.03**	**44.99**	-30.95 (-32.49, -29.54)[-36.25, -24.60]	44.82 (41.92, 48.12)[32.63, 56.87]
20	4	2	5	**15**	20	**-35.31**	**54.58**	-35.22 (-36.91, -33.60)[-41.30, -27.74]	54.38 (50.60, 58.50)[38.38, 70.36]
20	4	2	5	**17.5**	20	**-39.37**	**64.94**	-39.31 (-41.11, -37.47)[-45.97, -30.83]	64.76 (59.94, 69.81)[44.56, 85.10]
20	4	2	5	**20**	20	**-43.19**	**76.03**	-43.10 (-45.03, -41.16)[-50.28, -33.83]	75.76 (69.94, 81.93)[51.12, 101.13]
20	4	2	5	10	**10**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**15**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**20**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**25**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**30**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**35**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**40**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**45**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]
20	4	2	5	10	**50**	-26.58	36.20	-26.48 (-27.81, -25.32)[-30.86, -21.32]	36.01 (33.91, 38.51)[27.10, 44.63]

C.5 Sensitivity analyses

Table C.19: Increased base n_1, n_2 and n_3: bias performance measures varying number of participants, observations and assessments.

Inputs						**Bias** ($\times 10^{-4}$)			**Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	10	3	0.5	1	2	-5.588	-12.005	-920.301	-0.112	-0.120	-4.602	-1.587	-1.071	-12.934
10	10	3	0.5	1	2	-4.665	-25.769	-481.143	-0.093	-0.258	-2.406	-1.902	-3.169	-9.901
20	10	3	0.5	1	2	-2.441	-17.728	-199.656	-0.049	-0.177	-0.998	-1.388	-3.105	-5.808
30	10	3	0.5	1	2	-2.279	3.671	-45.753	-0.046	0.037	-0.229	-1.670	0.785	-1.692
40	10	3	0.5	1	2	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
60	10	3	0.5	1	2	0.784	2.038	-138.884	0.016	0.020	-0.694	0.774	0.623	-7.622
100	10	3	0.5	1	2	2.746	6.136	-55.357	0.055	0.061	-0.277	3.480	2.440	-3.855
40	**2**	3	0.5	1	2	-7.830	-19.476	-36.351	-0.157	-0.195	-0.182	-2.802	-1.649	-1.391
40	**4**	3	0.5	1	2	-1.400	-10.807	-89.394	-0.028	-0.108	-0.447	-0.726	-1.543	-3.766
40	**6**	3	0.5	1	2	-3.844	1.572	-104.411	-0.077	0.016	-0.522	-2.331	0.293	-4.436
40	**8**	3	0.5	1	2	-3.145	35.565	-92.752	-0.063	0.356	-0.464	-2.357	7.691	-3.912
40	**10**	3	0.5	1	2	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
40	**12**	3	0.5	1	2	-1.961	12.203	-145.731	-0.039	0.122	-0.729	-1.723	3.430	-6.361
40	**20**	3	0.5	1	2	2.967	3.636	-177.572	0.059	0.036	-0.888	3.368	1.336	-8.030
40	10	**2**	0.5	1	2	7.405	-10.869	-120.926	0.148	-0.109	-0.605	4.155	-2.578	-5.369
40	10	**3**	0.5	1	2	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
40	10	**4**	0.5	1	2	1.421	0.418	-134.866	0.028	0.004	-0.674	1.378	0.105	-5.714
40	10	**6**	0.5	1	2	3.511	-19.392	-88.779	0.070	-0.194	-0.444	4.365	-4.908	-3.861
40	10	**10**	0.5	1	2	3.495	-11.608	-81.574	0.070	-0.116	-0.408	5.925	-2.959	-3.512

Table C.20: Increased base n_1, n_2 and n_3: accuracy and coverage performance measures varying number of participants, observations and assessments.

Inputs						Accuracy and coverage								
						Mean squared error ($\times 10^{-4}$)			Coverage			Mean 95% CI width		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	10	3	0.5	1	2	12.400	125.689	5147.336	0.950	0.958	0.964[a]	0.141	0.466	4.494[a]
10	10	3	0.5	1	2	6.020	66.185	2384.739	0.960	0.946	0.941	0.099	0.323	2.282
20	10	3	0.5	1	2	3.095	32.625	1185.815	0.943	0.951	0.938	0.070	0.226	1.424
30	10	3	0.5	1	2	1.862	21.897	731.023	0.967	0.951	0.950	0.057	0.184	1.123
40	10	3	0.5	1	2	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
60	10	3	0.5	1	2	1.026	10.696	333.975	0.951	0.941	0.958	0.040	0.130	0.759
100	10	3	0.5	1	2	0.623	6.328	206.503	0.950	0.955	0.952	0.031	0.100	0.581
40	**2**	3	0.5	1	2	7.813	139.575	683.199	0.950	0.949	0.950	0.111	0.496	1.068
40	**4**	3	0.5	1	2	3.722	49.046	564.187	0.951	0.944	0.954	0.078	0.278	0.989
40	**6**	3	0.5	1	2	2.721	28.757	555.095	0.946	0.956	0.948	0.063	0.215	0.967
40	**8**	3	0.5	1	2	1.781	21.512	562.881	0.962	0.948	0.946	0.055	0.181	0.957
40	**10**	3	0.5	1	2	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
40	**12**	3	0.5	1	2	1.296	12.671	526.946	0.958	0.959	0.957	0.045	0.144	0.944
40	**20**	3	0.5	1	2	0.777	7.411	492.217	0.950	0.949	0.957	0.035	0.109	0.934
40	10	**2**	0.5	1	2	3.182	17.793	508.794	0.952	0.948	0.964	0.070	0.166	0.950
40	10	**3**	0.5	1	2	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
40	10	**4**	0.5	1	2	1.063	15.957	558.942	0.943	0.941	0.946	0.040	0.156	0.948
40	10	**6**	0.5	1	2	0.648	15.648	529.377	0.944	0.947	0.956	0.031	0.153	0.950
40	10	**10**	0.5	1	2	0.349	15.399	540.239	0.950	0.944	0.955	0.023	0.150	0.950

[a]9 CIs could not be calculated.

Table C.21: Increased base n_1, n_2 and n_3: bias performance measures varying CV_A, CV_I and CV_G.

			Inputs			**Bias** ($\times 10^4$)			**Bias Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
40	10	3	**0.125**	1	2	0.789	9.355	-40.326	0.063	0.094	-0.202	2.497	2.476	-1.760
40	10	3	**0.25**	1	2	1.578	8.397	-37.901	0.063	0.084	-0.190	2.497	2.189	-1.653
40	10	3	**0.375**	1	2	2.367	7.109	-35.555	0.063	0.071	-0.178	2.497	1.807	-1.550
40	10	3	**0.5**	1	2	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
40	10	3	**0.625**	1	2	3.945	3.464	-31.102	0.063	0.035	-0.156	2.497	0.813	-1.352
40	10	3	**0.75**	1	2	4.734	1.057	-28.995	0.063	0.011	-0.145	2.497	0.235	-1.258
40	10	3	**0.875**	1	2	5.521	-1.785	-26.961	0.063	-0.018	-0.135	2.496	-0.374	-1.167
40	10	3	**1**	1	2	6.310	-5.123	-24.962	0.063	-0.051	-0.125	2.497	-1.005	-1.078
40	10	3	0.5	**0.5**	2	3.155	-2.562	-34.241	0.063	-0.051	-0.171	2.497	-1.005	-1.524
40	10	3	0.5	**0.75**	2	3.156	2.031	-33.408	0.063	0.027	-0.167	2.497	0.625	-1.473
40	10	3	0.5	**1**	2	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
40	10	3	0.5	**1.25**	2	3.156	8.512	-33.994	0.063	0.068	-0.170	2.497	1.720	-1.456
40	10	3	0.5	**1.5**	2	3.156	11.364	-35.576	0.063	0.076	-0.178	2.497	1.944	-1.496
40	10	3	0.5	**1.75**	2	3.156	14.111	-38.097	0.063	0.081	-0.190	2.497	2.089	-1.568
40	10	3	0.5	**2**	2	3.156	16.795	-41.632	0.063	0.084	-0.208	2.497	2.189	-1.672
40	10	3	0.5	1	**1**	3.156	5.472	-16.524	0.063	0.055	-0.165	2.497	1.342	-1.317
40	10	3	0.5	1	**1.5**	3.156	5.472	-23.607	0.063	0.055	-0.157	2.497	1.342	-1.338
40	10	3	0.5	1	**2**	3.156	5.472	-33.289	0.063	0.055	-0.166	2.497	1.342	-1.449
40	10	3	0.5	1	**2.5**	3.156	5.472	-43.885	0.063	0.055	-0.176	2.497	1.342	-1.546
40	10	3	0.5	1	**3**	3.156	5.472	-54.913	0.063	0.055	-0.183	2.497	1.342	-1.622
40	10	3	0.5	1	**3.5**	3.156	5.472	-66.181	0.063	0.055	-0.189	2.497	1.342	-1.683
40	10	3	0.5	1	**4**	3.156	5.472	-77.595	0.063	0.055	-0.194	2.497	1.342	-1.731
40	10	3	0.5	1	**4.5**	3.156	5.472	-89.106	0.063	0.055	-0.198	2.497	1.342	-1.770
40	10	3	0.5	1	**5**	3.156	5.472	-100.683	0.063	0.055	-0.201	2.497	1.342	-1.802

Table C.22: Increased base n_1, n_2 and n_3: accuracy and coverage performance measure varying CV_A, CV_I and CV_G.

						Accuracy and coverage								
			Inputs			Mean squared error ($\times 10^{-4}$)			Coverage			Mean 95% CI width		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
40	10	3	**0.125**	1	2	0.100	14.289	524.883	0.958	0.948	0.951	0.012	0.148	0.951
40	10	3	**0.25**	1	2	0.399	14.717	525.613	0.958	0.951	0.951	0.025	0.150	0.952
40	10	3	**0.375**	1	2	0.899	15.485	526.640	0.958	0.951	0.952	0.037	0.154	0.952
40	10	3	**0.5**	1	2	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
40	10	3	**0.625**	1	2	2.497	18.167	529.598	0.958	0.954	0.952	0.061	0.167	0.954
40	10	3	**0.75**	1	2	3.596	20.217	531.570	0.958	0.955	0.953	0.074	0.175	0.956
40	10	3	**0.875**	1	2	4.895	22.787	533.778	0.958	0.956	0.953	0.086	0.186	0.958
40	10	3	**1**	1	2	6.392	26.012	536.361	0.958	0.952	0.950	0.098	0.199	0.959
40	10	3	0.5	**0.5**	2	1.598	6.503	504.777	0.958	0.952	0.955	0.049	0.099	0.936
40	10	3	0.5	**0.75**	2	1.598	10.571	514.623	0.958	0.953	0.954	0.049	0.127	0.943
40	10	3	0.5	**1**	2	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
40	10	3	0.5	**1.25**	2	1.598	24.498	544.933	0.958	0.952	0.952	0.049	0.194	0.966
40	10	3	0.5	**1.5**	2	1.598	34.176	565.762	0.958	0.951	0.951	0.049	0.229	0.982
40	10	3	0.5	**1.75**	2	1.598	45.632	590.694	0.958	0.951	0.954	0.049	0.264	1.001
40	10	3	0.5	**2**	2	1.598	58.873	620.049	0.958	0.951	0.958	0.049	0.300	1.023
40	10	3	0.5	1	**1**	1.598	16.619	157.547	0.958	0.951	0.958	0.049	0.159	0.515
40	10	3	0.5	1	**1.5**	1.598	16.618	311.529	0.958	0.951	0.952	0.049	0.159	0.730
40	10	3	0.5	1	**2**	1.598	16.619	527.984	0.958	0.951	0.952	0.049	0.159	0.953
40	10	3	0.5	1	**2.5**	1.598	16.619	806.191	0.958	0.951	0.952	0.049	0.159	1.180
40	10	3	0.5	1	**3**	1.598	16.619	1146.075	0.958	0.951	0.954	0.049	0.159	1.409
40	10	3	0.5	1	**3.5**	1.598	16.619	1547.576	0.958	0.951	0.955	0.049	0.159	1.638
40	10	3	0.5	1	**4**	1.598	16.619	2010.697	0.958	0.951	0.954	0.049	0.159	1.868
40	10	3	0.5	1	**4.5**	1.598	16.619	2535.358	0.958	0.951	0.956	0.049	0.159	2.099
40	10	3	0.5	1	**5**	1.598	16.619	3121.509	0.958	0.951	0.955	0.049	0.159	2.330

Table C.23: Increased base n_1, n_2 and n_3: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	10	3	0.5	1	2	0.50 (0.47, 0.52)[0.40, 0.63]	1.00 (0.93, 1.07)[0.66, 1.44]	1.84 (1.38, 2.38)[0.00, 4.55]
10	10	3	0.5	1	2	0.50 (0.48, 0.52)[0.43, 0.58]	1.00 (0.94, 1.05)[0.70, 1.26]	1.94 (1.63, 2.27)[0.69, 3.70]
20	10	3	0.5	1	2	0.50 (0.49, 0.51)[0.45, 0.55]	1.00 (0.96, 1.04)[0.79, 1.17]	1.96 (1.75, 2.20)[0.91, 3.23]
30	10	3	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.83, 1.15]	1.99 (1.81, 2.17)[1.18, 3.08]
40	10	3	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
60	10	3	0.5	1	2	0.50 (0.49, 0.51)[0.47, 0.53]	1.00 (0.98, 1.02)[0.88, 1.12]	1.99 (1.86, 2.11)[1.46, 2.55]
100	10	3	0.5	1	2	0.50 (0.49, 0.51)[0.48, 0.52]	1.00 (0.98, 1.02)[0.92, 1.08]	1.99 (1.90, 2.09)[1.58, 2.49]
40	**2**	3	0.5	1	2	0.50 (0.48, 0.52)[0.42, 0.60]	1.00 (0.92, 1.08)[0.61, 1.37]	1.99 (1.84, 2.17)[1.01, 2.78]
40	**4**	3	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.56]	1.00 (0.95, 1.05)[0.79, 1.21]	1.98 (1.83, 2.16)[1.35, 2.71]
40	**6**	3	0.5	1	2	0.50 (0.49, 0.51)[0.45, 0.57]	1.00 (0.96, 1.04)[0.82, 1.14]	1.99 (1.83, 2.14)[1.24, 2.85]
40	**8**	3	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.04)[0.86, 1.17]	1.99 (1.83, 2.15)[1.27, 2.80]
40	**10**	3	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
40	**12**	3	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.98, 1.03)[0.90, 1.10]	1.99 (1.83, 2.14)[1.25, 2.71]
40	**20**	3	0.5	1	2	0.50 (0.49, 0.51)[0.47, 0.53]	1.00 (0.98, 1.02)[0.91, 1.09]	1.98 (1.83, 2.12)[1.27, 2.68]
40	10	**2**	0.5	1	2	0.50 (0.49, 0.51)[0.44, 0.56]	1.00 (0.97, 1.03)[0.86, 1.14]	1.99 (1.83, 2.15)[1.29, 2.76]
40	10	**3**	0.5	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	**4**	0.5	1	2	0.50 (0.49, 0.51)[0.47, 0.53]	1.00 (0.97, 1.03)[0.89, 1.12]	1.99 (1.83, 2.14)[1.32, 3.00]
40	10	**6**	0.5	1	2	0.50 (0.49, 0.51)[0.48, 0.53]	1.00 (0.97, 1.02)[0.88, 1.13]	2.00 (1.84, 2.16)[1.35, 2.71]
40	10	**10**	0.5	1	2	0.50 (0.50, 0.50)[0.48, 0.52]	1.00 (0.97, 1.02)[0.89, 1.13]	1.98 (1.83, 2.15)[1.12, 2.69]

Table C.24: Increased base n_1, n_2 and n_3: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
5	10	3	5	10	20	0.56	30.99	5.03 (4.62, 5.41)[3.65, 7.43]	10.02 (9.05, 10.97)[5.82, 17.69]	18.58 (13.83, 24.04)[0.00, 45.52]
10	10	3	5	10	20	0.56	30.99	5.01 (4.74, 5.29)[3.92, 6.91]	10.03 (9.33, 10.77)[6.90, 14.65]	19.48 (16.24, 22.80)[6.71, 39.75]
20	10	3	5	10	20	0.56	30.99	4.99 (4.81, 5.20)[4.19, 6.13]	9.97 (9.50, 10.45)[7.75, 12.68]	19.71 (17.48, 22.03)[9.02, 33.96]
30	10	3	5	10	20	0.56	30.99	5.00 (4.84, 5.15)[4.41, 5.97]	10.01 (9.57, 10.41)[8.06, 12.14]	19.86 (18.08, 21.79)[11.13, 30.42]
40	10	3	5	10	20	0.56	30.99	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
60	10	3	5	10	20	0.56	30.99	5.00 (4.89, 5.10)[4.57, 5.72]	10.00 (9.73, 10.27)[8.69, 11.44]	19.87 (18.47, 21.20)[14.08, 26.01]
100	10	3	5	10	20	0.56	30.99	5.00 (4.91, 5.08)[4.56, 5.43]	10.00 (9.77, 10.22)[9.00, 11.05]	19.93 (18.92, 20.89)[15.65, 24.71]
40	**2**	3	5	10	20	0.56	30.99	4.99 (4.79, 5.20)[3.86, 6.07]	10.05 (9.20, 10.82)[5.88, 13.77]	19.96 (18.23, 21.73)[9.97, 28.43]
40	**4**	3	5	10	20	0.56	30.99	4.99 (4.83, 5.17)[4.18, 5.89]	9.97 (9.46, 10.50)[7.73, 13.51]	19.85 (18.17, 21.62)[13.35, 27.73]
40	**6**	3	5	10	20	0.56	30.99	4.99 (4.84, 5.15)[4.30, 5.75]	10.00 (9.54, 10.44)[7.99, 12.20]	19.86 (18.25, 21.50)[12.50, 29.45]
40	**8**	3	5	10	20	0.56	30.99	4.99 (4.85, 5.12)[4.37, 5.75]	10.01 (9.63, 10.39)[8.26, 11.92]	19.93 (18.19, 21.53)[12.50, 29.09]
40	**10**	3	5	10	20	0.56	30.99	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
40	**12**	3	5	10	20	0.56	30.99	5.00 (4.87, 5.13)[4.41, 5.75]	10.01 (9.68, 10.34)[8.54, 11.60]	19.89 (18.17, 21.43)[12.01, 28.41]
40	**20**	3	5	10	20	0.56	30.99	5.00 (4.88, 5.12)[4.37, 5.68]	9.99 (9.71, 10.29)[8.71, 11.74]	19.73 (18.23, 21.29)[13.12, 26.47]
40	10	**2**	5	10	20	0.56	30.99	5.00 (4.85, 5.18)[4.39, 5.75]	10.01 (9.62, 10.38)[8.43, 11.88]	19.91 (18.23, 21.41)[12.77, 28.87]
40	10	**3**	5	10	20	0.56	30.99	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
40	10	**4**	5	10	20	0.56	30.99	4.99 (4.87, 5.13)[4.53, 5.58]	10.00 (9.65, 10.35)[8.30, 12.10]	19.88 (18.21, 21.53)[13.15, 29.42]
40	10	**6**	5	10	20	0.56	30.99	5.01 (4.89, 5.13)[4.48, 5.64]	9.98 (9.67, 10.34)[8.47, 11.47]	19.93 (18.34, 21.59)[13.01, 28.22]
40	10	**10**	5	10	20	0.56	30.99	5.00 (4.90, 5.12)[4.54, 5.61]	9.97 (9.68, 10.33)[8.60, 11.87]	19.88 (18.26, 21.48)[11.79, 27.68]

Table C.25: Increased base n_1, n_2 and n_3: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
5	10	3	5	10	20	0.56	30.99	0.60 (0.46, 0.82)[0.24, 13732.67]	31.00 (28.45, 33.73)[19.41, 52.09]	9.99 (9.40, 10.63)[7.26, 12.85]
10	10	3	5	10	20	0.56	30.99	0.57 (0.49, 0.69)[0.29, 1.75]	31.03 (29.17, 33.11)[23.48, 43.54]	9.97 (9.53, 10.41)[7.89, 11.96]
20	10	3	5	10	20	0.56	30.99	0.57 (0.50, 0.64)[0.34, 1.20]	30.94 (29.68, 32.27)[24.95, 38.48]	10.01 (9.73, 10.29)[8.27, 11.34]
30	10	3	5	10	20	0.56	30.99	0.56 (0.51, 0.62)[0.37, 0.93]	31.03 (29.81, 32.15)[25.84, 37.29]	10.00 (9.75, 10.27)[8.89, 11.13]
40	10	3	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
60	10	3	5	10	20	0.56	30.99	0.56 (0.53, 0.60)[0.42, 0.80]	30.97 (30.20, 31.72)[27.51, 35.30]	10.01 (9.85, 10.18)[9.17, 10.89]
100	10	3	5	10	20	0.56	30.99	0.56 (0.54, 0.59)[0.44, 0.72]	30.98 (30.35, 31.59)[28.36, 33.88]	10.01 (9.86, 10.16)[9.25, 10.68]
40	**2**	3	5	10	20	0.56	30.99	0.56 (0.50, 0.63)[0.36, 1.18]	31.13 (29.06, 33.13)[21.32, 40.57]	9.99 (9.77, 10.22)[8.96, 11.18]
40	**4**	3	5	10	20	0.56	30.99	0.56 (0.51, 0.62)[0.40, 0.85]	30.92 (29.54, 32.28)[25.12, 40.38]	10.00 (9.78, 10.25)[8.95, 11.03]
40	**6**	3	5	10	20	0.56	30.99	0.56 (0.52, 0.62)[0.39, 0.91]	31.02 (29.85, 32.14)[25.86, 37.03]	10.01 (9.78, 10.23)[9.01, 10.91]
40	**8**	3	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.39, 0.86]	31.01 (30.01, 31.99)[26.18, 36.23]	10.02 (9.80, 10.24)[8.94, 10.96]
40	**10**	3	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
40	**12**	3	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.41, 0.91]	31.02 (30.16, 31.93)[27.08, 35.90]	10.00 (9.79, 10.22)[8.89, 10.96]
40	**20**	3	5	10	20	0.56	30.99	0.56 (0.53, 0.61)[0.42, 0.88]	30.93 (30.22, 31.84)[27.27, 35.74]	10.02 (9.79, 10.23)[8.88, 11.08]
40	10	**2**	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.40, 0.86]	31.04 (30.00, 31.99)[26.75, 36.44]	10.00 (9.79, 10.22)[9.05, 11.19]
40	10	**3**	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
40	10	**4**	5	10	20	0.56	30.99	0.56 (0.52, 0.62)[0.39, 0.84]	30.94 (30.04, 31.96)[26.48, 36.78]	10.01 (9.79, 10.24)[9.05, 10.83]
40	10	**6**	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.42, 0.85]	30.93 (30.10, 31.93)[26.72, 35.28]	10.00 (9.79, 10.20)[9.01, 11.06]
40	10	**10**	5	10	20	0.56	30.99	0.56 (0.52, 0.61)[0.41, 0.97]	30.94 (30.13, 31.88)[27.08, 35.94]	10.00 (9.79, 10.21)[8.92, 11.20]

Table C.26: Increased base n_1, n_2 and n_3: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
40	10	3	**0.125**	1	2	0.13 (0.12, 0.13)[0.11, 0.14]	1.00 (0.98, 1.03)[0.87, 1.13]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	**0.25**	1	2	0.25 (0.25, 0.25)[0.23, 0.27]	1.00 (0.97, 1.03)[0.87, 1.13]	1.99 (1.83, 2.15)[1.35, 2.70]
40	10	3	**0.375**	1	2	0.38 (0.37, 0.38)[0.34, 0.41]	1.00 (0.97, 1.03)[0.87, 1.14]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	**0.5**	1	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	**0.625**	1	2	0.63 (0.62, 0.64)[0.57, 0.68]	1.00 (0.97, 1.03)[0.86, 1.16]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	**0.75**	1	2	0.75 (0.74, 0.76)[0.68, 0.82]	1.00 (0.97, 1.03)[0.86, 1.17]	1.98 (1.83, 2.16)[1.35, 2.70]
40	10	3	**0.875**	1	2	0.88 (0.86, 0.89)[0.80, 0.95]	1.00 (0.97, 1.03)[0.86, 1.18]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	**1**	1	2	1.00 (0.98, 1.02)[0.91, 1.09]	1.00 (0.97, 1.04)[0.85, 1.20]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	0.5	**0.5**	2	0.50 (0.49, 0.51)[0.46, 0.54]	0.50 (0.48, 0.52)[0.42, 0.60]	1.99 (1.83, 2.14)[1.37, 2.68]
40	10	3	0.5	**0.75**	2	0.50 (0.49, 0.51)[0.46, 0.54]	0.75 (0.73, 0.77)[0.65, 0.87]	1.99 (1.83, 2.15)[1.36, 2.69]
40	10	3	0.5	**1**	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	0.5	**1.25**	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.25 (1.22, 1.29)[1.08, 1.42]	1.99 (1.83, 2.16)[1.34, 2.71]
40	10	3	0.5	**1.5**	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.50 (1.46, 1.54)[1.30, 1.70]	1.99 (1.83, 2.17)[1.32, 2.72]
40	10	3	0.5	**1.75**	2	0.50 (0.49, 0.51)[0.46, 0.54]	1.75 (1.71, 1.80)[1.52, 1.97]	1.98 (1.83, 2.17)[1.30, 2.73]
40	10	3	0.5	**2**	2	0.50 (0.49, 0.51)[0.46, 0.54]	2.00 (1.95, 2.05)[1.73, 2.25]	1.99 (1.82, 2.17)[1.27, 2.74]
40	10	3	0.5	1	**1**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	0.99 (0.91, 1.09)[0.63, 1.37]
40	10	3	0.5	1	**1.5**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.49 (1.37, 1.62)[1.00, 2.04]
40	10	3	0.5	1	**2**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	1.99 (1.83, 2.16)[1.35, 2.70]
40	10	3	0.5	1	**2.5**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	2.48 (2.29, 2.69)[1.70, 3.37]
40	10	3	0.5	1	**3**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	2.98 (2.75, 3.22)[2.04, 4.04]
40	10	3	0.5	1	**3.5**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	3.48 (3.20, 3.75)[2.39, 4.70]
40	10	3	0.5	1	**4**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	3.98 (3.66, 4.28)[2.73, 5.37]
40	10	3	0.5	1	**4.5**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	4.48 (4.12, 4.82)[3.08, 6.03]
40	10	3	0.5	1	**5**	0.50 (0.49, 0.51)[0.46, 0.54]	1.00 (0.97, 1.03)[0.86, 1.15]	4.97 (4.57, 5.35)[3.43, 6.70]

Table C.27: Increased base n_1, n_2 and n_3: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
40	10	3	**1.25**	10	20	**0.50**	**27.93**	1.25 (1.22, 1.28)[1.12, 1.43]	10.01 (9.68, 10.31)[8.60, 11.51]	19.88 (18.32, 21.62)[13.31, 27.31]
40	10	3	**2.5**	10	20	**0.52**	**28.57**	2.50 (2.44, 2.57)[2.23, 2.86]	10.01 (9.67, 10.33)[8.61, 11.51]	19.88 (18.31, 21.64)[13.36, 27.32]
40	10	3	**3.75**	10	20	**0.53**	**29.60**	3.76 (3.65, 3.85)[3.35, 4.28]	10.01 (9.66, 10.34)[8.62, 11.50]	19.86 (18.30, 21.62)[13.39, 27.34]
40	10	3	**5**	10	20	**0.56**	**30.99**	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
40	10	3	**6.25**	10	20	**0.59**	**32.69**	6.26 (6.09, 6.42)[5.57, 7.14]	9.99 (9.64, 10.35)[8.47, 11.69]	19.85 (18.28, 21.66)[13.46, 27.37]
40	10	3	**7.5**	10	20	**0.63**	**34.65**	7.51 (7.31, 7.71)[6.69, 8.58]	9.99 (9.64, 10.37)[8.36, 11.84]	19.85 (18.26, 21.67)[13.48, 27.39]
40	10	3	**8.75**	10	20	**0.66**	**36.83**	8.76 (8.53, 8.99)[7.79, 10.01]	9.99 (9.62, 10.38)[8.23, 11.99]	19.83 (18.26, 21.69)[13.42, 27.40]
40	10	3	**10**	10	20	**0.71**	**39.20**	10.02 (9.75, 10.28)[8.90, 11.45]	9.99 (9.60, 10.41)[8.09, 12.15]	19.81 (18.27, 21.69)[13.35, 27.42]
40	10	3	5	**5**	20	**0.35**	**19.60**	5.01 (4.87, 5.14)[4.46, 5.71]	4.99 (4.80, 5.21)[4.04, 6.08]	19.89 (18.24, 21.55)[13.59, 27.34]
40	10	3	5	**7.5**	20	**0.45**	**24.99**	5.01 (4.87, 5.14)[4.46, 5.71]	7.49 (7.23, 7.77)[6.32, 8.81]	19.90 (18.26, 21.61)[13.61, 27.35]
40	10	3	5	**10**	20	**0.56**	**30.99**	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
40	10	3	5	**12.5**	20	**0.67**	**37.32**	5.01 (4.87, 5.14)[4.46, 5.71]	12.51 (12.07, 12.93)[10.77, 14.40]	19.81 (18.24, 21.67)[13.23, 27.37]
40	10	3	5	**15**	20	**0.79**	**43.83**	5.01 (4.87, 5.14)[4.46, 5.71]	15.01 (14.49, 15.51)[12.88, 17.32]	19.82 (18.20, 21.69)[13.03, 27.38]
40	10	3	5	**17.5**	20	**0.91**	**50.45**	5.01 (4.87, 5.14)[4.46, 5.72]	17.52 (16.90, 18.09)[14.97, 20.25]	19.89 (18.16, 21.72)[12.67, 27.39]
40	10	3	5	**20**	20	**1.03**	**57.14**	5.01 (4.87, 5.14)[4.46, 5.74]	20.04 (19.32, 20.66)[17.05, 23.19]	19.92 (18.13, 21.76)[12.17, 27.41]
40	10	3	5	10	**10**	**1.12**	30.99	5.01 (4.91, 5.10)[4.58, 5.56]	10.01 (9.70, 10.30)[8.71, 11.42]	9.96 (9.08, 10.87)[6.21, 13.68]
40	10	3	5	10	**15**	**0.75**	30.99	5.01 (4.89, 5.12)[4.52, 5.64]	10.00 (9.69, 10.31)[8.67, 11.45]	14.88 (13.68, 16.24)[10.02, 20.27]
40	10	3	5	10	**20**	**0.56**	30.99	5.01 (4.87, 5.14)[4.46, 5.71]	10.00 (9.65, 10.34)[8.57, 11.55]	19.85 (18.30, 21.63)[13.43, 27.36]
40	10	3	5	10	**25**	**0.45**	30.99	5.01 (4.85, 5.16)[4.41, 5.85]	10.00 (9.62, 10.37)[8.47, 11.84]	24.82 (22.82, 27.08)[16.71, 34.85]
40	10	3	5	10	**30**	**0.37**	30.99	5.01 (4.83, 5.18)[4.34, 6.03]	9.99 (9.59, 10.41)[8.37, 12.14]	29.73 (27.40, 32.49)[19.91, 42.63]
40	10	3	5	10	**35**	**0.32**	30.99	5.00 (4.81, 5.20)[4.28, 6.23]	9.99 (9.55, 10.44)[8.27, 12.46]	34.67 (31.99, 37.93)[23.03, 50.72]
40	10	3	5	10	**40**	**0.28**	30.99	5.00 (4.79, 5.23)[4.21, 6.44]	9.98 (9.51, 10.49)[8.12, 12.79]	39.65 (36.50, 43.44)[26.06, 59.13]
40	10	3	5	10	**45**	**0.25**	30.99	4.99 (4.77, 5.26)[4.14, 6.66]	9.99 (9.46, 10.53)[7.96, 13.14]	44.59 (41.00, 49.12)[29.02, 67.88]
40	10	3	5	10	**50**	**0.22**	30.99	4.99 (4.75, 5.29)[4.06, 6.90]	9.98 (9.42, 10.59)[7.81, 13.52]	49.46 (45.48, 54.72)[31.91, 77.00]

Table C.28: Increased base n_1, n_2 and n_3: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
40	10	3	**1.25**	10	20	**0.5**	**27.93**	0.51 (0.47, 0.55)[0.36, 0.76]	27.96 (27.06, 28.80)[24.08, 32.13]	10.00 (9.80, 10.21)[8.83, 10.91]
40	10	3	**2.5**	10	20	**0.52**	**28.57**	0.52 (0.48, 0.56)[0.37, 0.78]	28.62 (27.69, 29.47)[24.78, 32.75]	10.00 (9.80, 10.20)[8.83, 10.91]
40	10	3	**3.75**	10	20	**0.53**	**29.60**	0.54 (0.49, 0.59)[0.38, 0.80]	29.63 (28.68, 30.50)[25.90, 33.78]	10.00 (9.80, 10.20)[8.84, 10.92]
40	10	3	**5**	10	20	**0.56**	**30.99**	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
40	10	3	**6.25**	10	20	**0.59**	**32.69**	0.59 (0.55, 0.64)[0.42, 0.89]	32.70 (31.74, 33.63)[28.92, 37.29]	10.00 (9.80, 10.20)[8.84, 10.93]
40	10	3	**7.5**	10	20	**0.63**	**34.65**	0.63 (0.58, 0.68)[0.45, 0.94]	34.62 (33.68, 35.62)[30.56, 39.59]	10.00 (9.79, 10.20)[8.84, 10.93]
40	10	3	**8.75**	10	20	**0.66**	**36.83**	0.67 (0.62, 0.73)[0.48, 1.00]	36.79 (35.83, 37.85)[32.41, 42.09]	10.01 (9.79, 10.20)[8.84, 10.94]
40	10	3	**10**	10	20	**0.71**	**39.20**	0.71 (0.65, 0.77)[0.51, 1.06]	39.14 (38.15, 40.32)[34.45, 44.78]	10.01 (9.79, 10.20)[8.84, 10.94]
40	10	3	5	**5**	20	**0.35**	**19.60**	0.36 (0.33, 0.39)[0.25, 0.52]	19.56 (19.09, 20.15)[17.16, 22.40]	10.01 (9.80, 10.20)[8.89, 10.97]
40	10	3	5	**7.5**	20	**0.45**	**24.99**	0.45 (0.42, 0.49)[0.32, 0.67]	24.99 (24.26, 25.67)[22.05, 28.53]	10.00 (9.80, 10.20)[8.86, 10.95]
40	10	3	5	**10**	20	**0.56**	**30.99**	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
40	10	3	5	**12.5**	20	**0.67**	**37.32**	0.68 (0.62, 0.74)[0.48, 1.02]	37.36 (36.15, 38.41)[32.63, 42.62]	10.01 (9.80, 10.21)[8.81, 10.90]
40	10	3	5	**15**	20	**0.79**	**43.83**	0.80 (0.73, 0.87)[0.56, 1.23]	43.89 (42.43, 45.17)[38.02, 50.29]	10.01 (9.80, 10.21)[8.78, 10.88]
40	10	3	5	**17.5**	20	**0.91**	**50.45**	0.92 (0.84, 1.00)[0.65, 1.45]	50.54 (48.84, 52.05)[43.50, 58.10]	10.01 (9.79, 10.22)[8.75, 10.86]
40	10	3	5	**20**	20	**1.03**	**57.14**	1.04 (0.95, 1.13)[0.73, 1.71]	57.25 (55.31, 58.96)[49.02, 66.02]	10.00 (9.79, 10.22)[8.72, 10.88]
40	10	3	5	10	**10**	**1.12**	30.99	1.12 (1.02, 1.23)[0.79, 1.87]	31.03 (30.24, 31.75)[27.76, 34.61]	10.00 (9.90, 10.11)[9.36, 10.45]
40	10	3	5	10	**15**	**0.75**	30.99	0.75 (0.69, 0.82)[0.53, 1.14]	31.01 (30.18, 31.80)[27.61, 34.71]	10.01 (9.85, 10.15)[9.10, 10.67]
40	10	3	5	10	**20**	**0.56**	30.99	0.56 (0.52, 0.61)[0.40, 0.84]	31.01 (30.06, 31.90)[27.33, 35.26]	10.00 (9.80, 10.20)[8.84, 10.92]
40	10	3	5	10	**25**	**0.45**	30.99	0.45 (0.41, 0.49)[0.32, 0.67]	31.00 (29.94, 32.07)[26.98, 36.13]	10.00 (9.75, 10.25)[8.57, 11.17]
40	10	3	5	10	**30**	**0.37**	30.99	0.38 (0.35, 0.41)[0.27, 0.55]	30.97 (29.83, 32.17)[26.42, 37.05]	10.01 (9.71, 10.30)[8.31, 11.42]
40	10	3	5	10	**35**	**0.32**	30.99	0.32 (0.30, 0.35)[0.23, 0.47]	30.94 (29.72, 32.33)[25.89, 38.02]	10.01 (9.66, 10.34)[8.04, 11.67]
40	10	3	5	10	**40**	**0.28**	30.99	0.28 (0.26, 0.31)[0.20, 0.41]	30.92 (29.59, 32.48)[25.37, 39.04]	10.01 (9.60, 10.39)[7.78, 11.92]
40	10	3	5	10	**45**	**0.25**	30.99	0.25 (0.23, 0.27)[0.18, 0.37]	30.90 (29.48, 32.64)[24.88, 40.12]	10.01 (9.56, 10.44)[7.52, 12.17]
40	10	3	5	10	**50**	**0.22**	30.99	0.23 (0.21, 0.24)[0.16, 0.33]	30.89 (29.36, 32.77)[24.40, 41.26]	10.02 (9.51, 10.49)[7.25, 12.42]

Table C.29: Increased base σ_A, σ_I and σ_G: bias performance measures varying number of participants, observations and assessments.

Inputs						**Bias** ($\times 10^{-4}$)			**Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	1	2	4	-56.371	-366.732	-2900.124	-0.564	-1.834	-7.250	-3.697	-8.783	-20.212
10	4	2	1	2	4	-62.263	-65.512	-1296.551	-0.623	-0.328	-3.241	-5.697	-2.299	-12.768
20	4	2	1	2	4	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
30	4	2	1	2	4	-12.795	-123.300	-339.590	-0.128	-0.616	-0.849	-2.067	-7.232	-5.973
40	4	2	1	2	4	-10.810	-31.211	-280.579	-0.108	-0.156	-0.701	-1.968	-2.076	-5.644
60	4	2	1	2	4	-17.174	15.257	-125.618	-0.172	0.076	-0.314	-3.862	1.270	-3.290
100	4	2	1	2	4	7.217	-25.441	-4.737	0.072	-0.127	-0.012	2.070	-2.791	-0.154
20	**2**	2	1	2	4	-30.433	-271.340	-538.296	-0.304	-1.357	-1.346	-2.688	-7.896	-7.200
20	**4**	2	1	2	4	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	**6**	2	1	2	4	-28.690	-17.194	-351.430	-0.287	-0.086	-0.879	-4.516	-1.073	-5.085
20	**8**	2	1	2	4	-23.023	-16.580	-844.688	-0.230	-0.083	-2.112	-4.158	-1.238	-12.109
20	**12**	2	1	2	4	-2.209	-70.996	-231.106	-0.022	-0.355	-0.578	-0.494	-6.433	-3.481
20	**20**	2	1	2	4	2.124	-3.066	-775.382	0.021	-0.015	-1.938	0.618	-0.380	-12.474
20	4	**2**	1	2	4	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	**3**	1	2	4	11.408	-107.822	-391.215	0.114	-0.539	-0.978	1.996	-5.384	-5.597
20	4	**4**	1	2	4	-3.644	-113.164	-166.299	-0.036	-0.566	-0.416	-0.795	-5.777	-2.302
20	4	**6**	1	2	4	-4.723	-79.078	-432.974	-0.047	-0.395	-1.082	-1.359	-4.260	-6.157
20	4	**10**	1	2	4	0.651	-99.337	-530.504	0.007	-0.497	-1.326	0.253	-5.474	-7.677

Table C.30: Increased base σ_A, σ_I and σ_G: accuracy and coverage performance measures varying number of participants, observations and assessments.

						Accuracy and coverage								
Inputs						**Mean squared error ($\times 10^{-4}$)**			**Coverage**			**Mean 95% CI width**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	1	2	4	232.754	1756.769	21429.835	0.956	0.962[a]	0.972[b]	0.675	1.835[a]	9.527[b]
10	4	2	1	2	4	119.843	812.126	10480.473	0.958	0.957	0.951[c]	0.456	1.216	4.755[c]
20	4	2	1	2	4	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
30	4	2	1	2	4	38.336	292.217	3243.548	0.961	0.945	0.957	0.257	0.671	2.329
40	4	2	1	2	4	30.190	226.138	2479.537	0.955	0.945	0.939	0.221	0.580	1.980
60	4	2	1	2	4	19.804	144.323	1459.003	0.957	0.948	0.953	0.180	0.472	1.590
100	4	2	1	2	4	12.166	83.132	940.326	0.952	0.950	0.953	0.139	0.364	1.215
20	**2**	2	1	2	4	128.258	1188.409	5619.268	0.939	0.953	0.957[d]	0.457	1.523	3.240[d]
20	**4**	2	1	2	4	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
20	**6**	2	1	2	4	40.434	256.991	4789.279	0.951	0.945	0.945	0.256	0.638	2.907
20	**8**	2	1	2	4	30.715	179.462	4937.400	0.954	0.951	0.934	0.221	0.537	2.841
20	**12**	2	1	2	4	19.990	122.319	4412.816	0.950	0.947	0.947	0.180	0.425	2.849
20	**20**	2	1	2	4	11.828	65.162	3923.862	0.955	0.953	0.960	0.139	0.323	2.785
20	4	**2**	1	2	4	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
20	4	**3**	1	2	4	32.670	402.270	4901.357	0.948	0.941	0.954	0.222	0.797	2.966
20	4	**4**	1	2	4	21.033	385.052	5220.485	0.958	0.949	0.934	0.180	0.780	2.977
20	4	**6**	1	2	4	12.083	345.141	4964.290	0.960	0.958	0.942	0.139	0.765	2.955
20	4	**10**	1	2	4	6.642	330.280	4802.942	0.949	0.956	0.944	0.104	0.752	2.945

[a]12 CIs could not be calculated; [b]78 CIs could not be calculated; [c]4 CIs could not be calculated; [d]2 CIs could not be calculated.

Table C.31: Increased base σ_A, σ_I and σ_G: bias performance measures varying CV_A, CV_I and CV_G.

			Inputs			**Bias** ($\times 10^{-4}$)			**Bias** **Percentage bias**			**Standardised bias**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	2	4	-2.188	-56.493	-620.326	-0.175	-0.282	-1.551	-2.272	-3.132	-9.042
20	4	2	**0.25**	2	4	-4.380	-55.916	-617.861	-0.175	-0.280	-1.545	-2.274	-3.092	-9.004
20	4	2	**0.375**	2	4	-6.564	-55.962	-615.907	-0.175	-0.280	-1.540	-2.272	-3.073	-8.972
20	4	2	**0.5**	2	4	-8.745	-56.614	-614.303	-0.175	-0.283	-1.536	-2.270	-3.073	-8.942
20	4	2	**0.625**	2	4	-10.924	-57.911	-613.067	-0.175	-0.290	-1.533	-2.269	-3.095	-8.915
20	4	2	**0.75**	2	4	-13.107	-59.902	-612.223	-0.175	-0.300	-1.531	-2.269	-3.138	-8.891
20	4	2	**0.875**	2	4	-15.291	-62.696	-611.736	-0.175	-0.313	-1.529	-2.269	-3.206	-8.870
20	4	2	**1**	2	4	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	**1.125**	2	4	-19.672	-71.120	-611.616	-0.175	-0.356	-1.529	-2.270	-3.424	-8.834
20	4	2	**1.25**	2	4	-21.870	-76.988	-612.051	-0.175	-0.385	-1.530	-2.271	-3.576	-8.819
20	4	2	1	**1**	4	-17.491	-73.461	-508.728	-0.175	-0.735	-1.272	-2.271	-5.178	-7.659
20	4	2	1	**1.5**	4	-17.501	-61.202	-553.799	-0.175	-0.408	-1.384	-2.272	-3.694	-8.207
20	4	2	1	**2**	4	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	1	**2.5**	4	-17.478	-76.141	-682.737	-0.175	-0.305	-1.707	-2.269	-3.162	-9.578
20	4	2	1	**3**	4	-17.479	-87.736	-769.232	-0.175	-0.292	-1.923	-2.269	-3.107	-10.384
20	4	2	1	**3.5**	4	-17.479	-100.266	-873.448	-0.175	-0.286	-2.184	-2.269	-3.084	-11.264
20	4	2	1	**4**	4	-17.477	-113.324	-998.791	-0.175	-0.283	-2.497	-2.269	-3.076	-12.220
20	4	2	1	2	**2**	-17.484	-66.409	-505.975	-0.175	-0.332	-2.530	-2.270	-3.302	-12.133
20	4	2	1	2	**2.5**	-17.484	-66.403	-507.678	-0.175	-0.332	-2.031	-2.270	-3.301	-10.657
20	4	2	1	2	**3**	-17.483	-66.399	-533.024	-0.175	-0.332	-1.777	-2.270	-3.301	-9.795
20	4	2	1	2	**3.5**	-17.481	-66.398	-569.272	-0.175	-0.332	-1.626	-2.269	-3.301	-9.237
20	4	2	1	2	**4**	-17.479	-66.399	-611.533	-0.175	-0.332	-1.529	-2.269	-3.301	-8.851
20	4	2	1	2	**4.5**	-17.478	-66.403	-657.493	-0.175	-0.332	-1.461	-2.269	-3.301	-8.571
20	4	2	1	2	**5**	-17.477	-66.409	-705.912	-0.175	-0.332	-1.412	-2.269	-3.302	-8.359

Table C.32: Increased base σ_A, σ_I and σ_G: accuracy and coverage performance measures varying CV_A, CV_I and CV_G.

						Accuracy and coverage								
			Inputs			**Mean squared error ($\times 10^{-4}$)**			**Coverage**			**Mean 95% CI width**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	2	4	0.928	325.629	4745.059	0.952	0.957	0.953	0.040	0.737	2.936
20	4	2	**0.25**	2	4	3.711	327.431	4746.499	0.952	0.957	0.951	0.079	0.741	2.937
20	4	2	**0.375**	2	4	8.348	332.050	4750.656	0.952	0.958	0.952	0.119	0.748	2.939
20	4	2	**0.5**	2	4	14.844	339.630	4757.399	0.952	0.959	0.951	0.158	0.759	2.942
20	4	2	**0.625**	2	4	23.191	350.417	4766.830	0.952	0.957	0.953	0.198	0.772	2.945
20	4	2	**0.75**	2	4	33.393	364.679	4778.943	0.952	0.959	0.954	0.237	0.789	2.949
20	4	2	**0.875**	2	4	45.450	382.730	4793.647	0.952	0.959	0.954	0.277	0.809	2.954
20	4	2	**1**	2	4	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
20	4	2	**1.125**	2	4	75.133	431.897	4831.060	0.952	0.960	0.954	0.356	0.861	2.965
20	4	2	**1.25**	2	4	92.759	464.004	4853.921	0.952	0.955	0.955	0.395	0.893	2.972
20	4	2	1	**1**	4	59.367	201.836	4437.353	0.952	0.961[a]	0.950	0.316	0.594[a]	2.831
20	4	2	1	**1.5**	4	59.364	274.827	4584.246	0.952	0.955	0.951	0.316	0.687	2.883
20	4	2	1	**2**	4	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
20	4	2	1	**2.5**	4	59.366	580.348	5127.269	0.952	0.959	0.959	0.316	0.995	3.060
20	4	2	1	**3**	4	59.366	798.212	5546.939	0.952	0.962	0.961	0.316	1.166	3.189
20	4	2	1	**3.5**	4	59.366	1057.746	6088.919	0.952	0.958	0.962[b]	0.316	1.340	3.357[b]
20	4	2	1	**4**	4	59.368	1358.504	6779.718	0.952	0.959	0.971[c]	0.316	1.517	3.550[c]
20	4	2	1	2	**2**	59.367	404.981	1764.673	0.952	0.958	0.971[d]	0.316	0.833	1.818[d]
20	4	2	1	2	**2.5**	59.367	404.985	2295.218	0.952	0.958	0.958[e]	0.316	0.833	2.056[e]
20	4	2	1	2	**3**	59.367	404.991	2989.920	0.952	0.958	0.958	0.316	0.833	2.339
20	4	2	1	2	**3.5**	59.368	404.994	3830.570	0.952	0.958	0.954	0.316	0.833	2.643
20	4	2	1	2	**4**	59.365	404.991	4810.967	0.952	0.958	0.953	0.316	0.833	2.959
20	4	2	1	2	**4.5**	59.366	404.985	5928.418	0.952	0.958	0.956	0.316	0.833	3.282
20	4	2	1	2	**5**	59.366	404.989	7181.564	0.952	0.958	0.954	0.316	0.833	3.610

[a]8 CIs could not be calculated; [b]2 CIs could not be calculated; [c]10 CIs could not be calculated; [d]18 CIs could not be calculated; [e]1 CIs could not be calculated.

Table C.33: Increased base σ_A, σ_I and σ_G: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
5	4	2	1	2	4	0.99 (0.89, 1.09)[0.50, 1.52]	1.97 (1.69, 2.26)[0.69, 3.07]	3.62 (2.74, 4.62)[0.00, 8.21]
10	4	2	1	2	4	0.99 (0.92, 1.07)[0.61, 1.43]	1.99 (1.78, 2.18)[1.13, 3.08]	3.83 (3.14, 4.52)[1.02, 7.40]
20	4	2	1	2	4	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
30	4	2	1	2	4	1.00 (0.96, 1.04)[0.80, 1.23]	1.99 (1.87, 2.09)[1.34, 2.64]	3.96 (3.59, 4.33)[2.05, 6.37]
40	4	2	1	2	4	1.00 (0.96, 1.04)[0.84, 1.16]	2.00 (1.90, 2.11)[1.46, 2.44]	3.97 (3.64, 4.30)[2.62, 5.61]
60	4	2	1	2	4	1.00 (0.97, 1.03)[0.86, 1.14]	2.00 (1.92, 2.08)[1.54, 2.39]	4.00 (3.74, 4.23)[2.80, 5.45]
100	4	2	1	2	4	1.00 (0.98, 1.02)[0.89, 1.12]	2.00 (1.93, 2.06)[1.72, 2.33]	3.99 (3.78, 4.23)[3.16, 4.94]
20	**2**	2	1	2	4	1.00 (0.92, 1.07)[0.61, 1.36]	1.96 (1.74, 2.20)[0.95, 3.10]	3.94 (3.40, 4.44)[1.48, 6.45]
20	**4**	2	1	2	4	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
20	**6**	2	1	2	4	1.00 (0.95, 1.04)[0.79, 1.19]	2.00 (1.90, 2.10)[1.39, 2.57]	3.97 (3.49, 4.40)[1.92, 6.50]
20	**8**	2	1	2	4	1.00 (0.96, 1.04)[0.80, 1.21]	1.99 (1.90, 2.09)[1.63, 2.41]	3.88 (3.44, 4.36)[1.96, 6.09]
20	**12**	2	1	2	4	1.00 (0.97, 1.03)[0.87, 1.15]	2.00 (1.92, 2.07)[1.64, 2.33]	3.94 (3.51, 4.41)[2.03, 6.27]
20	**20**	2	1	2	4	1.00 (0.98, 1.02)[0.89, 1.12]	2.00 (1.94, 2.05)[1.73, 2.24]	3.89 (3.49, 4.34)[1.98, 5.84]
20	4	**2**	1	2	4	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
20	4	**3**	1	2	4	1.00 (0.96, 1.04)[0.84, 1.16]	1.99 (1.87, 2.12)[1.37, 2.56]	3.94 (3.46, 4.43)[1.54, 6.55]
20	4	**4**	1	2	4	1.00 (0.97, 1.03)[0.87, 1.19]	1.99 (1.85, 2.13)[1.48, 2.64]	3.96 (3.48, 4.49)[2.05, 6.29]
20	4	**6**	1	2	4	1.00 (0.98, 1.02)[0.88, 1.13]	1.99 (1.87, 2.11)[1.33, 2.62]	3.96 (3.43, 4.40)[2.12, 6.55]
20	4	**10**	1	2	4	1.00 (0.98, 1.02)[0.92, 1.07]	1.98 (1.86, 2.11)[1.48, 2.61]	3.93 (3.49, 4.41)[1.92, 6.19]

Table C.34: Increased base σ_A, σ_I and σ_G: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
5	4	2	10	20	40	0.56	61.98	10.01 (8.41, 11.56)[4.62, 31.23]	19.76 (16.07, 23.64)[5.78, 49.27]	36.51 (26.71, 48.35)[0.00, 217.76]
10	4	2	10	20	40	0.56	61.98	9.99 (8.84, 11.25)[5.86, 19.40]	20.05 (17.50, 22.66)[9.96, 40.25]	38.50 (31.28, 47.03)[11.42, 92.20]
20	4	2	10	20	40	0.56	61.98	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
30	4	2	10	20	40	0.56	61.98	9.98 (9.37, 10.70)[7.32, 14.10]	19.88 (18.49, 21.59)[13.39, 29.90]	39.74 (35.39, 44.47)[20.01, 64.13]
40	4	2	10	20	40	0.56	61.98	10.01 (9.48, 10.62)[7.76, 12.87]	19.94 (18.57, 21.42)[14.18, 28.40]	39.65 (35.92, 43.79)[25.03, 58.96]
60	4	2	10	20	40	0.56	61.98	9.96 (9.54, 10.48)[7.91, 12.68]	20.06 (19.01, 21.13)[15.58, 26.06]	39.91 (36.94, 42.87)[27.10, 53.83]
100	4	2	10	20	40	0.56	61.98	9.99 (9.65, 10.40)[8.58, 11.82]	20.01 (19.19, 20.82)[16.25, 23.64]	39.89 (37.53, 42.49)[30.93, 50.92]
20	**2**	2	10	20	40	0.56	61.98	10.00 (9.04, 11.04)[5.75, 16.25]	19.75 (17.21, 22.57)[9.46, 35.27]	39.63 (33.38, 45.55)[13.86, 73.23]
20	**4**	2	10	20	40	0.56	61.98	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
20	**6**	2	10	20	40	0.56	61.98	10.01 (9.28, 10.76)[6.95, 14.97]	19.99 (18.48, 21.79)[12.37, 30.33]	39.55 (34.80, 44.98)[19.20, 71.34]
20	**8**	2	10	20	40	0.56	61.98	9.97 (9.35, 10.75)[7.09, 15.12]	19.90 (18.46, 21.76)[13.44, 32.67]	39.05 (34.06, 44.46)[19.08, 72.99]
20	**12**	2	10	20	40	0.56	61.98	9.99 (9.36, 10.68)[7.33, 13.70]	20.02 (18.56, 21.36)[14.03, 27.32]	39.52 (34.75, 44.96)[20.27, 72.08]
20	**20**	2	10	20	40	0.56	61.98	10.01 (9.40, 10.73)[7.41, 14.70]	20.02 (18.78, 21.38)[15.35, 28.32]	39.10 (34.59, 44.10)[17.37, 67.65]
20	4	**2**	10	20	40	0.56	61.98	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
20	4	**3**	10	20	40	0.56	61.98	10.06 (9.28, 10.81)[7.08, 13.81]	19.88 (18.16, 21.87)[12.45, 32.91]	39.39 (34.49, 45.00)[15.99, 67.61]
20	4	**4**	10	20	40	0.56	61.98	9.97 (9.36, 10.68)[7.35, 14.15]	19.95 (18.04, 21.84)[13.02, 32.18]	39.87 (34.23, 45.35)[19.01, 69.79]
20	4	**6**	10	20	40	0.56	61.98	10.03 (9.41, 10.75)[7.66, 14.49]	19.98 (18.34, 21.83)[12.59, 31.28]	39.59 (34.34, 44.87)[18.49, 67.57]
20	4	**10**	10	20	40	0.56	61.98	10.02 (9.37, 10.69)[7.50, 13.52]	19.88 (18.31, 21.74)[13.09, 29.42]	39.38 (34.37, 44.52)[18.65, 68.28]

Table C.35: Increased base σ_A, σ_I and σ_G: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying number of participants, observations and assessments. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
5	4	2	10	20	40	0.56	61.98	0.61 (0.46, 0.83)[0.20, 24030.27]	61.26 (51.62, 72.16)[26.39, 161.48]	9.98 (8.74, 11.37)[3.21, 15.02]
10	4	2	10	20	40	0.56	61.98	0.58 (0.48, 0.72)[0.27, 2.14]	62.36 (55.34, 69.57)[33.27, 123.84]	9.98 (9.10, 10.83)[5.26, 14.22]
20	4	2	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
30	4	2	10	20	40	0.56	61.98	0.56 (0.51, 0.63)[0.36, 1.13]	61.73 (57.78, 66.49)[43.93, 88.57]	10.01 (9.47, 10.50)[7.70, 12.78]
40	4	2	10	20	40	0.56	61.98	0.56 (0.51, 0.62)[0.35, 0.93]	61.93 (58.22, 65.91)[47.75, 85.16]	9.98 (9.54, 10.44)[8.25, 12.00]
60	4	2	10	20	40	0.56	61.98	0.56 (0.52, 0.61)[0.39, 0.78]	62.19 (59.35, 65.27)[49.72, 78.22]	9.98 (9.62, 10.35)[8.60, 11.87]
100	4	2	10	20	40	0.56	61.98	0.56 (0.53, 0.59)[0.43, 0.74]	62.08 (59.72, 64.28)[51.10, 72.34]	9.99 (9.70, 10.30)[8.67, 11.44]
20	**2**	2	10	20	40	0.56	61.98	0.56 (0.47, 0.68)[0.28, 1.97]	61.51 (54.79, 68.72)[34.55, 104.05]	10.01 (9.33, 10.64)[6.53, 12.87]
20	**4**	2	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
20	**6**	2	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.35, 1.18]	62.02 (57.67, 67.02)[42.18, 90.54]	9.98 (9.35, 10.62)[7.04, 12.64]
20	**8**	2	10	20	40	0.56	61.98	0.57 (0.51, 0.65)[0.36, 1.21]	61.75 (57.61, 66.73)[42.37, 99.79]	9.99 (9.39, 10.63)[6.58, 13.04]
20	**12**	2	10	20	40	0.56	61.98	0.56 (0.50, 0.64)[0.35, 1.10]	61.99 (57.73, 65.89)[44.74, 83.17]	10.01 (9.42, 10.60)[7.50, 12.81]
20	**20**	2	10	20	40	0.56	61.98	0.57 (0.51, 0.64)[0.37, 1.10]	62.19 (58.39, 66.17)[48.12, 86.99]	9.97 (9.37, 10.57)[7.16, 12.87]
20	4	**2**	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
20	4	**3**	10	20	40	0.56	61.98	0.56 (0.50, 0.65)[0.30, 1.59]	61.96 (56.88, 67.38)[42.61, 98.57]	9.99 (9.35, 10.63)[7.26, 12.80]
20	4	**4**	10	20	40	0.56	61.98	0.56 (0.49, 0.65)[0.32, 1.13]	61.79 (56.78, 67.15)[41.94, 97.21]	10.02 (9.39, 10.62)[6.94, 12.72]
20	4	**6**	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.33, 1.05]	61.96 (57.54, 67.26)[43.46, 93.59]	9.97 (9.35, 10.55)[6.98, 12.60]
20	4	**10**	10	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.34, 1.22]	61.72 (57.22, 66.80)[42.57, 89.12]	9.98 (9.40, 10.67)[7.34, 13.01]

Table C.36: Increased base σ_A, σ_I and σ_G: SD estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G.

Inputs						**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
20	4	2	**0.125**	2	4	0.12 (0.12, 0.13)[0.10, 0.15]	1.99 (1.87, 2.12)[1.42, 2.52]	3.92 (3.45, 4.42)[1.86, 6.49]
20	4	2	**0.25**	2	4	0.25 (0.24, 0.26)[0.19, 0.31]	1.99 (1.87, 2.12)[1.42, 2.51]	3.92 (3.45, 4.41)[1.86, 6.50]
20	4	2	**0.375**	2	4	0.37 (0.35, 0.39)[0.29, 0.46]	1.99 (1.87, 2.12)[1.42, 2.50]	3.92 (3.45, 4.41)[1.86, 6.50]
20	4	2	**0.5**	2	4	0.50 (0.47, 0.52)[0.39, 0.61]	1.99 (1.87, 2.12)[1.42, 2.50]	3.92 (3.44, 4.41)[1.86, 6.51]
20	4	2	**0.625**	2	4	0.62 (0.59, 0.65)[0.48, 0.77]	1.99 (1.87, 2.13)[1.42, 2.49]	3.92 (3.43, 4.41)[1.86, 6.51]
20	4	2	**0.75**	2	4	0.75 (0.71, 0.78)[0.58, 0.92]	1.99 (1.87, 2.13)[1.42, 2.49]	3.92 (3.43, 4.40)[1.86, 6.52]
20	4	2	**0.875**	2	4	0.87 (0.83, 0.92)[0.67, 1.07]	1.99 (1.86, 2.13)[1.42, 2.48]	3.93 (3.43, 4.40)[1.86, 6.52]
20	4	2	**1**	2	4	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
20	4	2	**1.125**	2	4	1.12 (1.06, 1.18)[0.87, 1.38]	1.99 (1.85, 2.14)[1.37, 2.49]	3.92 (3.42, 4.41)[1.85, 6.53]
20	4	2	**1.25**	2	4	1.25 (1.18, 1.31)[0.96, 1.53]	1.99 (1.85, 2.14)[1.35, 2.51]	3.92 (3.43, 4.41)[1.85, 6.53]
20	4	2	1	**1**	2	1.00 (0.95, 1.05)[0.77, 1.23]	0.99 (0.90, 1.09)[0.48, 1.37]	3.92 (3.47, 4.40)[1.97, 6.43]
20	4	2	1	**1.5**	2	1.00 (0.95, 1.05)[0.77, 1.23]	1.49 (1.38, 1.61)[0.99, 1.89]	3.92 (3.45, 4.40)[1.92, 6.48]
20	4	2	1	**2**	2	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
20	4	2	1	**2.5**	2	1.00 (0.95, 1.05)[0.77, 1.23]	2.49 (2.33, 2.66)[1.77, 3.11]	3.92 (3.41, 4.42)[1.77, 6.57]
20	4	2	1	**3**	2	1.00 (0.95, 1.05)[0.77, 1.23]	2.99 (2.80, 3.19)[2.13, 3.74]	3.91 (3.38, 4.44)[1.66, 6.60]
20	4	2	1	**3.5**	2	1.00 (0.95, 1.05)[0.77, 1.23]	3.49 (3.27, 3.71)[2.48, 4.37]	3.91 (3.37, 4.44)[1.53, 6.64]
20	4	2	1	**4**	2	1.00 (0.95, 1.05)[0.77, 1.23]	3.99 (3.73, 4.24)[2.83, 5.00]	3.90 (3.34, 4.45)[1.36, 6.67]
20	4	2	1	2	**2**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	1.95 (1.66, 2.23)[0.66, 3.35]
20	4	2	1	2	**2.5**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	2.44 (2.10, 2.77)[1.01, 4.15]
20	4	2	1	2	**3**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	2.94 (2.55, 3.31)[1.30, 4.94]
20	4	2	1	2	**3.5**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.44 (2.99, 3.86)[1.58, 5.73]
20	4	2	1	2	**4**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	3.93 (3.43, 4.41)[1.85, 6.52]
20	4	2	1	2	**4.5**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	4.41 (3.87, 4.95)[2.12, 7.31]
20	4	2	1	2	**5**	1.00 (0.95, 1.05)[0.77, 1.23]	1.99 (1.86, 2.14)[1.39, 2.48]	4.91 (4.32, 5.49)[2.39, 8.10]

Table C.37: Increased base σ_A, σ_I and σ_G: CV estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	CV_A	CV_I	CV_G
20	4	2	**1.25**	20	40	0.50	55.55	1.25 (1.16, 1.35)[0.81, 2.10]	20.10 (18.30, 21.90)[12.34, 31.40]	39.46 (34.22, 45.39)[17.81, 81.84]
20	4	2	**2.5**	20	40	0.50	55.87	2.51 (2.32, 2.71)[1.61, 4.21]	20.09 (18.29, 21.85)[12.35, 31.69]	39.48 (34.24, 45.34)[17.82, 81.37]
20	4	2	**3.75**	20	40	0.51	56.40	3.76 (3.47, 4.07)[2.42, 6.32]	20.13 (18.34, 21.82)[12.36, 31.97]	39.49 (34.27, 45.32)[17.83, 80.90]
20	4	2	**5**	20	40	0.52	57.14	5.02 (4.63, 5.42)[3.23, 8.43]	20.13 (18.36, 21.84)[12.38, 32.24]	39.52 (34.27, 45.33)[17.84, 80.43]
20	4	2	**6.25**	20	40	0.52	58.08	6.27 (5.79, 6.78)[4.03, 10.55]	20.14 (18.30, 21.87)[12.41, 32.50]	39.51 (34.24, 45.28)[17.85, 79.98]
20	4	2	**7.5**	20	40	0.53	59.21	7.53 (6.95, 8.14)[4.84, 12.67]	20.14 (18.31, 21.89)[12.44, 32.74]	39.54 (34.23, 45.32)[17.85, 79.53]
20	4	2	**8.75**	20	40	0.55	60.51	8.78 (8.11, 9.49)[5.65, 14.80]	20.13 (18.25, 21.95)[12.48, 32.97]	39.55 (34.24, 45.34)[17.86, 79.08]
20	4	2	**10**	20	40	0.56	61.98	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
20	4	2	**11.25**	20	40	0.57	63.61	11.28 (10.42, 12.20)[7.26, 19.05]	20.06 (18.14, 22.06)[12.15, 33.40]	39.54 (34.21, 45.34)[17.88, 78.20]
20	4	2	**12.5**	20	40	0.59	65.37	12.55 (11.58, 13.55)[8.07, 21.19]	20.06 (18.10, 22.08)[11.92, 33.60]	39.52 (34.21, 45.41)[17.89, 77.78]
20	4	2	10	**10**	40	**0.35**	**39.20**	10.02 (9.28, 10.84)[6.52, 17.16]	10.05 (8.87, 11.15)[4.35, 17.26]	39.46 (34.62, 45.35)[18.99, 78.84]
20	4	2	10	**15**	40	**0.45**	**49.97**	10.02 (9.28, 10.83)[6.49, 17.04]	15.04 (13.53, 16.53)[8.82, 25.25]	39.54 (34.55, 45.35)[18.51, 78.70]
20	4	2	10	**20**	40	**0.56**	**61.98**	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
20	4	2	10	**25**	40	**0.67**	**74.63**	10.05 (9.26, 10.85)[6.43, 16.81]	25.18 (22.86, 27.40)[15.53, 41.15]	39.43 (33.92, 45.35)[17.05, 78.64]
20	4	2	10	**30**	40	**0.79**	**87.65**	10.03 (9.23, 10.86)[6.40, 16.78]	30.22 (27.37, 32.87)[18.51, 49.15]	39.49 (33.65, 45.40)[16.01, 78.70]
20	4	2	10	**35**	40	**0.91**	**100.90**	10.04 (9.22, 10.88)[6.37, 16.84]	35.18 (31.88, 38.39)[21.46, 57.20]	39.20 (33.43, 45.39)[14.72, 78.82]
20	4	2	10	**40**	40	**1.03**	**114.29**	10.04 (9.20, 10.91)[6.34, 16.91]	40.24 (36.44, 44.02)[24.39, 65.31]	39.34 (33.16, 45.32)[13.10, 78.99]
20	4	2	10	20	**20**	**1.12**	61.98	9.99 (9.43, 10.63)[7.10, 13.60]	19.93 (18.52, 21.60)[13.14, 26.92]	19.50 (16.66, 22.37)[6.57, 32.73]
20	4	2	10	20	**25**	**0.89**	61.98	10.01 (9.37, 10.67)[6.93, 14.27]	19.98 (18.48, 21.70)[12.94, 28.26]	24.37 (21.02, 28.03)[9.87, 42.42]
20	4	2	10	20	**30**	**0.75**	61.98	10.03 (9.34, 10.72)[6.76, 15.05]	20.03 (18.37, 21.84)[12.74, 29.73]	29.47 (25.36, 33.74)[12.68, 53.19]
20	4	2	10	20	**35**	**0.64**	61.98	10.02 (9.30, 10.77)[6.61, 15.93]	20.05 (18.29, 21.96)[12.55, 31.37]	34.57 (29.89, 39.58)[15.33, 65.19]
20	4	2	10	20	**40**	**0.56**	61.98	10.03 (9.26, 10.85)[6.46, 16.92]	20.10 (18.26, 21.97)[12.37, 33.19]	39.56 (34.23, 45.38)[17.87, 78.64]
20	4	2	10	20	**45**	**0.50**	61.98	10.03 (9.23, 10.89)[6.31, 18.04]	20.13 (18.15, 22.09)[12.19, 35.24]	44.46 (38.54, 51.31)[20.35, 93.81]
20	4	2	10	20	**50**	**0.45**	61.98	10.03 (9.19, 10.99)[6.18, 19.32]	20.17 (18.03, 22.16)[12.02, 37.57]	49.45 (42.65, 57.09)[22.77, 111.04]

Table C.38: Increased base σ_A, σ_I and σ_G: II, RCV and mean estimates, median (Q1, Q3)[minimum, maximum] from biological variability data simulations varying σ_A, σ_I and σ_G. CVs and RCVs are displayed as percentages.

Inputs								**Median (Q1, Q3)[minimum, maximum]**		
n_1	n_2	n_3	CV_A	CV_I	CV_G	II	RCV	II	RCV	$Mean$
20	4	2	**1.25**	20	40	0.50	55.55	0.51 (0.45, 0.59)[0.29, 1.15]	55.83 (50.82, 60.79)[34.34, 87.23]	9.97 (9.32, 10.58)[6.74, 13.09]
20	4	2	**2.5**	20	40	0.50	55.87	0.51 (0.45, 0.59)[0.29, 1.16]	56.16 (51.08, 60.98)[34.77, 88.61]	9.97 (9.33, 10.57)[6.75, 13.10]
20	4	2	**3.75**	20	40	0.51	56.40	0.52 (0.45, 0.60)[0.30, 1.17]	56.71 (51.73, 61.47)[35.49, 90.31]	9.97 (9.33, 10.57)[6.75, 13.11]
20	4	2	**5**	20	40	0.52	57.14	0.53 (0.46, 0.61)[0.30, 1.18]	57.45 (52.46, 62.38)[36.48, 92.32]	9.96 (9.33, 10.57)[6.76, 13.11]
20	4	2	**6.25**	20	40	0.52	58.08	0.54 (0.47, 0.61)[0.30, 1.20]	58.43 (53.44, 63.37)[37.71, 94.60]	9.95 (9.33, 10.56)[6.76, 13.12]
20	4	2	**7.5**	20	40	0.53	59.21	0.55 (0.48, 0.62)[0.31, 1.22]	59.58 (54.46, 64.55)[39.18, 97.14]	9.95 (9.34, 10.56)[6.77, 13.13]
20	4	2	**8.75**	20	40	0.55	60.51	0.56 (0.49, 0.64)[0.32, 1.24]	60.89 (55.76, 65.92)[40.85, 99.92]	9.95 (9.34, 10.56)[6.77, 13.14]
20	4	2	**10**	20	40	0.56	61.98	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
20	4	2	**11.25**	20	40	0.57	63.61	0.58 (0.52, 0.67)[0.34, 1.30]	64.02 (58.57, 69.35)[44.72, 106.11]	9.95 (9.35, 10.57)[6.77, 13.16]
20	4	2	**12.5**	20	40	0.59	65.37	0.60 (0.53, 0.69)[0.35, 1.33]	65.78 (60.26, 71.52)[46.07, 109.49]	9.94 (9.35, 10.56)[6.76, 13.17]
20	4	2	10	**10**	40	**0.35**	**39.20**	0.36 (0.32, 0.41)[0.22, 0.74]	39.28 (36.42, 42.61)[27.84, 66.21]	9.97 (9.36, 10.55)[6.68, 12.96]
20	4	2	10	**15**	40	**0.45**	**49.97**	0.46 (0.41, 0.52)[0.27, 0.98]	50.20 (46.18, 54.62)[35.05, 83.72]	9.96 (9.35, 10.55)[6.73, 13.06]
20	4	2	10	**20**	40	**0.56**	**61.98**	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
20	4	2	10	**25**	40	**0.67**	**74.63**	0.69 (0.60, 0.79)[0.38, 1.61]	75.03 (68.66, 81.61)[49.63, 123.11]	9.96 (9.33, 10.57)[6.75, 13.25]
20	4	2	10	**30**	40	**0.79**	**87.65**	0.81 (0.70, 0.94)[0.44, 2.02]	88.22 (80.58, 95.95)[56.89, 143.96]	9.95 (9.31, 10.58)[6.73, 13.34]
20	4	2	10	**35**	40	**0.91**	**100.90**	0.93 (0.80, 1.09)[0.50, 2.53]	101.57 (92.29, 110.66)[64.33, 165.28]	9.95 (9.32, 10.61)[6.70, 13.44]
20	4	2	10	**40**	40	**1.03**	**114.29**	1.06 (0.90, 1.25)[0.55, 3.23]	114.73 (104.37, 125.66)[71.86, 186.99]	9.94 (9.29, 10.65)[6.67, 13.53]
20	4	2	10	20	**20**	**1.12**	61.98	1.15 (0.99, 1.35)[0.61, 3.79]	61.93 (58.08, 65.98)[45.21, 83.48]	9.96 (9.64, 10.31)[8.36, 11.81]
20	4	2	10	20	**25**	**0.89**	61.98	0.92 (0.80, 1.06)[0.50, 2.34]	62.14 (57.82, 66.36)[44.55, 87.62]	9.96 (9.57, 10.37)[7.96, 12.14]
20	4	2	10	20	**30**	**0.75**	61.98	0.76 (0.66, 0.88)[0.43, 1.81]	62.20 (57.64, 66.82)[43.92, 92.19]	9.97 (9.50, 10.43)[7.57, 12.48]
20	4	2	10	20	**35**	**0.64**	61.98	0.65 (0.57, 0.75)[0.37, 1.49]	62.36 (57.35, 67.16)[43.30, 97.26]	9.95 (9.42, 10.50)[7.17, 12.82]
20	4	2	10	20	**40**	**0.56**	61.98	0.57 (0.50, 0.65)[0.33, 1.27]	62.42 (57.02, 67.53)[42.70, 102.92]	9.95 (9.35, 10.57)[6.78, 13.15]
20	4	2	10	20	**45**	**0.50**	61.98	0.51 (0.45, 0.58)[0.29, 1.11]	62.38 (56.81, 67.98)[42.12, 109.28]	9.95 (9.26, 10.63)[6.36, 13.49]
20	4	2	10	20	**50**	**0.45**	61.98	0.45 (0.40, 0.52)[0.27, 0.99]	62.32 (56.60, 68.60)[41.55, 116.48]	9.95 (9.18, 10.70)[5.93, 13.83]

Appendix D

The impact of outlier detection and removal on studies of biological variability

D.1 Outlier detection methods with outlier simulation–poor test performance

Table D.1: Increased CV_A: outlier detection methods with no outlier simulation–outliers removed by each detection method; median (Q1, Q3) [minimum, maximum]. Maximum number of measurements is 160 and maximum number of individuals is 20.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	2 (0, 4)[0, 30]	1 (0, 3)[0, 19]	1 (0, 2)[0, 5]	5 (0, 10)[0, 25]
Cochran C test partial	2 (0, 4)[0, 10]	1 (0, 3)[0, 6]	1 (0, 2)[0, 5]	5 (0, 10)[0, 25]
Fraser-Harris method	2 (2, 8)[0, 20]	1 (1, 5)[0, 13]	1 (1, 2)[0, 6]	5 (5, 10)[0, 30]
Reed's criterion for means	0 (0, 8)[0, 16]	0 (0, 5)[0, 10]	0 (0, 1)[0, 2]	0 (0, 5)[0, 10]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	0 (0, 2)[0, 20]	0 (0, 1)[0, 13]	0 (0, 1)[0, 7]	0 (0, 5)[0, 35]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]
± 3SD	0 (0, 0)[0, 7]	0 (0, 0)[0, 4]	0 (0, 0)[0, 3]	0 (0, 0)[0, 15]

Table D.2: Increased CV_A: outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-1.981**	**-5.513**	**-33.393**	**-0.265**	**-0.553**	**-1.686**	**-3.364**	**-4.693**	**-9.682**
Cochran C test	-14.960	-9.415	-32.597	-1.997	-0.944	-1.646	-24.651	-7.641	-9.437
Cochran C test partial	-16.036	-3.017	-32.763	-2.141	-0.302	-1.654	-26.419	-2.545	-9.494
Fraser-Harris method	-16.175	-4.729	-73.572	-2.160	-0.474	-3.715	-26.430	-3.959	-20.954
Reed's criterion for means	-2.178	-7.307	-74.139	-0.291	-0.732	-3.744	-3.667	-6.164	-21.152
Reed's criterion for measurements	-1.981	-5.513	-33.393	-0.265	-0.553	-1.686	-3.364	-4.693	-9.682
Tukey IQR rule	-4.992	-20.381	-88.199	-0.667	-2.043	-4.454	-8.397	-17.088	-24.778
Dixon's Q test	-1.981	-5.513	-33.393	-0.265	-0.553	-1.686	-3.364	-4.693	-9.682
Grubbs's test	-2.099	-6.015	-34.139	-0.280	-0.603	-1.724	-3.564	-5.114	-9.893
± 3SD	-2.940	-9.895	-42.425	-0.393	-0.992	-2.142	-4.988	-8.398	-12.231

Table D.3: Increased CV_A: outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.347**	**1.383**	**12.007**	**0.954**	**0.955**	**0.952**	**0.024**	**0.048**	**0.148**
Cochran C test	0.391	1.527	12.039	0.938	0.938	0.951	0.023	0.048	0.148
Cochran C test partial	0.394	1.407	12.016	0.936	0.949	0.951	0.023	0.048	0.148
Fraser-Harris method	0.401	1.429	12.869	0.938	0.949	0.943	0.023	0.048	0.147
Reed's criterion for means	0.353	1.411	12.835	0.953	0.956	0.944	0.024	0.049	0.147
Reed's criterion for measurements	0.347	1.383	12.007	0.954	0.955	0.952	0.024	0.048	0.148
Tukey IQR rule	0.356	1.464	13.448	0.949	0.949	0.935	0.024	0.048	0.144
Dixon's Q test	0.347	1.383	12.007	0.954	0.955	0.952	0.024	0.048	0.148
Grubbs's test	0.347	1.387	12.025	0.955	0.955	0.952	0.024	0.048	0.148
± 3SD	0.348	1.398	12.212	0.952	0.955	0.948	0.024	0.048	0.147

Table D.4: Increased CV_A: outlier detection methods with no outlier simulation–SD, Median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.075, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.07 (0.07, 0.08)[0.05, 0.10]**	**0.10 (0.09, 0.11)[0.05, 0.14]**	**0.19 (0.17, 0.22)[0.06, 0.31]**
Cochran C test	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]
Cochran C test partial	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]
Fraser-Harris method	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.21)[0.06, 0.31]
Reed's criterion for means	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.21)[0.06, 0.31]
Reed's criterion for measurements	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]
Tukey IQR rule	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.21)[0.06, 0.31]
Dixon's Q test	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]
Grubbs's test	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]
± 3SD	0.07 (0.07, 0.08)[0.05, 0.10]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.06, 0.31]

Table D.5: Increased CV_A: outlier detection methods with no outlier simulation–CV, Median (Q1, Q3)[minimum, maximum]. True value of CV_A is 7.5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**7.47 (7.08, 7.88)[5.50, 9.68]**	**9.95 (9.15, 10.76)[4.88, 14.56]**	**19.61 (17.25, 22.03)[6.07, 32.13]**
Cochran C test	7.34 (6.93, 7.76)[5.32, 9.74]	9.93 (9.08, 10.76)[5.29, 13.89]	19.60 (17.25, 22.07)[5.88, 32.13]
Cochran C test partial	7.33 (6.92, 7.74)[5.28, 9.74]	9.98 (9.17, 10.79)[5.29, 13.89]	19.59 (17.25, 22.07)[5.88, 32.13]
Fraser-Harris method	7.33 (6.91, 7.74)[5.39, 9.74]	9.97 (9.13, 10.78)[5.29, 13.85]	19.20 (16.82, 21.64)[5.88, 31.51]
Reed's criterion for means	7.46 (7.07, 7.88)[5.47, 9.68]	9.93 (9.12, 10.75)[4.88, 14.56]	19.23 (16.80, 21.64)[6.07, 31.39]
Reed's criterion for measurements	7.47 (7.08, 7.88)[5.50, 9.68]	9.95 (9.15, 10.76)[4.88, 14.56]	19.61 (17.25, 22.03)[6.07, 32.13]
Tukey IQR rule	7.44 (7.05, 7.85)[5.25, 9.68]	9.81 (8.99, 10.59)[4.88, 14.26]	19.03 (16.63, 21.48)[6.08, 31.39]
Dixon's Q test	7.47 (7.08, 7.88)[5.50, 9.68]	9.95 (9.15, 10.76)[4.88, 14.56]	19.61 (17.25, 22.03)[6.07, 32.13]
Grubbs's test	7.47 (7.07, 7.88)[5.50, 9.68]	9.94 (9.14, 10.75)[4.88, 14.56]	19.60 (17.24, 22.02)[6.07, 32.13]
± 3SD	7.46 (7.07, 7.86)[5.45, 9.68]	9.91 (9.10, 10.71)[4.88, 14.56]	19.52 (17.17, 21.93)[6.08, 32.13]

Table D.6: Increased CV_A: outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.625, RCV is 34.65% and mean is 10. RCVs are displayed as percentages.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.64 (0.56, 0.73)[0.34, 2.03]**	**34.53 (32.81, 36.31)[25.10, 44.44]**	**9.97 (9.94, 10.01)[9.80, 10.13]**
Cochran C test	0.63 (0.55, 0.73)[0.34, 2.09]	34.24 (32.45, 36.13)[23.95, 44.10]	9.97 (9.94, 10.01)[9.80, 10.14]
Cochran C test partial	0.63 (0.56, 0.73)[0.34, 2.09]	34.34 (32.62, 36.18)[25.08, 44.10]	9.97 (9.94, 10.01)[9.80, 10.14]
Fraser-Harris method	0.65 (0.56, 0.75)[0.34, 2.09]	34.33 (32.57, 36.17)[25.04, 44.10]	9.97 (9.94, 10.01)[9.81, 10.16]
Reed's criterion for means	0.65 (0.57, 0.75)[0.34, 2.03]	34.49 (32.76, 36.29)[24.93, 44.44]	9.97 (9.94, 10.01)[9.81, 10.16]
Reed's criterion for measurements	0.64 (0.56, 0.73)[0.34, 2.03]	34.53 (32.81, 36.31)[25.10, 44.44]	9.97 (9.94, 10.01)[9.80, 10.13]
Tukey IQR rule	0.65 (0.57, 0.75)[0.34, 1.90]	34.16 (32.43, 35.94)[25.10, 44.05]	9.97 (9.94, 10.01)[9.80, 10.13]
Dixon's Q test	0.64 (0.56, 0.73)[0.34, 2.03]	34.53 (32.81, 36.31)[25.10, 44.44]	9.97 (9.94, 10.01)[9.80, 10.13]
Grubbs's test	0.64 (0.56, 0.73)[0.34, 2.03]	34.52 (32.79, 36.31)[25.10, 44.44]	9.97 (9.94, 10.01)[9.81, 10.13]
± 3SD	0.64 (0.56, 0.73)[0.34, 1.90]	34.41 (32.70, 36.19)[25.10, 44.44]	9.97 (9.94, 10.01)[9.80, 10.13]

Table D.7: Increased CV_A: outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -29.19% and upper bound is +41.22%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-29.11 (-30.34, -27.89)[-35.70, -22.16]**	**41.06 (38.68, 43.56)[28.46, 55.52]**
Cochran C test	-28.90 (-30.22, -27.63)[-35.48, -21.26]	40.65 (38.18, 43.30)[27.00, 55.00]
Cochran C test partial	-28.97 (-30.25, -27.75)[-35.48, -22.14]	40.79 (38.41, 43.37)[28.44, 55.00]
Fraser-Harris method	-28.96 (-30.24, -27.71)[-35.48, -22.11]	40.77 (38.34, 43.35)[28.39, 55.00]
Reed's criterion for means	-29.07 (-30.33, -27.85)[-35.70, -22.03]	40.99 (38.60, 43.53)[28.25, 55.52]
Reed's criterion for measurements	-29.11 (-30.34, -27.89)[-35.70, -22.16]	41.06 (38.68, 43.56)[28.46, 55.52]
Tukey IQR rule	-28.84 (-30.09, -27.62)[-35.45, -22.16]	40.54 (38.15, 43.04)[28.46, 54.93]
Dixon's Q test	-29.11 (-30.34, -27.89)[-35.70, -22.16]	41.06 (38.68, 43.56)[28.46, 55.52]
Grubbs's test	-29.10 (-30.34, -27.87)[-35.70, -22.16]	41.04 (38.65, 43.55)[28.46, 55.52]
± 3SD	-29.02 (-30.26, -27.81)[-35.70, -22.16]	40.89 (38.52, 43.39)[28.46, 55.52]

Table D.8: Increased CV_I: outlier detection methods with no outlier simulation–outliers removed by each detection method; median (Q1, Q3) [minimum, maximum]. Maximum number of measurements is 320 and maximum number of individuals is 40.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	2 (0, 4)[0, 28]	1 (0, 3)[0, 18]	1 (0, 2)[0, 6]	5 (0, 10)[0, 30]
Cochran C test partial	2 (0, 4)[0, 12]	1 (0, 3)[0, 8]	1 (0, 2)[0, 6]	5 (0, 10)[0, 30]
Fraser-Harris method	2 (2, 8)[0, 22]	1 (1, 5)[0, 14]	1 (1, 2)[0, 7]	5 (5, 10)[0, 35]
Reed's criterion for means	0 (0, 8)[0, 16]	0 (0, 5)[0, 10]	0 (0, 1)[0, 2]	0 (0, 5)[0, 10]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	0 (0, 2)[0, 16]	0 (0, 1)[0, 10]	0 (0, 1)[0, 6]	0 (0, 5)[0, 30]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]
± 3SD	0 (0, 0)[0, 5]	0 (0, 0)[0, 3]	0 (0, 0)[0, 2]	0 (0, 0)[0, 10]

Table D.9: Increased CV_I: outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-0.984**	**-5.900**	**-36.921**	**-0.197**	**-0.396**	**-1.864**	**-2.518**	**-4.100**	**-9.938**
Cochran C test	-9.072	-15.264	-35.824	-1.816	-1.023	-1.809	-22.478	-10.207	-9.582
Cochran C test partial	-10.175	-7.992	-35.647	-2.036	-0.536	-1.800	-25.163	-5.483	-9.570
Fraser-Harris method	-10.364	-10.839	-74.148	-2.074	-0.727	-3.744	-25.492	-7.384	-19.591
Reed's criterion for means	-1.150	-8.818	-75.580	-0.230	-0.591	-3.816	-2.922	-6.084	-20.031
Reed's criterion for measurements	-0.984	-5.900	-36.921	-0.197	-0.396	-1.864	-2.518	-4.100	-9.938
Tukey IQR rule	-2.013	-28.842	-103.334	-0.403	-1.934	-5.218	-5.122	-19.753	-26.934
Dixon's Q test	-0.984	-5.900	-36.921	-0.197	-0.396	-1.864	-2.518	-4.100	-9.938
Grubbs's test	-1.009	-6.286	-37.241	-0.202	-0.421	-1.880	-2.581	-4.370	-10.022
± 3SD	-1.246	-12.262	-49.829	-0.249	-0.822	-2.516	-3.187	-8.519	-13.329

Table D.10: Increased CV_I: outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.153**	**2.074**	**13.940**	**0.953**	**0.953**	**0.950**[a]	**0.016**	**0.058**	**0.158**[a]
Cochran C test	0.171	2.260	14.106	0.938	0.944	0.948[a]	0.016	0.058	0.158[a]
Cochran C test partial	0.174	2.131	14.002	0.936	0.949	0.944[a]	0.016	0.058	0.156[a]
Fraser-Harris method	0.176	2.166	14.874	0.937	0.949	0.944[b]	0.016	0.058	0.156[b]
Reed's criterion for means	0.155	2.109	14.808	0.952	0.953	0.946[c]	0.016	0.058	0.156[c]
Reed's criterion for measurements	0.153	2.074	13.940	0.953	0.953	0.950[a]	0.016	0.058	0.158[a]
Tukey IQR rule	0.155	2.215	15.787	0.951	0.946	0.933[b]	0.016	0.057	0.154[b]
Dixon's Q test	0.153	2.074	13.940	0.953	0.953	0.950[a]	0.016	0.058	0.158[a]
Grubbs's test	0.153	2.074	13.945	0.952	0.953	0.950[a]	0.016	0.058	0.158[a]
± 3SD	0.153	2.087	14.224	0.952	0.952	0.948[d]	0.016	0.058	0.157[d]

[a]1 CIs could not be calculated;[b]5 CIs could not be calculated, [c]3 CIs could not be calculated, [d]2 CIs could not be calculated.

Table D.11: Increased CV_I: outlier detection methods with no outlier simulation–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.15 and σ_G is 0.2.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.05 (0.05, 0.05)[0.04, 0.06]**	**0.10 (0.09, 0.11)[0.06, 0.14]**	**0.19 (0.17, 0.22)[0.08, 0.33]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.07, 0.33]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.07, 0.33]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.07, 0.33]
Reed's criterion for means	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.07, 0.33]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.06, 0.14]	0.19 (0.16, 0.21)[0.07, 0.32]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.33]

Table D.12: Increased CV_I: outlier detection methods with no outlier simulation–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**4.99 (4.72, 5.25)[3.53, 6.40]**	**9.94 (9.26, 10.65)[6.29, 13.72]**	**19.58 (17.21, 22.01)[7.55, 34.11]**
Cochran C test	4.91 (4.63, 5.18)[3.52, 6.48]	9.90 (9.18, 10.62)[6.02, 13.98]	19.59 (17.21, 22.03)[7.45, 34.20]
Cochran C test partial	4.90 (4.62, 5.17)[3.50, 6.40]	9.94 (9.25, 10.65)[6.24, 13.83]	19.59 (17.24, 22.03)[7.45, 34.20]
Fraser-Harris method	4.89 (4.61, 5.17)[3.50, 6.40]	9.94 (9.23, 10.65)[6.24, 14.12]	19.15 (16.73, 21.67)[7.45, 34.20]
Reed's criterion for means	4.98 (4.72, 5.25)[3.53, 6.40]	9.94 (9.25, 10.65)[6.29, 14.00]	19.15 (16.72, 21.67)[7.31, 34.11]
Reed's criterion for measurements	4.99 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.58 (17.21, 22.01)[7.55, 34.11]
Tukey IQR rule	4.97 (4.70, 5.24)[3.53, 6.37]	9.83 (9.15, 10.52)[6.29, 13.67]	19.00 (16.57, 21.51)[6.64, 32.74]
Dixon's Q test	4.99 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.58 (17.21, 22.01)[7.55, 34.11]
Grubbs's test	4.98 (4.72, 5.25)[3.53, 6.40]	9.94 (9.26, 10.65)[6.29, 13.72]	19.57 (17.21, 22.01)[7.55, 34.11]
± 3SD	4.98 (4.72, 5.25)[3.53, 6.40]	9.91 (9.23, 10.61)[6.29, 13.69]	19.48 (17.15, 21.93)[7.55, 34.11]

Table D.13: Increased CV_I: outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.79, RCV is 43.83% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.57 (0.50, 0.65)[0.31, 1.48]**	**30.86 (29.16, 32.63)[22.33, 40.37]**	**9.97 (9.94, 10.01)[9.82, 10.14]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.51]	30.65 (28.88, 32.47)[21.93, 40.59]	9.97 (9.94, 10.01)[9.82, 10.17]
Cochran C test partial	0.57 (0.50, 0.65)[0.30, 1.51]	30.74 (29.02, 32.56)[22.04, 40.59]	9.97 (9.94, 10.01)[9.82, 10.14]
Fraser-Harris method	0.58 (0.50, 0.67)[0.30, 1.65]	30.73 (28.97, 32.52)[22.04, 40.87]	9.98 (9.94, 10.01)[9.80, 10.16]
Reed's criterion for means	0.58 (0.51, 0.67)[0.31, 1.69]	30.83 (29.10, 32.61)[22.33, 40.87]	9.98 (9.94, 10.01)[9.81, 10.15]
Reed's criterion for measurements	0.57 (0.50, 0.65)[0.31, 1.48]	30.86 (29.16, 32.63)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
Tukey IQR rule	0.58 (0.51, 0.67)[0.31, 1.65]	30.53 (28.87, 32.33)[22.33, 39.66]	9.97 (9.94, 10.01)[9.82, 10.15]
Dixon's Q test	0.57 (0.50, 0.65)[0.31, 1.48]	30.86 (29.16, 32.63)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
Grubbs's test	0.57 (0.50, 0.65)[0.31, 1.48]	30.85 (29.15, 32.62)[22.33, 40.37]	9.97 (9.94, 10.01)[9.82, 10.14]
± 3SD	0.57 (0.50, 0.66)[0.31, 1.48]	30.76 (29.05, 32.56)[22.33, 39.96]	9.97 (9.94, 10.01)[9.82, 10.14]

Table D.14: Increased CV_I: outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -35.31% and upper bound is +54.58%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-26.48 (-27.76, -25.23)[-33.07, -19.98]**	**36.02 (33.75, 38.43)[24.97, 49.42]**
Cochran C test	-26.33 (-27.65, -25.02)[-33.22, -19.66]	35.74 (33.37, 38.21)[24.48, 49.74]
Cochran C test partial	-26.39 (-27.71, -25.13)[-33.22, -19.76]	35.86 (33.56, 38.33)[24.62, 49.74]
Fraser-Harris method	-26.38 (-27.68, -25.09)[-33.41, -19.76]	35.84 (33.50, 38.28)[24.62, 50.16]
Reed's criterion for means	-26.46 (-27.74, -25.19)[-33.41, -19.98]	35.98 (33.66, 38.39)[24.97, 50.16]
Reed's criterion for measurements	-26.48 (-27.76, -25.23)[-33.07, -19.98]	36.02 (33.75, 38.43)[24.97, 49.42]
Tukey IQR rule	-26.24 (-27.54, -25.02)[-32.60, -19.98]	35.57 (33.36, 38.02)[24.97, 48.37]
Dixon's Q test	-26.48 (-27.76, -25.23)[-33.07, -19.98]	36.02 (33.75, 38.43)[24.97, 49.42]
Grubbs's test	-26.48 (-27.75, -25.23)[-33.07, -19.98]	36.02 (33.74, 38.41)[24.97, 49.42]
± 3SD	-26.41 (-27.71, -25.15)[-32.80, -19.98]	35.89 (33.61, 38.33)[24.97, 48.82]

Table D.15: Increased CV_G: outlier detection methods with no outlier simulation–outliers removed by each detection method; median (Q1, Q3) [minimum, maximum]. Maximum number of measurements is 160 and maximum number of individuals is 20.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	2 (0, 4)[0, 26]	1 (0, 3)[0, 16]	1 (0, 2)[0, 6]	5 (0, 10)[0, 30]
Cochran C test partial	2 (0, 4)[0, 12]	1 (0, 3)[0, 8]	1 (0, 2)[0, 6]	5 (0, 10)[0, 30]
Fraser-Harris method	4 (2, 8)[0, 20]	3 (1, 5)[0, 13]	1 (1, 2)[0, 7]	5 (5, 10)[0, 35]
Reed's criterion for means	0 (0, 8)[0, 16]	0 (0, 5)[0, 10]	0 (0, 1)[0, 2]	0 (0, 5)[0, 10]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	0 (0, 3)[0, 26]	0 (0, 2)[0, 16]	0 (0, 1)[0, 6]	0 (0, 5)[0, 30]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]
± 3SD	0 (0, 0)[0, 8]	0 (0, 0)[0, 5]	0 (0, 0)[0, 2]	0 (0, 0)[0, 10]

Table D.16: Increased CV_G: outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-2.137**	**-5.691**	**-18.633**	**-0.428**	**-0.571**	**-0.635**	**-5.402**	**-5.469**	**-3.776**
Cochran C test	-10.448	-11.423	-18.674	-2.091	-1.145	-0.636	-25.577	-10.495	-3.768
Cochran C test partial	-11.546	-6.188	-18.523	-2.311	-0.620	-0.631	-28.242	-5.865	-3.751
Fraser-Harris method	-11.586	-6.665	-85.085	-2.319	-0.668	-2.898	-28.106	-6.272	-16.947
Reed's criterion for means	-2.157	-6.104	-85.492	-0.432	-0.612	-2.912	-5.408	-5.826	-17.060
Reed's criterion for measurements	-2.137	-5.691	-18.633	-0.428	-0.571	-0.635	-5.402	-5.469	-3.776
Tukey IQR rule	-3.084	-13.281	-103.731	-0.617	-1.331	-3.534	-7.745	-12.585	-19.899
Dixon's Q test	-2.137	-5.691	-18.633	-0.428	-0.571	-0.635	-5.402	-5.469	-3.776
Grubbs's test	-2.153	-5.803	-18.792	-0.431	-0.582	-0.640	-5.441	-5.577	-3.808
± 3SD	-2.335	-7.651	-25.544	-0.467	-0.767	-0.870	-5.901	-7.343	-5.158

Table D.17: Increased CV_G: outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.157**	**1.086**	**24.381**	**0.949**	**0.950**	**0.947**	**0.016**	**0.041**	**0.211**
Cochran C test	0.178	1.198	24.592	0.936	0.938	0.947	0.016	0.041	0.211
Cochran C test partial	0.180	1.117	24.420	0.934	0.947	0.947	0.015	0.041	0.211
Fraser-Harris method	0.183	1.134	25.931	0.933	0.946	0.945	0.016	0.042	0.208
Reed's criterion for means	0.160	1.102	25.843	0.948	0.950	0.946	0.016	0.042	0.208
Reed's criterion for measurements	0.157	1.086	24.381	0.949	0.950	0.947	0.016	0.041	0.211
Tukey IQR rule	0.160	1.131	28.251	0.948	0.947	0.932	0.016	0.041	0.205
Dixon's Q test	0.157	1.086	24.381	0.949	0.950	0.947	0.016	0.041	0.211
Grubbs's test	0.157	1.086	24.389	0.950	0.950	0.947	0.016	0.041	0.211
± 3SD	0.157	1.091	24.593	0.949	0.950	0.947	0.016	0.041	0.210

Table D.18: Increased CV_G: outlier detection methods with no outlier simulation–SD, Median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.29.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.05 (0.05, 0.05)[0.04, 0.06]**	**0.10 (0.09, 0.11)[0.06, 0.13]**	**0.29 (0.26, 0.32)[0.13, 0.49]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.29 (0.26, 0.32)[0.12, 0.50]
Cochran C test partial	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.29 (0.26, 0.32)[0.12, 0.50]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.28 (0.25, 0.32)[0.12, 0.50]
Reed's criterion for means	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.28 (0.25, 0.32)[0.12, 0.49]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.29 (0.26, 0.32)[0.13, 0.49]
Tukey IQR rule	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.28 (0.25, 0.32)[0.11, 0.49]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.29 (0.26, 0.32)[0.13, 0.49]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.29 (0.26, 0.32)[0.13, 0.49]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.29 (0.26, 0.32)[0.13, 0.49]

Table D.19: Increased CV_G: outlier detection methods with no outlier simulation–CV, Median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 30%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**4.97 (4.71, 5.25)[3.54, 6.23]**	**9.93 (9.21, 10.65)[5.68, 13.45]**	**29.58 (26.25, 33.33)[12.68, 52.00]**
Cochran C test	4.89 (4.61, 5.17)[3.50, 6.22]	9.88 (9.16, 10.63)[5.70, 13.66]	29.59 (26.22, 33.36)[12.36, 53.11]
Cochran C test partial	4.88 (4.61, 5.16)[3.50, 6.22]	9.92 (9.22, 10.65)[5.77, 13.68]	29.61 (26.23, 33.36)[12.36, 52.93]
Fraser-Harris method	4.88 (4.60, 5.16)[3.44, 6.28]	9.91 (9.21, 10.66)[5.81, 13.68]	28.97 (25.42, 32.73)[12.18, 52.93]
Reed's criterion for means	4.97 (4.70, 5.25)[3.54, 6.27]	9.92 (9.21, 10.65)[5.79, 13.66]	28.97 (25.43, 32.73)[12.17, 52.00]
Reed's criterion for measurements	4.97 (4.71, 5.25)[3.54, 6.23]	9.93 (9.21, 10.65)[5.68, 13.45]	29.58 (26.25, 33.33)[12.68, 52.00]
Tukey IQR rule	4.96 (4.69, 5.24)[3.50, 6.29]	9.85 (9.14, 10.58)[5.68, 13.76]	28.80 (25.10, 32.67)[10.92, 52.00]
Dixon's Q test	4.97 (4.71, 5.25)[3.54, 6.23]	9.93 (9.21, 10.65)[5.68, 13.45]	29.58 (26.25, 33.33)[12.68, 52.00]
Grubbs's test	4.97 (4.71, 5.25)[3.54, 6.23]	9.93 (9.21, 10.65)[5.68, 13.45]	29.58 (26.25, 33.33)[12.68, 52.00]
± 3SD	4.97 (4.70, 5.24)[3.53, 6.23]	9.91 (9.20, 10.63)[5.68, 13.45]	29.52 (26.18, 33.25)[12.68, 52.00]

Table D.20: Increased CV_G: outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.373, RCV is 30.99% and mean is 10. RCVs are displayed as percentages.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.37 (0.33, 0.43)[0.20, 0.84]**	**30.81 (29.06, 32.60)[20.70, 39.65]**	**9.95 (9.91, 10.00)[9.70, 10.18]**
Cochran C test	0.37 (0.33, 0.42)[0.18, 0.84]	30.57 (28.79, 32.45)[20.54, 40.10]	9.95 (9.91, 10.00)[9.70, 10.21]
Cochran C test partial	0.37 (0.33, 0.43)[0.19, 0.84]	30.68 (28.90, 32.49)[20.54, 40.12]	9.95 (9.91, 10.00)[9.70, 10.18]
Fraser-Harris method	0.38 (0.33, 0.44)[0.19, 0.93]	30.67 (28.89, 32.49)[20.90, 40.12]	9.95 (9.90, 10.00)[9.71, 10.20]
Reed's criterion for means	0.38 (0.34, 0.44)[0.20, 0.93]	30.80 (29.03, 32.58)[21.05, 39.69]	9.95 (9.90, 10.00)[9.71, 10.20]
Reed's criterion for measurements	0.37 (0.33, 0.43)[0.20, 0.84]	30.81 (29.06, 32.60)[20.70, 39.65]	9.95 (9.91, 10.00)[9.70, 10.18]
Tukey IQR rule	0.38 (0.34, 0.44)[0.21, 0.99]	30.59 (28.84, 32.41)[20.70, 39.96]	9.95 (9.91, 10.00)[9.69, 10.18]
Dixon's Q test	0.37 (0.33, 0.43)[0.20, 0.84]	30.81 (29.06, 32.60)[20.70, 39.65]	9.95 (9.91, 10.00)[9.70, 10.18]
Grubbs's test	0.37 (0.33, 0.43)[0.20, 0.84]	30.81 (29.05, 32.59)[20.70, 39.65]	9.95 (9.91, 10.00)[9.70, 10.18]
± 3SD	0.37 (0.33, 0.43)[0.20, 0.84]	30.74 (28.99, 32.54)[20.70, 39.65]	9.95 (9.91, 10.00)[9.71, 10.18]

Table D.21: Increased CV_G: outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-26.45 (-27.74, -25.16)[-32.60, -18.68]**	**35.96 (33.61, 38.39)[22.97, 48.36]**
Cochran C test	-26.27 (-27.63, -24.96)[-32.89, -18.55]	35.64 (33.25, 38.18)[22.77, 49.02]
Cochran C test partial	-26.35 (-27.66, -25.04)[-32.91, -18.55]	35.78 (33.40, 38.24)[22.77, 49.04]
Fraser-Harris method	-26.35 (-27.66, -25.03)[-32.91, -18.83]	35.77 (33.39, 38.24)[23.20, 49.04]
Reed's criterion for means	-26.44 (-27.73, -25.14)[-32.63, -18.96]	35.95 (33.58, 38.36)[23.39, 48.43]
Reed's criterion for measurements	-26.45 (-27.74, -25.16)[-32.60, -18.68]	35.96 (33.61, 38.39)[22.97, 48.36]
Tukey IQR rule	-26.29 (-27.60, -24.99)[-32.80, -18.68]	35.66 (33.32, 38.12)[22.97, 48.82]
Dixon's Q test	-26.45 (-27.74, -25.16)[-32.60, -18.68]	35.96 (33.61, 38.39)[22.97, 48.36]
Grubbs's test	-26.45 (-27.73, -25.15)[-32.60, -18.68]	35.95 (33.61, 38.38)[22.97, 48.36]
± 3SD	-26.40 (-27.70, -25.10)[-32.60, -18.68]	35.87 (33.52, 38.30)[22.97, 48.36]

Table D.22: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–outliers removed by each detection method; median (Q1, Q3) [minimum, maximum]. Maximum number of measurements is 160 and maximum number of individuals is 20.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	2 (0, 4)[0, 26]	1 (0, 3)[0, 16]	1 (0, 2)[0, 5]	5 (0, 10)[0, 25]
Cochran C test partial	2 (0, 4)[0, 10]	1 (0, 3)[0, 6]	1 (0, 2)[0, 5]	5 (0, 10)[0, 25]
Fraser-Harris method	4 (2, 8)[0, 22]	3 (1, 5)[0, 14]	1 (1, 2)[0, 6]	5 (5, 10)[0, 30]
Reed's criterion for means	0 (0, 8)[0, 16]	0 (0, 5)[0, 10]	0 (0, 1)[0, 2]	0 (0, 5)[0, 10]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	0 (0, 2)[0, 22]	0 (0, 1)[0, 14]	0 (0, 1)[0, 6]	0 (0, 5)[0, 30]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 1]	0 (0, 0)[0, 5]
± 3SD	0 (0, 0)[0, 5]	0 (0, 0)[0, 3]	0 (0, 0)[0, 2]	0 (0, 0)[0, 10]

Table D.23: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	**Bias ($\times 10^{-4}$)**			**Bias Percentage bias**			**Standardised bias**		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-2.096**	**-9.559**	**-35.933**	**-0.280**	**-0.641**	**-1.224**	**-3.622**	**-6.244**	**-6.969**
Cochran C test	-14.982	-16.181	-36.318	-2.000	-1.085	-1.237	-25.103	-10.186	-7.018
Cochran C test partial	-16.097	-9.033	-35.255	-2.149	-0.606	-1.201	-26.972	-5.856	-6.821
Fraser-Harris method	-16.095	-11.294	-99.356	-2.149	-0.757	-3.385	-26.751	-7.253	-19.025
Reed's criterion for means	-2.150	-11.887	-99.810	-0.287	-0.797	-3.400	-3.684	-7.691	-19.148
Reed's criterion for measurements	-2.096	-9.559	-35.933	-0.280	-0.641	-1.224	-3.622	-6.244	-6.969
Tukey IQR rule	-3.832	-25.736	-114.342	-0.512	-1.725	-3.895	-6.583	-16.617	-21.486
Dixon's Q test	-2.096	-9.559	-35.933	-0.280	-0.641	-1.224	-3.622	-6.244	-6.969
Grubbs's test	-2.128	-9.948	-36.364	-0.284	-0.667	-1.239	-3.677	-6.496	-7.050
± 3SD	-2.516	-14.116	-46.415	-0.336	-0.946	-1.581	-4.341	-9.186	-8.972

Table D.24: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	**Mean squared error ($\times 10^{-4}$)**			**Accuracy and coverage Coverage**			**Mean 95% CI width**		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.335**	**2.353**	**26.716**	**0.955**	**0.949**	**0.946**	**0.024**	**0.062**	**0.218**
Cochran C test	0.379	2.550	26.912	0.936	0.940	0.944	0.023	0.062	0.218
Cochran C test partial	0.382	2.388	26.835	0.935	0.947	0.945	0.023	0.062	0.218
Fraser-Harris method	0.388	2.437	28.259	0.934	0.947	0.942	0.023	0.062	0.216
Reed's criterion for means	0.341	2.403	28.166	0.954	0.948	0.942	0.024	0.062	0.216
Reed's criterion for measurements	0.335	2.353	26.716	0.955	0.949	0.946	0.024	0.062	0.218
Tukey IQR rule	0.340	2.465	29.627	0.954	0.943	0.936	0.024	0.062	0.213
Dixon's Q test	0.335	2.353	26.716	0.955	0.949	0.946	0.024	0.062	0.218
Grubbs's test	0.335	2.355	26.741	0.955	0.949	0.946	0.024	0.062	0.218
± 3SD	0.337	2.381	26.981	0.954	0.947	0.944	0.024	0.062	0.217

Table D.25: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.075, σ_I is 0.15 and σ_G is 0.29.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.07 (0.07, 0.08)[0.06, 0.10]**	**0.15 (0.14, 0.16)[0.09, 0.20]**	**0.29 (0.25, 0.32)[0.11, 0.48]**
Cochran C test	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.21]	0.29 (0.25, 0.32)[0.11, 0.47]
Cochran C test partial	0.07 (0.07, 0.08)[0.05, 0.10]	0.15 (0.14, 0.16)[0.09, 0.21]	0.29 (0.25, 0.32)[0.11, 0.47]
Fraser-Harris method	0.07 (0.07, 0.08)[0.05, 0.10]	0.15 (0.14, 0.16)[0.09, 0.21]	0.28 (0.25, 0.32)[0.10, 0.47]
Reed's criterion for means	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.28 (0.25, 0.32)[0.11, 0.47]
Reed's criterion for measurements	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.29 (0.25, 0.32)[0.11, 0.48]
Tukey IQR rule	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.28 (0.25, 0.32)[0.10, 0.48]
Dixon's Q test	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.29 (0.25, 0.32)[0.11, 0.48]
Grubbs's test	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.29 (0.25, 0.32)[0.11, 0.48]
± 3SD	0.07 (0.07, 0.08)[0.06, 0.10]	0.15 (0.14, 0.16)[0.09, 0.20]	0.29 (0.25, 0.32)[0.11, 0.48]

Table D.26: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 7.5%, CV_I is 15% and CV_G is 30%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**7.48 (7.08, 7.86)[5.52, 9.77]**	**14.89 (13.88, 15.95)[8.67, 20.37]**	**29.49 (25.91, 33.19)[10.92, 50.61]**
Cochran C test	7.35 (6.95, 7.74)[5.51, 9.54]	14.84 (13.80, 15.89)[8.67, 20.86]	29.52 (25.91, 33.19)[10.92, 50.30]
Cochran C test partial	7.34 (6.94, 7.73)[5.48, 9.54]	14.90 (13.88, 15.94)[8.67, 20.80]	29.51 (25.91, 33.23)[10.92, 50.30]
Fraser-Harris method	7.34 (6.94, 7.73)[5.48, 9.54]	14.89 (13.85, 15.93)[8.67, 21.05]	28.86 (25.24, 32.60)[10.23, 49.47]
Reed's criterion for means	7.48 (7.08, 7.87)[5.52, 9.75]	14.87 (13.85, 15.94)[8.67, 20.49]	28.87 (25.23, 32.61)[10.92, 49.47]
Reed's criterion for measurements	7.48 (7.08, 7.86)[5.52, 9.77]	14.89 (13.88, 15.95)[8.67, 20.37]	29.49 (25.91, 33.19)[10.92, 50.61]
Tukey IQR rule	7.46 (7.07, 7.85)[5.52, 9.77]	14.71 (13.70, 15.80)[8.67, 20.37]	28.71 (25.04, 32.52)[9.89, 50.61]
Dixon's Q test	7.48 (7.08, 7.86)[5.52, 9.77]	14.89 (13.88, 15.95)[8.67, 20.37]	29.49 (25.91, 33.19)[10.92, 50.61]
Grubbs's test	7.48 (7.08, 7.86)[5.52, 9.77]	14.89 (13.87, 15.94)[8.67, 20.37]	29.48 (25.91, 33.18)[10.92, 50.61]
± 3SD	7.47 (7.08, 7.86)[5.52, 9.77]	14.84 (13.82, 15.89)[8.67, 20.37]	29.38 (25.83, 33.10)[10.92, 50.61]

Table D.27: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 46.49% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.56 (0.50, 0.65)[0.29, 1.60]**	**46.20 (43.75, 48.85)[32.56, 60.33]**	**9.94 (9.89, 9.99)[9.69, 10.18]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.60]	45.94 (43.34, 48.50)[30.26, 61.24]	9.94 (9.89, 9.99)[9.69, 10.18]
Cochran C test partial	0.56 (0.49, 0.65)[0.29, 1.60]	46.05 (43.53, 48.66)[32.61, 61.04]	9.94 (9.89, 9.99)[9.69, 10.18]
Fraser-Harris method	0.58 (0.50, 0.67)[0.30, 1.76]	46.01 (43.44, 48.65)[32.61, 61.54]	9.94 (9.89, 9.99)[9.67, 10.18]
Reed's criterion for means	0.58 (0.50, 0.67)[0.30, 1.75]	46.17 (43.61, 48.80)[32.56, 60.54]	9.94 (9.89, 9.99)[9.67, 10.18]
Reed's criterion for measurements	0.56 (0.50, 0.65)[0.29, 1.60]	46.20 (43.75, 48.85)[32.56, 60.33]	9.94 (9.89, 9.99)[9.69, 10.18]
Tukey IQR rule	0.58 (0.50, 0.66)[0.30, 1.66]	45.79 (43.26, 48.41)[32.56, 60.33]	9.94 (9.89, 9.99)[9.69, 10.18]
Dixon's Q test	0.56 (0.50, 0.65)[0.29, 1.60]	46.20 (43.75, 48.85)[32.56, 60.33]	9.94 (9.89, 9.99)[9.69, 10.18]
Grubbs's test	0.56 (0.50, 0.65)[0.29, 1.60]	46.20 (43.72, 48.84)[32.56, 60.33]	9.94 (9.89, 9.99)[9.69, 10.18]
± 3SD	0.56 (0.50, 0.65)[0.29, 1.60]	46.09 (43.58, 48.73)[32.56, 60.33]	9.94 (9.89, 9.99)[9.69, 10.18]

Table D.28: Increased CV_A, CV_I and CV_G: outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -36.98% and upper bound is +58.67%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-36.80 (-38.42, -35.26)[-44.92, -27.71]**	**58.23 (54.46, 62.38)[38.34, 81.55]**
Cochran C test	-36.64 (-38.21, -35.00)[-45.40, -26.04]	57.83 (53.85, 61.83)[35.21, 83.15]
Cochran C test partial	-36.71 (-38.30, -35.12)[-45.29, -27.74]	58.00 (54.13, 62.07)[38.40, 82.80]
Fraser-Harris method	-36.68 (-38.30, -35.06)[-45.56, -27.74]	57.93 (54.00, 62.06)[38.40, 83.68]
Reed's criterion for means	-36.78 (-38.39, -35.17)[-45.03, -27.71]	58.18 (54.25, 62.30)[38.34, 81.91]
Reed's criterion for measurements	-36.80 (-38.42, -35.26)[-44.92, -27.71]	58.23 (54.46, 62.38)[38.34, 81.55]
Tukey IQR rule	-36.55 (-38.15, -34.95)[-44.92, -27.71]	57.59 (53.73, 61.68)[38.34, 81.55]
Dixon's Q test	-36.80 (-38.42, -35.26)[-44.92, -27.71]	58.23 (54.46, 62.38)[38.34, 81.55]
Grubbs's test	-36.80 (-38.41, -35.24)[-44.92, -27.71]	58.22 (54.43, 62.37)[38.34, 81.55]
± 3SD	-36.73 (-38.35, -35.15)[-44.92, -27.71]	58.05 (54.21, 62.20)[38.34, 81.55]

D.2 Outlier detection methods with outlier simulation–increased

n

Table D.29: Increased n: outlier detection methods with no outlier simulation–outliers removed by each detection method; median (Q1, Q3) [minimum, maximum]. Maximum number of measurements is 320 and maximum number of individuals is 40.

Outlier strategy	**Measurements removed**		**Individuals with measurements removed**	
	median (Q1, Q3) [minimum, maximum]			
	n	%	n	%
Cochran C test	6 (2, 8)[0, 28]	2 (1, 3)[0, 9]	2 (1, 3)[0, 9]	5 (3, 8)[0, 23]
Cochran C test partial	4 (2, 6)[0, 18]	1 (1, 2)[0, 6]	2 (1, 3)[0, 9]	5 (3, 8)[0, 23]
Fraser-Harris method	4 (2, 8)[0, 22]	1 (1, 3)[0, 7]	2 (1, 3)[0, 9]	5 (3, 8)[0, 23]
Reed's criterion for means	0 (0, 0)[0, 16]	0 (0, 0)[0, 5]	0 (0, 0)[0, 2]	0 (0, 0)[0, 5]
Reed's criterion for measurements	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Tukey IQR rule	2 (0, 4)[0, 25]	1 (0, 1)[0, 8]	1 (0, 2)[0, 8]	3 (0, 5)[0, 20]
Dixon's Q test	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]	0 (0, 0)[0, 0]
Grubbs's test	0 (0, 0)[0, 1]	0 (0, 0)[0, 0]	0 (0, 0)[0, 1]	0 (0, 0)[0, 3]
± 3SD	0 (0, 1)[0, 8]	0 (0, 0)[0, 3]	0 (0, 1)[0, 3]	0 (0, 3)[0, 8]

Table D.30: Increased n: outlier detection methods with no outlier simulation–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**-0.777**	**-3.033**	**-16.814**	**-0.156**	**-0.304**	**-0.849**	**-2.749**	**-4.064**	**-6.946**
Cochran C test	-8.684	-5.259	-16.218	-1.738	-0.527	-0.819	-29.846	-6.810	-6.682
Cochran C test partial	-9.661	-2.107	-16.498	-1.933	-0.211	-0.833	-33.146	-2.776	-6.810
Fraser-Harris method	-9.647	-2.497	-30.035	-1.931	-0.250	-1.517	-33.062	-3.284	-12.248
Reed's criterion for means	-0.774	-3.434	-30.511	-0.155	-0.344	-1.541	-2.735	-4.593	-12.443
Reed's criterion for measurements	-0.777	-3.033	-16.814	-0.156	-0.304	-0.849	-2.749	-4.064	-6.946
Tukey IQR rule	-1.855	-13.043	-63.927	-0.371	-1.308	-3.228	-6.528	-17.421	-25.525
Dixon's Q test	-0.777	-3.033	-16.814	-0.156	-0.304	-0.849	-2.749	-4.064	-6.946
Grubbs's test	-0.800	-3.176	-17.031	-0.160	-0.318	-0.860	-2.831	-4.258	-7.034
± 3SD	-1.164	-6.513	-28.258	-0.233	-0.653	-1.427	-4.118	-8.738	-11.605

Table D.31: Increased n: outlier detection methods with no outlier simulation–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**0.080**	**0.558**	**5.888**	**0.944**	**0.943**	**0.947**	**0.011**	**0.029**	**0.098**
Cochran C test	0.092	0.599	5.918	0.927	0.936	0.949	0.011	0.029	0.098
Cochran C test partial	0.094	0.576	5.896	0.925	0.940	0.949	0.011	0.029	0.098
Fraser-Harris method	0.094	0.579	6.103	0.926	0.940	0.946	0.011	0.029	0.098
Reed's criterion for means	0.080	0.560	6.105	0.944	0.943	0.944	0.011	0.029	0.098
Reed's criterion for measurements	0.080	0.558	5.888	0.944	0.943	0.947	0.011	0.029	0.098
Tukey IQR rule	0.081	0.578	6.681	0.945	0.939	0.932	0.011	0.029	0.096
Dixon's Q test	0.080	0.558	5.888	0.944	0.943	0.947	0.011	0.029	0.098
Grubbs's test	0.080	0.557	5.891	0.944	0.944	0.947	0.011	0.029	0.098
± 3SD	0.080	0.560	6.009	0.944	0.943	0.946	0.011	0.029	0.097

Table D.32: Increased n: outlier detection methods with no outlier simulation–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.05 (0.05, 0.05)[0.04, 0.06]**	**0.10 (0.09, 0.10)[0.07, 0.13]**	**0.20 (0.18, 0.21)[0.11, 0.29]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.11, 0.29]
Cochran C test partial	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.11, 0.29]
Fraser-Harris method	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.19 (0.18, 0.21)[0.11, 0.29]
Reed's criterion for means	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.19 (0.18, 0.21)[0.11, 0.29]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.11, 0.29]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.19 (0.17, 0.21)[0.10, 0.29]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.11, 0.29]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.20 (0.18, 0.21)[0.11, 0.29]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.07, 0.13]	0.19 (0.18, 0.21)[0.11, 0.29]

Table D.33: Increased n: outlier detection methods with no outlier simulation–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**4.99 (4.81, 5.19)[4.00, 6.19]**	**9.98 (9.47, 10.48)[7.21, 13.08]**	**19.79 (18.12, 21.49)[11.14, 29.60]**
Cochran C test	4.91 (4.71, 5.11)[3.94, 6.09]	9.96 (9.43, 10.48)[7.14, 13.20]	19.79 (18.14, 21.51)[11.16, 29.63]
Cochran C test partial	4.90 (4.70, 5.10)[3.94, 6.09]	9.99 (9.46, 10.50)[7.08, 13.19]	19.79 (18.15, 21.49)[11.15, 29.59]
Fraser-Harris method	4.90 (4.71, 5.10)[3.94, 6.09]	9.99 (9.46, 10.50)[7.08, 13.19]	19.66 (18.00, 21.39)[11.10, 29.59]
Reed's criterion for means	4.99 (4.80, 5.19)[4.00, 6.19]	9.97 (9.47, 10.48)[7.26, 13.08]	19.65 (17.99, 21.38)[11.01, 29.60]
Reed's criterion for measurements	4.99 (4.81, 5.19)[4.00, 6.19]	9.98 (9.47, 10.48)[7.21, 13.08]	19.79 (18.12, 21.49)[11.14, 29.60]
Tukey IQR rule	4.98 (4.79, 5.17)[4.00, 6.19]	9.88 (9.37, 10.39)[7.15, 13.00]	19.33 (17.60, 21.01)[10.35, 29.11]
Dixon's Q test	4.99 (4.81, 5.19)[4.00, 6.19]	9.98 (9.47, 10.48)[7.21, 13.08]	19.79 (18.12, 21.49)[11.14, 29.60]
Grubbs's test	4.99 (4.81, 5.19)[4.00, 6.19]	9.97 (9.47, 10.48)[7.21, 13.08]	19.79 (18.12, 21.49)[11.14, 29.60]
± 3SD	4.99 (4.80, 5.18)[4.01, 6.19]	9.94 (9.44, 10.45)[7.17, 13.08]	19.68 (18.01, 21.36)[11.14, 29.49]

Table D.34: Increased n: outlier detection methods with no outlier simulation–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**0.56 (0.51, 0.62)[0.35, 1.03]**	**30.95 (29.67, 32.19)[24.00, 38.33]**	**9.97 (9.95, 10.00)[9.85, 10.09]**
Cochran C test	0.56 (0.51, 0.62)[0.34, 1.04]	30.83 (29.47, 32.09)[23.72, 38.53]	9.97 (9.95, 10.00)[9.85, 10.10]
Cochran C test partial	0.56 (0.51, 0.62)[0.34, 1.03]	30.88 (29.54, 32.12)[23.56, 38.48]	9.97 (9.95, 10.00)[9.85, 10.10]
Fraser-Harris method	0.57 (0.51, 0.62)[0.34, 1.08]	30.87 (29.52, 32.11)[23.56, 38.48]	9.97 (9.95, 10.00)[9.85, 10.10]
Reed's criterion for means	0.57 (0.52, 0.63)[0.35, 1.08]	30.95 (29.66, 32.18)[24.10, 38.33]	9.97 (9.95, 10.00)[9.85, 10.09]
Reed's criterion for measurements	0.56 (0.51, 0.62)[0.35, 1.03]	30.95 (29.67, 32.19)[24.00, 38.33]	9.97 (9.95, 10.00)[9.85, 10.09]
Tukey IQR rule	0.57 (0.52, 0.63)[0.34, 1.08]	30.68 (29.43, 31.93)[23.82, 38.12]	9.97 (9.95, 10.00)[9.85, 10.09]
Dixon's Q test	0.56 (0.51, 0.62)[0.35, 1.03]	30.95 (29.67, 32.19)[24.00, 38.33]	9.97 (9.95, 10.00)[9.85, 10.09]
Grubbs's test	0.56 (0.51, 0.62)[0.35, 1.03]	30.95 (29.67, 32.18)[24.00, 38.33]	9.97 (9.95, 10.00)[9.85, 10.09]
± 3SD	0.57 (0.52, 0.62)[0.34, 1.03]	30.86 (29.61, 32.08)[23.85, 38.33]	9.97 (9.95, 10.00)[9.85, 10.09]

Table D.35: Increased n: outlier detection methods with no outlier simulation–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-26.55 (-27.44, -25.61)[-31.71, -21.30]**	**36.14 (34.43, 37.83)[27.07, 46.44]**
Cochran C test	-26.46 (-27.37, -25.46)[-31.85, -21.08]	35.98 (34.16, 37.68)[26.71, 46.73]
Cochran C test partial	-26.50 (-27.40, -25.51)[-31.82, -20.96]	36.05 (34.25, 37.74)[26.51, 46.66]
Fraser-Harris method	-26.49 (-27.39, -25.50)[-31.82, -20.96]	36.04 (34.23, 37.72)[26.51, 46.66]
Reed's criterion for means	-26.55 (-27.44, -25.60)[-31.71, -21.38]	36.14 (34.42, 37.81)[27.20, 46.44]
Reed's criterion for measurements	-26.55 (-27.44, -25.61)[-31.71, -21.30]	36.14 (34.43, 37.83)[27.07, 46.44]
Tukey IQR rule	-26.35 (-27.26, -25.43)[-31.57, -21.16]	35.78 (34.10, 37.47)[26.85, 46.14]
Dixon's Q test	-26.55 (-27.44, -25.61)[-31.71, -21.30]	36.14 (34.43, 37.83)[27.07, 46.44]
Grubbs's test	-26.55 (-27.44, -25.61)[-31.71, -21.30]	36.14 (34.43, 37.82)[27.07, 46.44]
± 3SD	-26.48 (-27.37, -25.57)[-31.71, -21.18]	36.03 (34.35, 37.68)[26.87, 46.44]

D.3 Outlier detection methods with outlier simulation

Table D.36: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**5447.728**	**-111.622**	**-107.027**	**1090.226**	**-11.190**	**-5.404**	**276.969**	**-19.279**	**-15.389**
Cochran C test	-8.924	-11.863	-32.001	-1.786	-1.189	-1.616	-21.512	-10.957	-9.622
Cochran C test partial	-9.797	-5.963	-32.368	-1.961	-0.598	-1.634	-23.657	-5.735	-9.730
Fraser-Harris method	-9.900	-7.355	-70.186	-1.981	-0.737	-3.544	-23.713	-7.037	-20.717
Reed's criterion for means	1973.328	-44.889	-65.560	394.912	-4.500	-3.310	69.754	-12.390	-13.263
Reed's criterion for measurements	-1.104	-5.357	-32.809	-0.221	-0.537	-1.657	-2.739	-5.296	-9.890
Tukey IQR rule	-2.399	-16.204	-84.458	-0.480	-1.624	-4.265	-5.934	-15.895	-24.483
Dixon's Q test	-1.104	-5.357	-32.809	-0.221	-0.537	-1.657	-2.739	-5.296	-9.890
Grubbs's test	-1.104	-5.357	-32.809	-0.221	-0.537	-1.657	-2.739	-5.296	-9.890
± 3SD	-1.104	-5.357	-32.809	-0.221	-0.537	-1.657	-2.739	-5.296	-9.890

Table D.37: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**3354.647**	**34.770**	**49.516**	**0.000**	$-^a$	**0.980**^{b}	**0.188**	$-^a$	**0.255**^{b}
Cochran C test	0.180	1.186	11.164	0.935	0.939	0.960	0.016	0.041	0.147
Cochran C test partial	0.181	1.085	11.171	0.934	0.950	0.958	0.016	0.041	0.146
Fraser-Harris method	0.184	1.098	11.970	0.934	0.949	0.951	0.016	0.042	0.145
Reed's criterion for means	1189.717	13.327	24.865	0.593	0.957^c	0.961^d	0.079	0.043^c	0.173^d
Reed's criterion for measurements	0.162	1.026	11.113	0.943	0.959	0.959	0.016	0.041	0.146
Tukey IQR rule	0.164	1.066	12.613	0.944	0.952	0.943	0.016	0.041	0.143
Dixon's Q test	0.162	1.026	11.113	0.943	0.959	0.959	0.016	0.041	0.146
Grubbs's test	0.162	1.026	11.113	0.943	0.959	0.959	0.016	0.041	0.146
± 3SD	0.162	1.026	11.113	0.943	0.959	0.959	0.016	0.041	0.146

[a]5000 CIs could not be calculated; [b]2776 CIs could not be calculated; [c]1865 CIs could not be calculated; [d]988 CIs could not be calculated.

Table D.38: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.74 (0.40, 0.79)[0.36, 0.86]**	**0.10 (0.04, 0.13)[0.00, 0.27]**	**0.19 (0.15, 0.23)[0.00, 0.41]**
Cochran C test	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.08, 0.34]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.08, 0.34]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.08, 0.34]
Reed's criterion for means	0.05 (0.05, 0.40)[0.04, 0.88]	0.10 (0.09, 0.11)[0.00, 0.27]	0.19 (0.17, 0.22)[0.00, 0.41]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.08, 0.34]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.10)[0.06, 0.13]	0.19 (0.17, 0.21)[0.08, 0.34]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.08, 0.34]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.08, 0.34]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.08, 0.34]

Table D.39: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**85.25 (41.29, 92.99)[37.63, 104.32]**	**9.86 (4.39, 13.23)[0.00, 27.04]**	**19.60 (15.20, 23.63)[0.00, 43.19]**
Cochran C test	4.91 (4.63, 5.19)[3.36, 6.51]	9.90 (9.15, 10.60)[5.03, 13.81]	19.63 (17.32, 21.96)[8.21, 34.98]
Cochran C test partial	4.90 (4.62, 5.18)[3.36, 6.51]	9.96 (9.23, 10.64)[5.58, 13.76]	19.62 (17.31, 21.95)[8.21, 34.98]
Fraser-Harris method	4.90 (4.62, 5.19)[3.41, 6.51]	9.94 (9.22, 10.62)[5.58, 13.76]	19.27 (16.90, 21.56)[7.57, 34.98]
Reed's criterion for means	5.33 (4.87, 41.75)[3.69, 107.75]	9.93 (8.84, 11.03)[0.00, 27.04]	19.56 (16.65, 22.36)[0.00, 43.19]
Reed's criterion for measurements	4.98 (4.71, 5.26)[3.68, 6.66]	9.94 (9.26, 10.64)[6.18, 13.31]	19.61 (17.31, 21.94)[8.28, 35.01]
Tukey IQR rule	4.97 (4.70, 5.25)[3.65, 6.58]	9.83 (9.16, 10.53)[6.15, 13.21]	19.10 (16.73, 21.48)[7.57, 34.58]
Dixon's Q test	4.98 (4.71, 5.26)[3.68, 6.66]	9.94 (9.26, 10.64)[6.18, 13.31]	19.61 (17.31, 21.94)[8.28, 35.01]
Grubbs's test	4.98 (4.71, 5.26)[3.68, 6.66]	9.94 (9.26, 10.64)[6.18, 13.31]	19.61 (17.31, 21.94)[8.28, 35.01]
± 3SD	4.98 (4.71, 5.26)[3.68, 6.66]	9.94 (9.26, 10.64)[6.18, 13.31]	19.61 (17.31, 21.94)[8.28, 35.01]

Table D.40: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**3.41 (2.19, 4.90)[1.15, ∞]**	**236.60 (118.00, 259.59)[109.17, 291.62]**	**9.99 (9.94, 10.04)[9.76, 10.20]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.46]	30.65 (28.80, 32.42)[20.71, 40.20]	9.97 (9.94, 10.01)[9.79, 10.14]
Cochran C test partial	0.56 (0.50, 0.65)[0.30, 1.43]	30.74 (29.01, 32.47)[20.54, 40.03]	9.97 (9.94, 10.01)[9.79, 10.13]
Fraser-Harris method	0.57 (0.51, 0.66)[0.30, 1.60]	30.71 (28.98, 32.45)[20.54, 40.03]	9.97 (9.94, 10.01)[9.79, 10.13]
Reed's criterion for means	0.68 (0.54, 2.42)[0.28, ∞]	32.98 (30.18, 119.19)[22.53, 301.74]	9.98 (9.94, 10.01)[9.79, 10.16]
Reed's criterion for measurements	0.57 (0.50, 0.65)[0.30, 1.39]	30.82 (29.18, 32.56)[22.64, 39.33]	9.97 (9.94, 10.01)[9.79, 10.13]
Tukey IQR rule	0.58 (0.51, 0.67)[0.30, 1.48]	30.54 (28.88, 32.30)[22.45, 39.33]	9.97 (9.94, 10.01)[9.81, 10.14]
Dixon's Q test	0.57 (0.50, 0.65)[0.30, 1.39]	30.82 (29.18, 32.56)[22.64, 39.33]	9.97 (9.94, 10.01)[9.79, 10.13]
Grubbs's test	0.57 (0.50, 0.65)[0.30, 1.39]	30.82 (29.18, 32.56)[22.64, 39.33]	9.97 (9.94, 10.01)[9.79, 10.13]
± 3SD	0.57 (0.50, 0.65)[0.30, 1.39]	30.82 (29.18, 32.56)[22.64, 39.33]	9.97 (9.94, 10.01)[9.79, 10.13]

Table D.41: Outlier detection methods with outlier simulation (0.5% and magnitude 2)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-87.13 (-88.92, -67.74)[-90.86, -65.10]**	**677.30 (209.95, 802.13)[186.51, 994.41]**
Cochran C test	-26.33 (-27.61, -24.97)[-32.96, -18.68]	35.74 (33.27, 38.14)[22.98, 49.16]
Cochran C test partial	-26.39 (-27.65, -25.12)[-32.85, -18.55]	35.86 (33.54, 38.21)[22.77, 48.93]
Fraser-Harris method	-26.37 (-27.63, -25.10)[-32.85, -18.55]	35.82 (33.51, 38.18)[22.77, 48.93]
Reed's criterion for means	-28.01 (-68.07, -25.98)[-91.38, -20.14]	38.91 (35.11, 213.21)[25.22, 1059.51]
Reed's criterion for measurements	-26.46 (-27.71, -25.25)[-32.39, -20.23]	35.97 (33.77, 38.34)[25.36, 47.90]
Tukey IQR rule	-26.25 (-27.53, -25.03)[-32.39, -20.08]	35.59 (33.38, 37.98)[25.13, 47.90]
Dixon's Q test	-26.46 (-27.71, -25.25)[-32.39, -20.23]	35.97 (33.77, 38.34)[25.36, 47.90]
Grubbs's test	-26.46 (-27.71, -25.25)[-32.39, -20.23]	35.97 (33.77, 38.34)[25.36, 47.90]
± 3SD	-26.46 (-27.71, -25.25)[-32.39, -20.23]	35.97 (33.77, 38.34)[25.36, 47.90]

Table D.42: Outlier detection methods with outlier simulation (1% and magnitude 2)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**8109.916**	**-135.070**	**-198.346**	**1622.996**	**-13.541**	**-10.015**	**410.328**	**-19.334**	**-21.982**
Cochran C test	-8.976	-12.832	-25.625	-1.796	-1.286	-1.294	-21.771	-11.762	-7.354
Cochran C test partial	-9.672	-6.384	-25.252	-1.936	-0.640	-1.275	-23.457	-6.059	-7.250
Fraser-Harris method	-9.788	-7.549	-62.722	-1.959	-0.757	-3.167	-23.580	-7.104	-17.737
Reed's criterion for means	5885.671	-117.541	-162.950	1177.869	-11.783	-8.228	230.936	-19.381	-21.253
Reed's criterion for measurements	5673.242	-116.930	-186.847	1135.357	-11.722	-9.435	191.965	-20.054	-25.302
Tukey IQR rule	-2.464	-16.088	-75.736	-0.493	-1.613	-3.824	-6.131	-15.470	-21.292
Dixon's Q test	5673.914	-116.891	-186.731	1135.491	-11.718	-9.429	192.047	-20.047	-25.286
Grubbs's test	4454.166	-168.542	-231.438	891.390	-16.896	-11.686	269.379	-32.246	-36.583
± 3SD	-1.263	-5.638	-25.636	-0.253	-0.565	-1.294	-3.171	-5.494	-7.370

Table D.43: Outlier detection methods with outlier simulation (1% and magnitude 2)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	**6967.708**	**50.630**	**85.348**	**0.000**	$-^a$	**0.353**^{b}	**0.273**	$-^a$	**0.980**^{b}
Cochran C test	0.178	1.207	12.207	0.935	0.936	0.946	0.016	0.041	0.147
Cochran C test partial	0.179	1.114	12.194	0.935	0.946	0.946	0.016	0.041	0.147
Fraser-Harris method	0.182	1.135	12.899	0.935	0.944	0.942	0.016	0.042	0.146
Reed's criterion for means	4113.656	38.164	61.442	0.024	0.976^c	0.975^d	0.206	0.044^c	0.262^d
Reed's criterion for measurements	4091.974	35.365	58.024	0.000	1.000^e	0.982^f	0.196	0.039^e	0.249$-^f$
Tukey IQR rule	0.162	1.107	13.226	0.944	0.945	0.939	0.016	0.041	0.144
Dixon's Q test	4092.206	35.365	58.019	0.000	$-^a$	0.981^f	0.196	$-^a$	0.249^f
Grubbs's test	2257.363	30.160	45.380	0.000	$-^a$	0.985^g	0.157	$-^a$	0.235^g
± 3SD	0.159	1.056	12.165	0.948	0.951	0.947	0.016	0.041	0.147

^{a}5000 CIs could not be calculated; ^{b}4344 CIs could not be calculated; ^{c}4875 CIs could not be calculated; ^{d}3037 CIs could not be calculated; ^{e}4999 CIs could not be calculated; ^{f}2452 CIs could not be calculated; ^{g}2051 CIs could not be calculated.

Table D.44: Outlier detection methods with outlier simulation (1% and magnitude 2)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.88 (0.83, 1.04)[0.53, 1.19]**	**0.09 (0.00, 0.14)[0.00, 0.29]**	**0.19 (0.13, 0.24)[0.00, 0.45]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.05, 0.14]	0.19 (0.17, 0.22)[0.10, 0.31]
Cochran C test partial	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.31]
Fraser-Harris method	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.09, 0.31]
Reed's criterion for means	0.57 (0.41, 0.82)[0.04, 1.19]	0.10 (0.03, 0.13)[0.00, 0.28]	0.19 (0.14, 0.23)[0.00, 0.42]
Reed's criterion for measurements	0.42 (0.40, 1.04)[0.05, 1.19]	0.10 (0.04, 0.13)[0.00, 0.29]	0.19 (0.14, 0.23)[0.00, 0.45]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.07, 0.31]
Dixon's Q test	0.42 (0.40, 1.04)[0.37, 1.19]	0.10 (0.04, 0.13)[0.00, 0.29]	0.19 (0.14, 0.23)[0.00, 0.45]
Grubbs's test	0.40 (0.40, 0.72)[0.37, 0.84]	0.09 (0.04, 0.12)[0.00, 0.23]	0.18 (0.14, 0.22)[0.00, 0.36]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.31]

Table D.45: Outlier detection methods with outlier simulation (1% and magnitude 2)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**108.43 (100.40, 140.40)[56.60, 175.51]**	**9.39 (0.04, 14.32)[0.00, 29.95]**	**19.21 (12.85, 24.48)[0.00, 47.16]**
Cochran C test	4.90 (4.63, 5.19)[3.56, 6.41]	9.88 (9.15, 10.60)[5.39, 13.89]	19.63 (17.29, 22.05)[9.80, 32.17]
Cochran C test partial	4.89 (4.63, 5.18)[3.54, 6.33]	9.93 (9.23, 10.65)[6.37, 13.82]	19.65 (17.29, 22.06)[9.39, 32.17]
Fraser-Harris method	4.89 (4.62, 5.18)[3.54, 6.40]	9.92 (9.21, 10.63)[6.37, 13.86]	19.27 (16.86, 21.75)[9.20, 32.17]
Reed's criterion for means	61.56 (42.65, 98.26)[4.07, 175.51]	9.68 (3.41, 13.27)[0.00, 29.04]	19.18 (14.34, 23.51)[0.00, 44.04]
Reed's criterion for measurements	43.90 (41.51, 140.40)[5.46, 175.51]	9.69 (4.23, 12.99)[0.00, 29.95]	18.89 (14.46, 22.98)[0.00, 47.16]
Tukey IQR rule	4.96 (4.70, 5.24)[3.67, 6.37]	9.83 (9.13, 10.55)[6.24, 13.69]	19.14 (16.66, 21.66)[7.28, 31.28]
Dixon's Q test	43.90 (41.51, 140.40)[38.21, 175.51]	9.69 (4.23, 12.99)[0.00, 29.95]	18.89 (14.46, 22.98)[0.00, 47.16]
Grubbs's test	42.11 (41.24, 82.78)[38.21, 100.82]	9.26 (4.39, 12.18)[0.00, 23.78]	18.27 (14.34, 21.97)[0.00, 37.43]
± 3SD	4.98 (4.72, 5.25)[3.67, 6.37]	9.94 (9.25, 10.65)[6.35, 13.69]	19.66 (17.27, 22.09)[9.02, 32.26]

Table D.46: Outlier detection methods with outlier simulation (1% and magnitude 2): II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**5.56 (3.98, 8.93)[1.44, ∞]**	**302.16 (278.39, 389.17)[159.64, 488.92]**	**10.01 (9.95, 10.06)[9.76, 10.28]**
Cochran C test	0.56 (0.49, 0.65)[0.27, 1.27]	30.57 (28.84, 32.43)[19.30, 40.72]	9.97 (9.94, 10.00)[9.81, 10.17]
Cochran C test partial	0.56 (0.49, 0.65)[0.30, 1.27]	30.70 (28.99, 32.52)[21.39, 40.60]	9.97 (9.94, 10.00)[9.80, 10.17]
Fraser-Harris method	0.57 (0.50, 0.66)[0.31, 1.30]	30.67 (28.94, 32.49)[21.39, 40.72]	9.97 (9.94, 10.01)[9.80, 10.15]
Reed's criterion for means	3.65 (2.33, 5.88)[0.34, ∞]	172.84 (121.51, 274.69)[25.84, 488.92]	9.99 (9.93, 10.04)[9.76, 10.26]
Reed's criterion for measurements	2.91 (2.16, 6.01)[0.70, ∞]	125.24 (118.37, 389.17)[28.97, 488.92]	9.95 (9.91, 10.03)[9.76, 10.28]
Tukey IQR rule	0.58 (0.50, 0.67)[0.30, 1.33]	30.58 (28.87, 32.34)[21.47, 40.26]	9.97 (9.94, 10.00)[9.81, 10.16]
Dixon's Q test	2.91 (2.16, 6.01)[1.10, ∞]	125.24 (118.37, 389.17)[109.08, 488.92]	9.95 (9.91, 10.03)[9.76, 10.28]
Grubbs's test	2.60 (2.16, 6.01)[0.70, ∞]	119.95 (117.39, 229.44)[109.08, 282.79]	9.96 (9.92, 10.00)[9.79, 10.22]
± 3SD	0.57 (0.50, 0.65)[0.30, 1.29]	30.84 (29.16, 32.59)[21.47, 40.26]	9.97 (9.94, 10.00)[9.80, 10.16]

Table D.47: Outlier detection methods with outlier simulation (1% and magnitude 2)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-91.40 (-94.46, -90.12)[-96.30, -77.32]**	**1062.23 (912.43, 1703.94)[340.84, 2599.75]**
Cochran C test	-26.27 (-27.62, -25.00)[-33.31, -17.53]	35.64 (33.33, 38.16)[21.25, 49.94]
Cochran C test partial	-26.36 (-27.68, -25.11)[-33.23, -19.23]	35.80 (33.52, 38.27)[23.81, 49.76]
Fraser-Harris method	-26.34 (-27.66, -25.07)[-33.31, -19.23]	35.77 (33.46, 38.24)[23.81, 49.94]
Reed's criterion for means	-79.58 (-89.90, -68.72)[-96.30, -22.72]	389.69 (219.67, 890.16)[29.41, 2599.75]
Reed's criterion for measurements	-69.72 (-94.46, -67.84)[-96.30, -25.09]	230.25 (210.97, 1703.94)[33.50, 2599.75]
Tukey IQR rule	-26.28 (-27.55, -25.02)[-33.00, -19.30]	35.65 (33.37, 38.03)[23.91, 49.25]
Dixon's Q test	-69.72 (-94.46, -67.84)[-96.30, -65.07]	230.25 (210.97, 1,703.94)[186.27, 2,599.75]
Grubbs's test	-68.28 (-86.50, -67.56)[-90.38, -65.07]	215.31 (208.27, 640.68)[186.27, 939.31]
± 3SD	-26.47 (-27.73, -25.23)[-33.00, -19.30]	35.99 (33.75, 38.37)[23.91, 49.25]

Table D.48: Outlier detection methods with outlier simulation (2% and magnitude 2)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**-159.199**	**-291.521**	**2356.935**	**-15.960**	**-14.720**	**596.095**	**-19.523**	**-25.591**
Cochran C test	-8.230	-13.100	-23.167	-1.647	-1.313	-1.170	-19.326	-11.478	-6.540
Cochran C test partial	-8.299	-5.968	-23.589	-1.661	-0.598	-1.191	-19.574	-5.463	-6.674
Fraser-Harris method	-8.470	-7.145	-57.669	-1.695	-0.716	-2.912	-19.833	-6.496	-16.000
Reed's criterion for means	†	-160.449	-293.786	2088.670	-16.085	-14.835	442.682	-20.333	-26.947
Reed's criterion for measurements	†	-163.613	-364.992	2177.340	-16.402	-18.430	350.754	-20.813	-33.483
Tukey IQR rule	-1.814	-13.921	-71.855	-0.363	-1.396	-3.628	-4.339	-13.072	-20.002
Dixon's Q test	†	-163.613	-364.992	2177.340	-16.402	-18.430	350.754	-20.813	-33.483
Grubbs's test	9006.759	-279.046	-557.880	1802.477	-27.974	-28.170	422.853	-39.779	-57.615
± 3SD	-0.692	-4.283	-22.750	-0.139	-0.429	-1.149	-1.669	-4.070	-6.490

† greater than 10^4.

Table D.49: Outlier detection methods with outlier simulation (2% and magnitude 2)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**69.026**	**138.262**	**0.000**	$-^a$	**0.976**^{b}	**0.389**	$-^a$	**0.563**^{b}
Cochran C test	0.188	1.320	12.602	0.926	0.916	0.945	0.016	0.041	0.147
Cochran C test partial	0.187	1.197	12.550	0.925	0.927	0.946	0.016	0.041	0.147
Fraser-Harris method	0.190	1.215	13.323	0.926	0.929	0.941	0.016	0.042	0.146
Reed's criterion for means	†	64.843	127.491	0.000	$-^a$	0.987^c	0.353	$-^a$	0.460^c
Reed's criterion for measurements	†	64.477	132.146	0.000	$-^a$	0.995^d	0.361	$-^a$	0.394^d
Tukey IQR rule	0.175	1.154	13.421	0.933	0.938	0.937	0.016	0.041	0.144
Dixon's Q test	†	64.477	132.146	0.000	$-^a$	0.995^d	0.361	$-^a$	0.394^d
Grubbs's test	8565.860	56.996	124.882	0.000	$-^a$	0.991^f	0.301	$-^a$	0.398^f
± 3SD	0.172	1.109	12.337	0.936	0.943	0.946	0.016	0.042	0.147

† greater than 10^4. ^{a}5000 CIs could not be calculated; ^{b}4959 CIs could not be calculated; ^{c}4848 CIs could not be calculated; ^{d}4788 CIs could not be calculated; ^{e}4959 CIs could not be calculated; ^{f}4766 CIs could not be calculated.

Table D.50: Outlier detection methods with outlier simulation (2% and magnitude 2)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**1.24 (1.05, 1.41)[0.76, 1.63]**	**0.08 (0.00, 0.15)[0.00, 0.33]**	**0.19 (0.08, 0.26)[0.00, 0.48]**
Cochran C test	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.06, 0.35]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.06, 0.35]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.06, 0.35]
Reed's criterion for means	1.06 (0.98, 1.24)[0.58, 1.62]	0.08 (0.00, 0.15)[0.00, 0.33]	0.19 (0.09, 0.25)[0.00, 0.47]
Reed's criterion for measurements	1.24 (0.71, 1.41)[0.66, 1.63]	0.08 (0.00, 0.15)[0.00, 0.33]	0.18 (0.08, 0.24)[0.00, 0.48]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.21)[0.06, 0.33]
Dixon's Q test	1.24 (0.71, 1.41)[0.66, 1.63]	0.08 (0.00, 0.15)[0.00, 0.33]	0.18 (0.08, 0.24)[0.00, 0.48]
Grubbs's test	0.96 (0.69, 1.16)[0.66, 1.41]	0.07 (0.00, 0.13)[0.00, 0.29]	0.16 (0.06, 0.22)[0.00, 0.42]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.11)[0.06, 0.13]	0.19 (0.17, 0.22)[0.07, 0.34]

Table D.51: Outlier detection methods with outlier simulation (2% and magnitude 2)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**192.61 (142.42, 249.50)[88.70, 362.45]**	**8.01 (0.04, 15.31)[0.01, 33.63]**	**18.87 (7.95, 25.94)[0.00, 50.62]**
Cochran C test	4.91 (4.62, 5.20)[3.20, 6.42]	9.89 (9.13, 10.64)[5.95, 13.36]	19.62 (17.31, 22.22)[6.14, 35.58]
Cochran C test partial	4.92 (4.62, 5.20)[3.14, 6.38]	9.94 (9.22, 10.67)[5.95, 13.39]	19.60 (17.29, 22.21)[6.14, 35.58]
Fraser-Harris method	4.91 (4.62, 5.20)[3.14, 6.44]	9.92 (9.20, 10.66)[5.95, 13.39]	19.23 (16.85, 21.96)[6.14, 36.00]
Reed's criterion for means	143.78 (126.21, 190.99)[62.90, 360.54]	8.37 (0.04, 15.10)[0.00, 33.63]	18.72 (8.97, 25.41)[0.00, 49.36]
Reed's criterion for measurements	192.61 (80.63, 249.50)[73.41, 362.45]	8.36 (0.04, 14.79)[0.00, 33.63]	17.80 (7.79, 24.53)[0.00, 50.62]
Tukey IQR rule	4.98 (4.69, 5.26)[3.52, 6.75]	9.87 (9.17, 10.56)[5.61, 13.31]	19.15 (16.69, 21.74)[5.54, 33.54]
Dixon's Q test	192.61 (80.63, 249.50)[73.41, 362.45]	8.36 (0.04, 14.79)[0.00, 33.63]	17.80 (7.79, 24.53)[0.00, 50.62]
Grubbs's test	122.73 (78.58, 168.92)[73.41, 252.69]	7.06 (0.03, 13.01)[0.00, 30.03]	15.72 (6.06, 21.95)[0.00, 44.07]
± 3SD	4.99 (4.71, 5.27)[3.52, 6.75]	9.94 (9.27, 10.66)[5.85, 13.31]	19.63 (17.28, 22.21)[6.92, 35.20]

Table D.52: Outlier detection methods with outlier simulation (2% and magnitude 2)–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**10.15 (6.79, 26.62)[2.77, ∞]**	**534.90 (396.42, 692.11)[250.08, 1007.45]**	**10.03 (9.96, 10.11)[9.72, 10.36]**
Cochran C test	0.56 (0.49, 0.65)[0.27, 1.88]	30.63 (28.77, 32.48)[21.11, 39.77]	9.97 (9.94, 10.00)[9.78, 10.13]
Cochran C test partial	0.56 (0.49, 0.65)[0.27, 1.88]	30.75 (29.00, 32.61)[21.35, 39.82]	9.97 (9.94, 10.00)[9.78, 10.13]
Fraser-Harris method	0.57 (0.50, 0.66)[0.27, 1.88]	30.72 (28.93, 32.60)[21.35, 39.82]	9.97 (9.94, 10.01)[9.80, 10.13]
Reed's criterion for means	8.40 (5.52, 18.17)[2.06, ∞]	400.07 (350.68, 530.43)[174.35, 1000.84]	10.01 (9.94, 10.09)[9.72, 10.33]
Reed's criterion for measurements	9.63 (5.99, 24.88)[2.02, ∞]	534.90 (224.15, 692.11)[205.52, 1007.45]	10.03 (9.91, 10.11)[9.72, 10.36]
Tukey IQR rule	0.58 (0.50, 0.67)[0.28, 1.92]	30.64 (28.95, 32.41)[20.79, 40.08]	9.97 (9.94, 10.00)[9.79, 10.14]
Dixon's Q test	9.63 (5.99, 24.88)[2.02, ∞]	534.90 (224.15, 692.11)[205.52, 1,007.45]	10.03 (9.91, 10.11)[9.72, 10.36]
Grubbs's test	7.82 (5.03, 21.51)[2.02, ∞]	341.19 (219.51, 468.75)[205.52, 703.55]	9.97 (9.91, 10.05)[9.72, 10.30]
± 3SD	0.57 (0.49, 0.65)[0.28, 1.69]	30.88 (29.18, 32.66)[21.47, 40.08]	9.97 (9.94, 10.00)[9.78, 10.13]

Table D.53: Outlier detection methods with outlier simulation (2% and magnitude 2)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-96.84 (-97.97, -94.63)[-98.91, -88.22]**	**3062.33 (1763.58, 4836.51)[749.16, 9043.41]**
Cochran C test	-26.31 (-27.66, -24.94)[-32.68, -19.01]	35.71 (33.23, 38.23)[23.46, 48.54]
Cochran C test partial	-26.40 (-27.74, -25.12)[-32.71, -19.20]	35.87 (33.54, 38.40)[23.76, 48.61]
Fraser-Harris method	-26.38 (-27.74, -25.06)[-32.71, -19.20]	35.84 (33.44, 38.39)[23.76, 48.61]
Reed's criterion for means	-94.72 (-96.79, -93.35)[-98.89, -79.82]	1794.01 (1402.69, 3016.04)[395.50, 8948.68]
Reed's criterion for measurements	-96.84 (-97.97, -86.00)[-98.91, -84.03]	3062.33 (614.36, 4836.51)[526.28, 9043.41]
Tukey IQR rule	-26.32 (-27.60, -25.07)[-32.88, -18.75]	35.73 (33.46, 38.13)[23.07, 48.99]
Dixon's Q test	-96.84 (-97.97, -86.00)[-98.91, -84.03]	3062.33 (614.36, 4836.51)[526.28, 9043.41]
Grubbs's test	-93.02 (-96.01, -85.54)[-98.03, -84.03]	1332.53 (591.73, 2406.19)[526.28, 4975.69]
± 3SD	-26.49 (-27.78, -25.25)[-32.88, -19.30]	36.04 (33.78, 38.47)[23.91, 48.99]

Table D.54: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**60.439**	**-4.044**	**7688.806**	**6.059**	**-0.204**	**120.325**	**4.735**	**-0.217**
Cochran C test	-10.144	-12.187	-27.384	-2.030	-1.222	-1.383	-24.896	-11.351	-8.004
Cochran C test partial	-10.975	-6.344	-27.451	-2.196	-0.636	-1.386	-26.952	-6.109	-8.044
Fraser-Harris method	-11.028	-7.082	-66.806	-2.207	-0.710	-3.373	-26.797	-6.775	-19.259
Reed's criterion for means	†	21.389	-22.074	2311.292	2.144	-1.115	46.602	2.887	-2.071
Reed's criterion for measurements	-2.052	-5.810	-28.096	-0.411	-0.582	-1.419	-5.187	-5.695	-8.261
Tukey IQR rule	-3.275	-16.824	-79.860	-0.655	-1.687	-4.032	-8.234	-16.357	-22.644
Dixon's Q test	-2.052	-5.810	-28.096	-0.411	-0.582	-1.419	-5.187	-5.695	-8.261
Grubbs's test	-2.052	-5.810	-28.096	-0.411	-0.582	-1.419	-5.187	-5.695	-8.261
± 3SD	-2.052	-5.810	-28.096	-0.411	-0.582	-1.419	-5.187	-5.695	-8.261

† greater than 10^4.

Table D.55: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**163.275**	**346.544**	**0.000**	$-^a$	**0.986**^{b}	**1.233**	$-^a$	**0.390**^{b}
Cochran C test	0.176	1.168	11.781	0.934	0.936	0.953	0.016	0.041	0.147
Cochran C test partial	0.178	1.082	11.721	0.931	0.947	0.953	0.015	0.041	0.147
Fraser-Harris method	0.182	1.097	12.479	0.929	0.946	0.946	0.016	0.042	0.146
Reed's criterion for means	†	54.952	113.647	0.651	0.951^c	0.957^d	0.384	0.043^c	0.163^d
Reed's criterion for measurements	0.157	1.044	11.645	0.948	0.951	0.953	0.016	0.041	0.147
Tukey IQR rule	0.159	1.086	13.076	0.948	0.947	0.938	0.016	0.041	0.143
Dixon's Q test	0.157	1.044	11.645	0.948	0.951	0.953	0.016	0.041	0.147
Grubbs's test	0.157	1.044	11.645	0.948	0.951	0.953	0.016	0.041	0.147
± 3SD	0.157	1.044	11.645	0.948	0.951	0.953	0.016	0.041	0.147

† greater than 10^4. ^{a}5000 CIs could not be calculated; ^{b}4496 CIs could not be calculated; ^{c}1563 CIs could not be calculated; ^{d}1389 CIs could not be calculated.

Table D.56: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**0.75 (0.71, 7.09)[0.66, 7.65]**	**0.07 (0.00, 0.16)[0.00, 0.66]**	**0.18 (0.01, 0.28)[0.00, 0.95]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.34]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.35]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.06, 0.35]
Reed's criterion for means	0.05 (0.05, 0.70)[0.04, 7.65]	0.10 (0.09, 0.11)[0.00, 0.63]	0.19 (0.16, 0.22)[0.00, 0.95]
Reed's criterion for measurements	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.34]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.10)[0.06, 0.14]	0.19 (0.17, 0.21)[0.06, 0.34]
Dixon's Q test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.34]
Grubbs's test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.34]
± 3SD	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.34]

Table D.57: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	**86.73 (80.86, †)[73.67, †]**	**7.33 (0.43, 15.66)[0.00, 73.22]**	**18.37 (0.85, 28.06)[0.00, 121.52]**
Cochran C test	4.89 (4.62, 5.17)[3.51, 6.33]	9.89 (9.15, 10.59)[5.72, 14.20]	19.65 (17.28, 22.00)[9.09, 35.44]
Cochran C test partial	4.89 (4.61, 5.16)[3.49, 6.32]	9.94 (9.21, 10.62)[6.23, 14.20]	19.63 (17.28, 22.01)[8.97, 35.61]
Fraser-Harris method	4.88 (4.61, 5.17)[3.49, 6.33]	9.93 (9.21, 10.62)[6.26, 14.20]	19.28 (16.89, 21.66)[6.38, 35.61]
Reed's criterion for means	5.22 (4.83, 80.24)[3.59, †]	9.87 (8.71, 10.89)[0.00, 69.48]	19.53 (16.15, 22.78)[0.00, 121.16]
Reed's criterion for measurements	4.98 (4.71, 5.25)[3.67, 6.36]	9.94 (9.24, 10.62)[6.13, 13.90]	19.61 (17.28, 22.02)[8.62, 35.53]
Tukey IQR rule	4.97 (4.69, 5.24)[3.66, 6.36]	9.82 (9.13, 10.53)[6.13, 13.90]	19.10 (16.73, 21.54)[6.00, 35.53]
Dixon's Q test	4.98 (4.71, 5.25)[3.67, 6.36]	9.94 (9.24, 10.62)[6.13, 13.90]	19.61 (17.28, 22.02)[8.62, 35.53]
Grubbs's test	4.98 (4.71, 5.25)[3.67, 6.36]	9.94 (9.24, 10.62)[6.13, 13.90]	19.61 (17.28, 22.02)[8.62, 35.53]
± 3SD	4.98 (4.71, 5.25)[3.67, 6.36]	9.94 (9.24, 10.62)[6.13, 13.90]	19.61 (17.28, 22.02)[8.62, 35.53]

† greater than 10^4.

Table D.58: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	**† (4.13, †)[1.84, ∞]**	**240.74 (226.30, †)[206.19, †]**	**10.04 (9.92, 10.53)[9.76, 10.70]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.35]	30.60 (28.81, 32.36)[21.14, 41.31]	9.97 (9.94, 10.00)[9.81, 10.13]
Cochran C test partial	0.56 (0.49, 0.65)[0.29, 1.34]	30.71 (28.93, 32.44)[22.21, 41.27]	9.97 (9.94, 10.00)[9.81, 10.13]
Fraser-Harris method	0.57 (0.50, 0.67)[0.29, 1.64]	30.70 (28.92, 32.44)[22.39, 41.27]	9.97 (9.94, 10.00)[9.81, 10.15]
Reed's criterion for means	0.64 (0.53, 3.71)[0.29, ∞]	32.39 (29.93, 224.89)[22.47, †]	9.97 (9.94, 10.02)[9.77, 10.68]
Reed's criterion for measurements	0.57 (0.50, 0.65)[0.29, 1.34]	30.80 (29.10, 32.54)[22.26, 40.63]	9.97 (9.94, 10.00)[9.80, 10.12]
Tukey IQR rule	0.58 (0.50, 0.67)[0.30, 1.75]	30.51 (28.82, 32.27)[22.04, 40.63]	9.97 (9.94, 10.00)[9.80, 10.13]
Dixon's Q test	0.57 (0.50, 0.65)[0.29, 1.34]	30.80 (29.10, 32.54)[22.26, 40.63]	9.97 (9.94, 10.00)[9.80, 10.12]
Grubbs's test	0.57 (0.50, 0.65)[0.29, 1.34]	30.80 (29.10, 32.54)[22.26, 40.63]	9.97 (9.94, 10.00)[9.80, 10.12]
± 3SD	0.57 (0.50, 0.65)[0.29, 1.34]	30.80 (29.10, 32.54)[22.26, 40.63]	9.97 (9.94, 10.00)[9.80, 10.12]

† greater than 10^4.

Table D.59: Outlier detection methods with outlier simulation (0.5% and magnitude 10)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-87.48 (-100.00, -86.21)[-100.00, -84.11]**	**698.96 (624.96, †)[529.33, †]**
Cochran C test	-26.29 (-27.57, -24.97)[-33.69, -19.03]	35.67 (33.28, 38.06)[23.50, 50.81]
Cochran C test partial	-26.37 (-27.63, -25.07)[-33.66, -19.89]	35.82 (33.45, 38.17)[24.83, 50.75]
Fraser-Harris method	-26.36 (-27.62, -25.05)[-33.66, -20.03]	35.80 (33.43, 38.16)[25.05, 50.75]
Reed's criterion for means	-27.59 (-86.07, -25.80)[-100.00, -20.10]	38.10 (34.77, 617.99)[25.15, †]
Reed's criterion for measurements	-26.44 (-27.70, -25.19)[-33.24, -19.93]	35.95 (33.67, 38.31)[24.89, 49.80]
Tukey IQR rule	-26.23 (-27.50, -24.98)[-33.24, -19.75]	35.55 (33.30, 37.93)[24.61, 49.80]
Dixon's Q test	-26.44 (-27.70, -25.19)[-33.24, -19.93]	35.95 (33.67, 38.31)[24.89, 49.80]
Grubbs's test	-26.44 (-27.70, -25.19)[-33.24, -19.93]	35.95 (33.67, 38.31)[24.89, 49.80]
± 3SD	-26.44 (-27.70, -25.19)[-33.24, -19.93]	35.95 (33.67, 38.31)[24.89, 49.80]

† greater than 10^4.

Table D.60: Outlier detection methods with outlier simulation (1% and magnitude 10)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**-73.068**	**-235.552**	†	**-7.325**	**-11.894**	**190.281**	**-5.204**	**-10.409**
Cochran C test	-9.787	-12.971	-25.385	-1.959	-1.300	-1.282	-23.670	-11.738	-7.326
Cochran C test partial	-10.384	-7.873	-25.072	-2.078	-0.789	-1.266	-25.099	-7.350	-7.258
Fraser-Harris method	-10.403	-8.747	-63.203	-2.082	-0.877	-3.191	-24.969	-8.124	-17.916
Reed's criterion for means	†	-74.900	-159.939	7106.731	-7.509	-8.076	100.335	-6.412	-9.166
Reed's criterion for measurements	†	-340.732	-679.873	6296.039	-34.158	-34.330	77.946	-45.727	-63.279
Tukey IQR rule	-2.845	-17.476	-75.277	-0.569	-1.752	-3.801	-7.018	-16.868	-21.164
Dixons Q test	†	-340.732	-679.873	6296.039	-34.158	-34.330	77.946	-45.727	-63.279
Grubbs's test	†	-235.845	-491.365	4602.308	-23.643	-24.811	83.192	-28.134	-41.467
± 3SD	3252.061	-85.324	-95.739	650.818	-8.554	-4.834	98.082	-18.428	-15.778

† greater than 10^4.

Table D.61: Outlier detection methods with outlier simulation (1% and magnitude 10)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**197.656**	**517.633**	**0.000**	$-^a$	**1.000**^{b}	**2.011**	$-^a$	**0.563**^{b}
Cochran C test	0.181	1.238	12.070	0.933	0.930	0.948	0.016	0.041	0.147
Cochran C test partial	0.182	1.153	11.996	0.932	0.938	0.947	0.016	0.041	0.147
Fraser-Harris method	0.184	1.167	12.844	0.931	0.940	0.942	0.016	0.042	0.146
Reed's criterion for means	†	136.997	307.001	0.012	0.966^c	0.986^d	1.158	0.044^c	0.382^d
Reed's criterion for measurements	†	67.134	161.658	0.000	$-^a$	0.989^e	1.013	$-^a$	0.401^e
Tukey IQR rule	0.165	1.104	13.218	0.943	0.946	0.935	0.016	0.041	0.144
Dixon's Q test	†	67.134	161.658	0.000	$-^a$	0.989^e	1.013	$-^a$	0.401^e
Grubbs's test	†	75.837	164.554	0.000	$-^a$	0.991^f	0.745	$-^a$	0.394^f
± 3SD	2156.957	22.165	37.737	0.481	0.954^g	0.947^h	0.119	0.041^g	0.119^h

† greater than 10^4. ^{a}5000 CIs could not be calculated; ^{b}4976 CIs could not be calculated; ^{c}4941 CIs could not be calculated; ^{d}4559 CIs could not be calculated; ^{e}4523 CIs could not be calculated; ^{f}4433 CIs could not be calculated; ^{g}2455 CIs could not be calculated; ^{h}2002 CIs could not be calculated.

Table D.62: Outlier detection methods with outlier simulation (1% and magnitude 10)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**7.13 (6.60, 9.69)[0.95, 10.53]**	**0.01 (0.00, 0.15)[0.00, 0.74]**	**0.03 (0.00, 0.31)[0.00, 1.03]**
Cochran C test	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.33]
Cochran C test partial	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.09, 0.33]
Fraser-Harris method	0.05 (0.05, 0.05)[0.03, 0.07]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.22)[0.07, 0.33]
Reed's criterion for means	0.76 (0.73, 7.22)[0.04, 10.61]	0.06 (0.00, 0.14)[0.00, 0.74]	0.17 (0.00, 0.25)[0.00, 0.93]
Reed's criterion for measurements	0.98 (0.71, 9.69)[0.67, 10.53]	0.03 (0.00, 0.12)[0.00, 0.74]	0.15 (0.00, 0.22)[0.00, 0.65]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.07]	0.10 (0.09, 0.10)[0.06, 0.14]	0.19 (0.17, 0.21)[0.06, 0.33]
Dixon's Q test	0.98 (0.71, 9.69)[0.67, 10.53]	0.03 (0.00, 0.12)[0.00, 0.74]	0.15 (0.00, 0.22)[0.00, 0.65]
Grubbs's test	0.72 (0.71, 6.75)[0.67, 7.41]	0.06 (0.00, 0.13)[0.00, 0.59]	0.16 (0.02, 0.22)[0.00, 0.87]
± 3SD	0.06 (0.05, 0.71)[0.04, 0.76]	0.10 (0.08, 0.11)[0.00, 0.24]	0.19 (0.16, 0.23)[0.00, 0.37]

Table D.63: Outlier detection methods with outlier simulation (1% and magnitude 2)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	† (†, †)**[121.75,** †]	**0.78 (0.40, 15.22)[0.01, 84.86]**	**3.15 (0.00, 32.26)[0.00, 136.86]**
Cochran C test	4.89 (4.63, 5.18)[3.48, 6.56]	9.86 (9.14, 10.61)[5.71, 13.97]	19.63 (17.27, 22.10)[9.24, 33.69]
Cochran C test partial	4.89 (4.62, 5.17)[3.48, 6.56]	9.90 (9.21, 10.64)[6.22, 13.86]	19.64 (17.28, 22.09)[9.24, 33.43]
Fraser-Harris method	4.89 (4.62, 5.17)[3.45, 6.67]	9.89 (9.20, 10.63)[6.14, 13.86]	19.24 (16.85, 21.76)[7.43, 33.43]
Reed's criterion for means	88.73 (83.52, †)[4.21, †]	5.99 (0.36, 13.96)[0.01, 84.86]	17.56 (0.01, 25.49)[0.00, 117.89]
Reed's criterion for measurements	126.29 (81.40, †)[74.88, †]	3.32 (0.28, 12.26)[0.00, 84.86]	14.94 (0.00, 22.06)[0.00, 72.17]
Tukey IQR rule	4.96 (4.70, 5.24)[3.66, 6.68]	9.80 (9.12, 10.51)[6.43, 13.95]	19.12 (16.76, 21.63)[6.41, 33.57]
Dixon's Q test	126.29 (81.40, †)[74.88, †]	3.32 (0.28, 12.26)[0.00, 84.86]	14.94 (0.00, 22.06)[0.00, 72.17]
Grubbs's test	82.72 (80.83, †)[74.88, †]	6.44 (0.25, 12.95)[0.00, 64.17]	16.11 (1.56, 22.59)[0.00, 105.68]
± 3SD	5.83 (4.96, 81.27)[3.71, 88.52]	9.84 (8.31, 11.18)[0.00, 24.70]	19.65 (16.27, 22.84)[0.00, 38.55]

† greater than 10^4.

Table D.64: Outlier detection methods with outlier simulation (1% and magnitude 10)–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	† (†, †)[†, ∞]	† (†, †)[**341.18**, †]	**10.48 (10.30, 11.00)[9.72, 11.27]**
Cochran C test	0.56 (0.49, 0.65)[0.29, 1.40]	30.54 (28.80, 32.43)[21.04, 41.04]	9.97 (9.94, 10.00)[9.78, 10.13]
Cochran C test partial	0.56 (0.49, 0.65)[0.29, 1.36]	30.64 (28.92, 32.49)[22.14, 40.72]	9.97 (9.94, 10.00)[9.79, 10.14]
Fraser-Harris method	0.57 (0.50, 0.66)[0.28, 1.59]	30.61 (28.90, 32.48)[22.14, 40.72]	9.97 (9.94, 10.00)[9.79, 10.16]
Reed's criterion for means	10.04 (4.14, †)[0.41, ∞]	246.71 (233.31, †)[24.11, †]	9.96 (9.90, 10.53)[9.72, 11.25]
Reed's criterion for measurements	6.82 (4.12, †)[2.00, ∞]	351.90 (227.40, †)[207.93, †]	9.92 (9.88, 11.00)[9.72, 11.27]
Tukey IQR rule	0.57 (0.50, 0.66)[0.29, 1.46]	30.50 (28.83, 32.30)[21.93, 40.99]	9.97 (9.94, 10.00)[9.79, 10.14]
Dixon's Q test	6.82 (4.12, †)[2.00, ∞]	351.90 (227.40, †)[207.93, †]	9.92 (9.88, 11.00)[9.72, 11.27]
Grubbs's test	5.75 (3.84, †)[2.00, ∞]	231.00 (225.99, †)[207.93, †]	9.94 (9.90, 10.44)[9.74, 10.68]
± 3SD	0.91 (0.56, 4.15)[0.31, ∞]	36.34 (30.72, 227.04)[21.93, 245.37]	9.95 (9.91, 9.98)[9.74, 10.13]

† greater than 10^4.

Table D.65: Outlier detection methods with outlier simulation (1% and magnitude 10)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-100.00 (-100.00, -100.00)[-100.00, -93.02]**	† (†, †)[**1332.44**, †]
Cochran C test	-26.25 (-27.62, -24.97)[-33.51, -18.95]	35.59 (33.28, 38.16)[23.38, 50.41]
Cochran C test partial	-26.32 (-27.66, -25.05)[-33.31, -19.83]	35.72 (33.43, 38.24)[24.73, 49.94]
Fraser-Harris method	-26.30 (-27.65, -25.04)[-33.31, -19.83]	35.69 (33.40, 38.22)[24.73, 49.94]
Reed's criterion for means	-87.96 (-100.00, -86.85)[-100.00, -21.39]	730.83 (660.35, †)[27.21, †]
Reed's criterion for measurements	-93.39 (-100.00, -86.31)[-100.00, -84.31]	1411.82 (630.43, †)[537.26, †]
Tukey IQR rule	-26.22 (-27.52, -24.99)[-33.48, -19.66]	35.54 (33.31, 37.97)[24.47, 50.33]
Dixon's Q test	-93.39 (-100.00, -86.31)[-100.00, -84.31]	1411.82 (630.43, †)[537.26, †]
Grubbs's test	-86.64 (-100.00, -86.18)[-100.00, -84.31]	648.58 (623.41, †)[537.26 †]
± 3SD	-30.36 (-86.28, -26.38)[-87.86, -19.66]	43.60 (35.83, 628.67)[24.47, 723.58]

† greater than 10^4.

Table D.66: Outlier detection methods with outlier simulation (2% and magnitude 10)–bias performance measures.

Outlier strategy	Bias ($\times 10^{-4}$)			Bias Percentage bias			Standardised bias		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**-357.863**	**-866.090**	†	**-35.876**	**-43.733**	**324.080**	**-27.114**	**-38.045**
Cochran C test	-9.595	-14.479	-33.887	-1.920	-1.451	-1.711	-23.061	-12.945	-9.741
Cochran C test partial	-9.818	-7.926	-34.784	-1.965	-0.795	-1.756	-23.687	-7.380	-10.028
Fraser-Harris method	-9.847	-9.097	-68.857	-1.971	-0.912	-3.477	-23.552	-8.413	-19.595
Reed's criterion for means	†	-208.348	-599.974	†	-20.887	-30.295	192.830	-15.225	-26.254
Reed's criterion for measurements	†	-550.412	-1546.350	†	-55.178	-78.082	170.189	-58.245	-151.753
Tukey IQR rule	-2.889	-14.945	-79.765	-0.578	-1.498	-4.028	-7.164	-14.377	-22.614
Dixon's Q test	†	-550.412	-1546.350	†	-55.178	-78.082	170.189	-58.245	-151.753
Grubbs's test	†	-374.486	-895.841	†	-37.542	-45.235	168.646	-33.776	-49.583
± 3SD	8213.669	-174.464	-445.310	1643.760	-17.490	-22.486	226.644	-25.187	-46.029

† greater than 10^4.

Table D.67: Outlier detection methods with outlier simulation (2% and magnitude 10)–accuracy and coverage performance measures.

Outlier strategy	Mean squared error ($\times 10^{-4}$)			Accuracy and coverage Coverage			Mean 95% CI width		
	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G	σ_A	σ_I	σ_G
No outlier detection	†	**187.000**	**593.239**	**0.000**	$-^a$	$-^a$	**3.046**	$-^a$	$-^a$
Cochran C test	0.182	1.272	12.217	0.932	0.928	0.950^b	0.016	0.041	0.147^b
Cochran C test partial	0.181	1.160	12.153	0.933	0.940	0.950^b	0.016	0.041	0.146^b
Fraser-Harris method	0.184	1.178	12.822	0.934	0.938	0.946^b	0.016	0.042	0.145^b
Reed's criterion for means	†	191.618	558.224	0.000	$-^a$	1.000^c	2.440	$-^a$	0.753^c
Reed's criterion for measurements	†	119.596	342.954	0.000	$-^a$	1.000^c	2.576	$-^a$	0.680^c
Tukey IQR rule	0.163	1.103	13.078	0.945	0.945	0.939^b	0.016	0.041	0.143^b
Dixons Q test	†	119.596	342.954	0.000	$-^a$	1.000^c	2.576	$-^a$	0.680^c
Grubbs's test	†	136.954	406.693	0.000	$-^a$	1.000^c	1.997	$-^a$	0.680^c
± 3SD	8059.803	51.022	113.426	0.120	0.961^d	0.958^e	0.276	0.042^d	0.213^e

† greater than 10^4. ^{a}5000 CIs could not be calculated; ^{b}1 CIs could not be calculated; ^{c}4998 CIs could not be calculated; ^{d}4362 CIs could not be calculated; ^{e}4146 CIs could not be calculated.

Table D.68: Outlier detection methods with outlier simulation (2% and magnitude 10)–SD; median (Q1, Q3)[minimum, maximum]. True value of σ_A is 0.05, σ_I is 0.10 and σ_G is 0.20.

Outlier strategy	σ_A	σ_I	σ_G
No outlier detection	**10.07 (7.31, 12.12)[1.38, 14.71]**	**0.01 (0.00, 0.01)[0.00, 0.74]**	**0.00 (0.00, 0.03)[0.00, 1.04]**
Cochran C test	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.05, 0.14]	0.20 (0.17, 0.22)[0.05, 0.34]
Cochran C test partial	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.20 (0.17, 0.22)[0.05, 0.34]
Fraser-Harris method	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.05, 0.34]
Reed's criterion for means	7.49 (7.06, 10.28)[1.03, 14.46]	0.01 (0.00, 0.11)[0.00, 0.71]	0.00 (0.00, 0.21)[0.00, 0.97]
Reed's criterion for measurements	10.07 (1.38, 12.12)[1.17, 14.71]	0.01 (0.00, 0.01)[0.00, 0.74]	0.00 (0.00, 0.00)[0.00, 0.94]
Tukey IQR rule	0.05 (0.05, 0.05)[0.04, 0.06]	0.10 (0.09, 0.11)[0.06, 0.14]	0.19 (0.17, 0.21)[0.04, 0.34]
Dixon's Q test	10.07 (1.38, 12.12)[1.17, 14.71]	0.01 (0.00, 0.01)[0.00, 0.74]	0.00 (0.00, 0.00)[0.00, 0.94]
Grubbs's test	7.10 (1.24, 9.86)[1.17, 12.63]	0.01 (0.00, 0.09)[0.00, 0.64]	0.00 (0.00, 0.17)[0.00, 0.92]
± 3SD	1.00 (0.72, 1.04)[0.04, 1.26]	0.09 (0.00, 0.13)[0.00, 0.30]	0.17 (0.09, 0.22)[0.00, 0.46]

Table D.69: Outlier detection methods with outlier simulation (2% and magnitude 2)–CV; median (Q1, Q3)[minimum, maximum]. True value of CV_A is 5%, CV_I is 10% and CV_G is 20%. CVs are displayed as percentages.

Outlier strategy	CV_A	CV_I	CV_G
No outlier detection	† (†, †)**[237.60,** †]	**0.60 (0.40, 1.35)[0.01, 84.74]**	**0.00 (0.00, 2.97)[0.00, 139.16]**
Cochran C test	4.90 (4.62, 5.18)[3.53, 6.36]	9.84 (9.11, 10.61)[5.44, 14.20]	19.70 (17.22, 22.06)[4.86, 35.15]
Cochran C test partial	4.89 (4.62, 5.18)[3.57, 6.36]	9.91 (9.21, 10.63)[6.02, 14.20]	19.69 (17.23, 22.05)[4.62, 35.14]
Fraser-Harris method	4.89 (4.61, 5.18)[3.52, 6.36]	9.90 (9.18, 10.63)[6.02, 14.20]	19.31 (16.81, 21.70)[4.62, 35.14]
Reed's criterion for means	† (†, †)[136.99, †]	0.64 (0.38, 11.06)[0.01, 81.58]	0.00 (0.00, 20.90)[0.00, 125.83]
Reed's criterion for measurements	† (240.13, †)[170.61, †]	0.54 (0.34, 1.36)[0.01, 84.74]	0.00 (0.00, 0.07)[0.00, 119.23]
Tukey IQR rule	4.96 (4.69, 5.24)[3.59, 6.39]	9.85 (9.15, 10.55)[6.28, 14.49]	19.21 (16.73, 21.61)[4.40, 35.19]
Dixon's Q test	† (240.13, †)[170.61, †]	0.54 (0.34, 1.36)[0.01, 84.74]	0.00 (0.00, 0.07)[0.00, 119.23]
Grubbs's test	† (191.95, †)[170.61, †]	0.55 (0.32, 8.68)[0.01, 70.70]	0.00 (0.00, 17.48)[0.00, 115.28]
± 3SD	131.76 (82.17, 139.86)[3.79, 198.48]	9.23 (0.04, 13.30)[0.00, 30.24]	17.53 (8.61, 22.61)[0.00, 48.32]

† greater than 10^4.

Table D.70: Outlier detection methods with outlier simulation (2% and magnitude 10)–II, RCV and mean; median (Q1, Q3)[minimum, maximum]. True value of II is 0.56, RCV is 30.99% and mean is 10.

Outlier strategy	II	RCV	Mean
No outlier detection	† (†, †)**[5.39,** ∞]	† (†, †)**[658.59,** †]	**10.98 (10.40, 11.56)[9.61, 12.40]**
Cochran C test	0.56 (0.49, 0.65)[0.31, 2.18]	30.51 (28.75, 32.39)[20.09, 41.64]	9.97 (9.94, 10.00)[9.80, 10.13]
Cochran C test partial	0.56 (0.49, 0.65)[0.30, 2.29]	30.66 (28.92, 32.49)[22.37, 41.64]	9.97 (9.94, 10.00)[9.80, 10.13]
Fraser-Harris method	0.57 (0.50, 0.67)[0.30, 2.29]	30.62 (28.89, 32.46)[22.37, 41.64]	9.97 (9.94, 10.00)[9.80, 10.14]
Reed's criterion for means	† (†, †)[3.86, ∞]	† (†, †)[385.92, †]	10.49 (10.34, 11.08)[9.62, 12.37]
Reed's criterion for measurements	† (†, †)[3.75, ∞]	† (667.46, †)[478.98, †]	10.98 (9.84, 11.56)[9.61, 12.40]
Tukey IQR rule	0.58 (0.50, 0.67)[0.30, 2.45]	30.60 (28.89, 32.30)[22.63, 42.51]	9.97 (9.94, 10.01)[9.80, 10.14]
Dixon's Q test	† (†, †)[3.75, ∞]	† (667.46, †)[478.98, †]	10.98 (9.84, 11.56)[9.61, 12.40]
Grubbs's test	† (†, †)[3.75, ∞]	† (532.46, †)[478.98, †]	10.42 (9.84, 11.00)[9.64, 11.82]
± 3SD	6.94 (4.14, 16.88)[0.30, ∞]	366.95 (229.78, 387.69)[24.17, 550.16]	9.87 (9.82, 9.92)[9.64, 10.11]

† greater than 10^4.

Table D.71: Outlier detection methods with outlier simulation (2% and magnitude 10)–asymmetric RCVs; median (Q1, Q3)[minimum, maximum]. True lower RCV bound is -26.58% and upper bound is +36.20%. RCVs are displayed as percentages.

Outlier strategy	RCV lower bound	RCV upper bound
No outlier detection	**-100.00 (-100.00, -100.00)[-100.00, -97.80]**	† (†, †)**[4435.91,** †]
Cochran C test	-26.23 (-27.59, -24.93)[-33.90, -18.18]	35.55 (33.21, 38.09)[22.22, 51.29]
Cochran C test partial	-26.34 (-27.66, -25.06)[-33.90, -20.02]	35.75 (33.44, 38.23)[25.02, 51.29]
Fraser-Harris method	-26.31 (-27.64, -25.03)[-33.90, -20.02]	35.70 (33.39, 38.20)[25.02, 51.29]
Reed's criterion for means	-100.00 (-100.00, -100.00)[-100.00, -94.37]	† (†, †)[1677.44, †]
Reed's criterion for measurements	-100.00 (-100.00, -97.85)[-100.00, -96.16]	† (4540.83, †)[2503.65, †]
Tukey IQR rule	-26.29 (-27.52, -25.03)[-34.47, -20.22]	35.67 (33.39, 37.97)[25.35, 52.60]
Dixon's Q test	-100.00 (-100.00, -97.85)[-100.00, -96.16]	† (4540.83, †)[2503.65, †]
Grubbs's test	-100.00 (-100.00, -96.81)[-100.00, -96.16]	† (3037.07, †)[2503.65, †]
± 3SD	-93.85 (-94.42, -86.53)[-96.99, -21.43]	1526.72 (642.39, 1691.81)[27.28, 3221.93]

† greater than 10^4.

Appendix E

A review of monitoring-related methodology literature

E.1 Summary of identified studies

Table E.1: Summary of reviewed monitoring and monitoring related methodology literature.

Reference	Design					Analysis					Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Monitoring												
Stevens et al (2010)[12]						✓	✓			✓		8
Thompson and Pocock (1990)[126]							✓					26
Buclin et al (2011)[9]	✓	✓	✓				✓					6
Glasziou et al (2007)[121]					✓							195
Bell et al (2008)[122]					✓							19
Glasziou et al (2008)[14]							✓					55
Bell et al (2011)[127]							✓					8
Bell et al (2009)[123]							✓					36
Bell et al (2009)[128]					✓		✓					13
Keenan et al (2009)[10]							✓					40
Powers et al (2011)[129]							✓					50

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design					Analysis						Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Takahashi et al (2012)[117]	✓	✓					✓					2
Takahashi et al (2010)[118]	✓						✓					16
Oke et al (2012)[11]	✓						✓					2
Proust-Lima et al (2014)[132]								✓		✓		3
Proust-Lima and Taylor (2009)[133]								✓				25
Li and Gatsonis (2012)[31]	✓							✓				1
Slate and Turnbull (2000)[134]								✓				44
Bellera et al (2008)[135]								✓				12
Bellera et al (2009)[119]	✓		✓					✓				5
Inoue et al (2004)[136]								✓				23
Subtil and Rabilloud (2010)[137]								✓				3

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design						Analysis					Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Taylor et al (2005)[138]								✓			✓	31
Thiébaut et al (2003)[130]							✓				✓	18
Wolbers et al (2010)[131]							✓				✓	22
Sölétormos et al (2000)[120]	✓		✓									7
Bellera et al (2008)[29]	✓		✓					✓				2
DeLong et al (1985)[139]											✓	20
When to start consortium (2009)[124]					✓	✓					✓	389
Cole et al (2004)[140]											✓	28
Ahdieh-Grant (2003)[125]					✓						✓	32
Screening												

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design					Analysis						Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Walter and Day (1983)[141]					✓							91
Day and Walter (1984)[142]					✓							119
Etzioni and Shen (1997)[143]					✓							9
Zelen (1993)[144]	✓											49
Lee and Zelen (1998)[145]	✓											31
Frame and Frame (1998)[146]	✓											27
Lee et al (2004)[147]	✓											13
McIntosh et al (2002)[148]		✓										40
McIntosh and Urban (2003)[149]		✓										36
Time-dependent ROC curves												
Pepe et al (2008)[30]										✓	✓	24
Cai et al (2006)[150]											✓	26

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design					Analysis						Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Zheng and Heagerty (2004)[151]											✓	21
Subtil et al (2009)[152]											✓	3
Parker and DeLong (2003)[153]											✓	14
Slate and Turnbull (2000)[134]								✓				44
Etzioni et al (1999)[154]											✓	42
Variability												
Macaskill (2008)[155]					✓					✓	✓	0
Sölétormos et al (2000)[28]			✓								✓	3
Smellie (2008)[57]										✓	✓	15
Petersen (2005)[156]											✓	-
Fraser (2001)[16]											✓	-
Fraser et al (1990)[56]											✓	57
Klee (2010)[158]											✓	39

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design					Analysis						Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Petersen et al (2012)[157]											✓	5
Omar et al (2008)[162]					✓						✓	10
Biosca et al (2006)[163]					✓						✓	3
Clerico and Emdin (2004)[164]											✓	239
SPC												
Macaskill (2008)[155]					✓					✓	✓	0
Tennant et al (2007)[159]											✓[d]	18
Thor et al (2007)[160]										✓	✓[d]	90
Gavit et al (2009)[161]											✓	3
Decision analytic models												
Karnon et al (2007)[165]			✓	✓					✓	✓		20+5
Sutton et al (2008)[168]					✓				✓		✓	26

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Reference	Design					Analysis						Citations†
	Test frequency	Test thresholds	Decision rules	Review of methods	Other	General data structure	Linear mixed effects modelling/SNR	Joint modelling	Cost effectiveness	Review of methods	Other	
Baker (1998)[166]	✓								✓			8
Parmigiani (1997)[167]	✓								✓			10
Real options approaches												
Palmer and Smith (2000)[169]					✓						✓	66
Driffield and Smith (2007)[170]					✓						✓	17
Meyer and Rees (2012)[171]					✓						✓	1
Shechter et al (2010)[172]					✓						✓	0
Whynes (1995)[173]					✓						✓	2
Lasserre et al (2006)[174]					✓						✓	6

SNR is Signal to Noise ratio.

† Citations from Scopus search 24th July 2014

[a]review of current practice; [b] reporting guidelines; [c] review of reporting standards; [d] review of literature.

Appendix F

Simulating monitoring data and evaluating monitoring strategies

F.1 Detailed simulation results

Table F.1: Monitoring simulation–results using retest monitoring strategy (strategy B) by observation point.

Observation month	Tests[a] (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[b] n (%)	PPV (%)	Sensitivity (%)
0	28357/20000	1233 (6.2)	2447 (12.2)	700 (3.5)	15620 (78.1)	3680 (18.4)	1933 (9.7)	33.5	63.8
6	22924/16320	144 (0.9)	796 (4.9)	711 (4.4)	14669 (89.9)	940 (5.8)	855 (5.2)	15.3	16.8
12	21549/15380	125 (0.8)	568 (3.7)	734 (4.8)	13953 (90.7)	693 (4.5)	859 (5.6)	18.0	14.6
18	20617/14687	97 (0.7)	494 (3.4)	775 (5.3)	13321 (90.7)	591 (4.0)	872 (5.9)	16.4	11.1
24	19913/14096	117 (0.8)	438 (3.1)	776 (5.5)	12765 (90.6)	555 (3.9)	893 (6.3)	21.1	13.1
30	19167/13541	103 (0.8)	408 (3.0)	803 (5.9)	12227 (90.3)	511 (3.8)	906 (6.7)	20.2	11.4
36	18555/13030	105 (0.8)	413 (3.2)	816 (6.3)	11696 (89.8)	518 (4.0)	921 (7.1)	20.3	11.4
42	17927/12512	105 (0.8)	422 (3.4)	849 (6.8)	11136 (89.0)	527 (4.2)	954 (7.6)	19.9	11.0
48	17272/11985	108 (0.9)	408 (3.4)	858 (7.2)	10611 (88.5)	516 (4.3)	966 (8.1)	20.9	11.2
54	16674/11469	126 (1.1)	392 (3.4)	855 (7.5)	10096 (88.0)	518 (4.5)	981 (8.6)	24.3	12.8
60	16019/10951	88 (0.8)	269 (2.5)	887 (8.1)	9707 (88.6)	357 (3.3)	975 (8.9)	24.6	9.0
All	218974/153971	2351 (1.5)	7055 (4.6)	8764 (5.7)	135801 (88.2)	9406 (6.1)	11115 (7.2)	25.0	21.2

[a] Tests performed/number of people tests performed on (number of results generated).
[b] Tests performed when the patient was diseased.

Table F.2: Monitoring simulation–results using reduced frequency of monitoring strategy (strategy C) by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1247 (6.2)	2735 (13.7)	686 (3.4)	15332 (76.7)	3982 (19.9)	1933 (9.7)	31.3	64.5
12	16018	277 (1.7)	1253 (7.8)	739 (4.6)	13749 (85.8)	1530 (9.6)	1016 (6.3)	18.1	27.3
24	14488	243 (1.7)	983 (6.8)	775 (5.3)	12487 (86.2)	1226 (8.5)	1018 (7.0)	19.8	23.9
36	13262	247 (1.9)	952 (7.2)	782 (5.9)	11281 (85.1)	1199 (9.0)	1029 (7.8)	20.6	24.0
48	12063	246 (2.0)	861 (7.1)	806 (6.7)	10150 (84.1)	1107 (9.2)	1052 (8.7)	22.2	23.4
60	10956	252 (2.3)	757 (6.9)	809 (7.4)	9138 (83.4)	1009 (9.2)	1061 (9.7)	25.0	23.8
All	86787	2512 (2.9)	7541 (8.7)	4597 (5.3)	72137 (83.1)	10053 (11.6)	7109 (8.2)	25.0	35.3

[a] Tests performed when the patient was diseased.

Table F.3: Monitoring simulation–results using absolute increase from start value monitoring strategy (strategy D) by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	50 (0.3)	564 (3.4)	890 (5.4)	15098 (90.9)	614 (3.7)	940 (5.7)	8.1	5.3
12	15988	66 (0.4)	517 (3.2)	1014 (6.3)	14391 (90.0)	583 (3.6)	1080 (6.8)	11.3	6.1
18	15405	95 (0.6)	550 (3.6)	1085 (7.0)	13675 (88.8)	645 (4.2)	1180 (7.7)	14.7	8.1
24	14760	116 (0.8)	582 (3.9)	1132 (7.7)	12930 (87.6)	698 (4.7)	1248 (8.5)	16.6	9.3
30	14062	169 (1.2)	559 (4.0)	1123 (8.0)	12211 (86.8)	728 (5.2)	1292 (9.2)	23.2	13.1
36	13334	187 (1.4)	603 (4.5)	1066 (8.0)	11478 (86.1)	790 (5.9)	1253 (9.4)	23.7	14.9
42	12544	174 (1.4)	586 (4.7)	1048 (8.4)	10736 (85.6)	760 (6.1)	1222 (9.7)	22.9	14.2
48	11784	212 (1.8)	583 (4.9)	955 (8.1)	10034 (85.1)	795 (6.7)	1167 (9.9)	26.7	18.2
54	10989	218 (2.0)	591 (5.4)	865 (7.9)	9315 (84.8)	809 (7.4)	1083 (9.9)	26.9	20.1
60	10180	204 (2.0)	574 (5.6)	765 (7.5)	8637 (84.8)	778 (7.6)	969 (9.5)	26.2	21.1
All	155648	2653 (1.7)	7945 (5.1)	10714 (6.9)	134336 (86.3)	10598 (6.8)	13367 (8.6)	25.0	19.8

[a] Tests performed when the patient was diseased.

Table F.4: Monitoring simulation–results using absolute increase from last value monitoring strategy (strategy E) by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	28 (0.2)	342 (2.1)	912 (5.5)	15320 (92.3)	370 (2.2)	940 (5.7)	7.6	3.0
12	16232	22 (0.1)	281 (1.7)	1086 (6.7)	14843 (91.4)	303 (1.9)	1108 (6.8)	7.3	2.0
18	15929	32 (0.2)	256 (1.6)	1230 (7.7)	14411 (90.5)	288 (1.8)	1262 (7.9)	11.1	2.5
24	15641	31 (0.2)	256 (1.6)	1378 (8.8)	13976 (89.4)	287 (1.8)	1409 (9.0)	10.8	2.2
30	15354	34 (0.2)	240 (1.6)	1538 (10.0)	13542 (88.2)	274 (1.8)	1572 (10.2)	12.4	2.2
36	15080	39 (0.3)	248 (1.6)	1682 (11.2)	13111 (86.9)	287 (1.9)	1721 (11.4)	13.6	2.3
42	14793	52 (0.4)	228 (1.5)	1860 (12.6)	12653 (85.5)	280 (1.9)	1912 (12.9)	18.6	2.7
48	14513	56 (0.4)	223 (1.5)	2014 (13.9)	12220 (84.2)	279 (1.9)	2070 (14.3)	20.1	2.7
54	14234	45 (0.3)	204 (1.4)	2212 (15.5)	11773 (82.7)	249 (1.7)	2257 (15.9)	18.1	2.0
60	13985	78 (0.6)	212 (1.5)	2423 (17.3)	11272 (80.6)	290 (2.1)	2501 (17.9)	26.9	3.1
All	172363	1579 (0.9)	4726 (2.7)	17106 (9.9)	148952 (86.4)	6305 (3.7)	18685 (10.8)	25.0	8.5

[a] Tests performed when the patient was diseased.

Table F.5: Monitoring simulation–results using relative increase from start value monitoring strategy (strategy F) by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	44 (0.3)	641 (3.9)	896 (5.4)	15021 (90.5)	685 (4.1)	940 (5.7)	6.4	4.7
12	15917	48 (0.3)	539 (3.4)	1041 (6.5)	14289 (89.8)	587 (3.7)	1089 (6.8)	8.2	4.4
18	15330	84 (0.5)	525 (3.4)	1125 (7.3)	13596 (88.7)	609 (4.0)	1209 (7.9)	13.8	6.9
24	14721	100 (0.7)	558 (3.8)	1193 (8.1)	12870 (87.4)	658 (4.5)	1293 (8.8)	15.2	7.7
30	14063	143 (1.0)	524 (3.7)	1217 (8.7)	12179 (86.6)	667 (4.7)	1360 (9.7)	21.4	10.5
36	13396	169 (1.3)	519 (3.9)	1184 (8.8)	11524 (86.0)	688 (5.1)	1353 (10.1)	24.6	12.5
42	12708	179 (1.4)	528 (4.2)	1164 (9.2)	10837 (85.3)	707 (5.6)	1343 (10.6)	25.3	13.3
48	12001	195 (1.6)	568 (4.7)	1099 (9.2)	10139 (84.5)	763 (6.4)	1294 (10.8)	25.6	15.1
54	11238	226 (2.0)	528 (4.7)	1009 (9.0)	9475 (84.3)	754 (6.7)	1235 (11.0)	30.0	18.3
60	10484	220 (2.1)	530 (5.1)	918 (8.8)	8816 (84.1)	750 (7.2)	1138 (10.9)	29.3	19.3
All	156460	2570 (1.6)	7696 (4.9)	11617 (7.4)	134577 (86.0)	10266 (6.6)	14187 (9.1)	25.0	18.1

[a] Tests performed when the patient was diseased.

Table F.6: Monitoring simulation–results using relative increase from last value monitoring strategy (strategy G) by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	16 (0.1)	263 (1.6)	924 (5.6)	15399 (92.8)	279 (1.7)	940 (5.7)	5.7	1.7
12	16323	9 (0.1)	206 (1.3)	1117 (6.8)	14991 (91.8)	215 (1.3)	1126 (6.9)	4.2	0.8
18	16108	18 (0.1)	199 (1.2)	1280 (7.9)	14611 (90.7)	217 (1.3)	1298 (8.1)	8.3	1.4
24	15891	13 (0.1)	184 (1.2)	1455 (9.2)	14239 (89.6)	197 (1.2)	1468 (9.2)	6.6	0.9
30	15694	16 (0.1)	170 (1.1)	1642 (10.5)	13866 (88.4)	186 (1.2)	1658 (10.6)	8.6	1.0
36	15508	13 (0.1)	172 (1.1)	1829 (11.8)	13494 (87.0)	185 (1.2)	1842 (11.9)	7.0	0.7
42	15323	22 (0.1)	161 (1.1)	2054 (13.4)	13086 (85.4)	183 (1.2)	2076 (13.5)	12.0	1.1
48	15140	25 (0.2)	140 (0.9)	2264 (15.0)	12711 (84.0)	165 (1.1)	2289 (15.1)	15.2	1.1
54	14975	20 (0.1)	134 (0.9)	2511 (16.8)	12310 (82.2)	154 (1.0)	2531 (16.9)	13.0	0.8
60	14821	21 (0.1)	138 (0.9)	2809 (19.0)	11853 (80.0)	159 (1.1)	2830 (19.1)	13.2	0.7
All	176385	1335 (0.8)	4003 (2.3)	18656 (10.6)	152391 (86.4)	5338 (3.0)	19991 (11.3)	25.0	6.7

[a] Tests performed when the patient was diseased.

Table F.7: Monitoring simulation–results using linear regression monitoring strategy (strategy H) by observation point.

Observation month	**Tests (n)**	**TP results n (%)**	**FP results n (%)**	**FN results n (%)**	**TN results n (%)**	**Positive results n (%)**	**Diseased**[a] **n (%)**	**PPV (%)**	**Sensitivity (%)**
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	280 (1.7)	1488 (9.0)	660 (4.0)	14174 (85.4)	1768 (10.6)	940 (5.7)	15.8	29.8
12	14834	146 (1.0)	741 (5.0)	650 (4.4)	13297 (89.6)	887 (6.0)	796 (5.4)	16.5	18.3
18	13947	92 (0.7)	492 (3.5)	685 (4.9)	12678 (90.9)	584 (4.2)	777 (5.6)	15.8	11.8
24	13363	87 (0.7)	385 (2.9)	706 (5.3)	12185 (91.2)	472 (3.5)	793 (5.9)	18.4	11.0
30	12891	93 (0.7)	307 (2.4)	742 (5.8)	11749 (91.1)	400 (3.1)	835 (6.5)	23.2	11.1
36	12491	85 (0.7)	284 (2.3)	772 (6.2)	11350 (90.9)	369 (3.0)	857 (6.9)	23.0	9.9
42	12122	68 (0.6)	268 (2.2)	838 (6.9)	10948 (90.3)	336 (2.8)	906 (7.5)	20.2	7.5
48	11786	100 (0.8)	273 (2.3)	859 (7.3)	10554 (89.5)	373 (3.2)	959 (8.1)	26.8	10.4
54	11413	107 (0.9)	277 (2.4)	880 (7.7)	10149 (88.9)	384 (3.4)	987 (8.6)	27.9	10.8
60	11029	111 (1.0)	260 (2.4)	906 (8.2)	9752 (88.4)	371 (3.4)	1017 (9.2)	29.9	10.9
All	150478	2331 (1.5)	7011 (4.7)	8469 (5.6)	132667 (88.2)	9342 (6.2)	10800 (7.2)	25.0	21.6

[a] Tests performed when the patient was diseased.

F.2 Detailed simulation results–sensitivity analyses

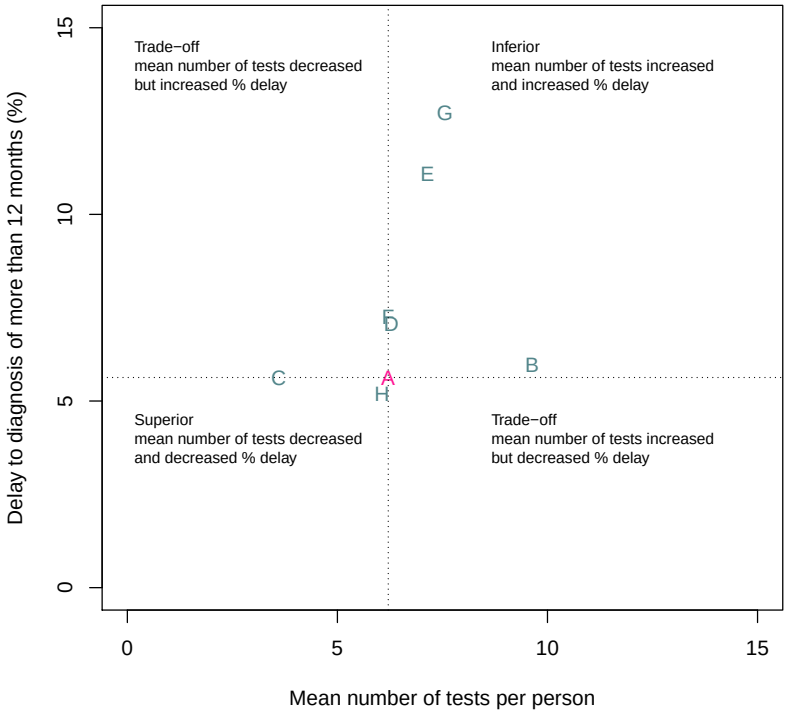


Figure F.1: Adjusted fibrosis progression rate–performance of various monitoring strategies on simulated monitoring data with PPV of 25%. A is the simple threshold strategy; B is the retest strategy; C is the decreased monitoring frequency strategy; D is the absolute increase from initial value strategy; E is the absolute increase from last value strategy; F is the relative increase from initial value strategy; G is the relative increase from last value strategy; H is the linear regression strategy.

Table F.8: Monitoring simulation–results using reference strategy with decreased measurement error by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1153 (5.8)	1931 (9.7)	750 (3.8)	16166 (80.8)	3084 (15.4)	1903 (9.5)	37.4	60.6
6	16916	113 (0.7)	478 (2.8)	810 (4.8)	15515 (91.7)	591 (3.5)	923 (5.5)	19.1	12.2
12	16325	92 (0.6)	371 (2.3)	881 (5.4)	14981 (91.8)	463 (2.8)	973 (6.0)	19.9	9.5
18	15862	109 (0.7)	341 (2.1)	934 (5.9)	14478 (91.3)	450 (2.8)	1043 (6.6)	24.2	10.5
24	15412	103 (0.7)	331 (2.1)	967 (6.3)	14011 (90.9)	434 (2.8)	1070 (6.9)	23.7	9.6
30	14978	93 (0.6)	356 (2.4)	1025 (6.8)	13504 (90.2)	449 (3.0)	1118 (7.5)	20.7	8.3
36	14529	129 (0.9)	353 (2.4)	1032 (7.1)	13015 (89.6)	482 (3.3)	1161 (8.0)	26.8	11.1
42	14047	128 (0.9)	356 (2.5)	1061 (7.6)	12502 (89.0)	484 (3.4)	1189 (8.5)	26.4	10.8
48	13563	112 (0.8)	341 (2.5)	1086 (8.0)	12024 (88.7)	453 (3.3)	1198 (8.8)	24.7	9.3
54	13110	137 (1.0)	334 (2.5)	1105 (8.4)	11534 (88.0)	471 (3.6)	1242 (9.5)	29.1	11.0
60	12639	164 (1.3)	370 (2.9)	1095 (8.7)	11010 (87.1)	534 (4.2)	1259 (10.0)	30.7	13.0
All	167381	2333 (1.4)	5562 (3.3)	10746 (6.4)	148740 (88.9)	7895 (4.7)	13079 (7.8)	29.6	17.8

[a] Tests performed when the patient was diseased.

Table F.9: Monitoring simulation–results using reference strategy with decreased measurement error by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1289 (6.4)	2755 (13.8)	614 (3.1)	15342 (76.7)	4044 (20.2)	1903 (9.5)	31.9	67.7
6	15956	114 (0.7)	665 (4.2)	645 (4.0)	14532 (91.1)	779 (4.9)	759 (4.8)	14.6	15.0
12	15177	93 (0.6)	441 (2.9)	686 (4.5)	13957 (92.0)	534 (3.5)	779 (5.1)	17.4	11.9
18	14643	97 (0.7)	431 (2.9)	726 (5.0)	13389 (91.4)	528 (3.6)	823 (5.6)	18.4	11.8
24	14115	83 (0.6)	438 (3.1)	761 (5.4)	12833 (90.9)	521 (3.7)	844 (6.0)	15.9	9.8
30	13594	111 (0.8)	410 (3.0)	772 (5.7)	12301 (90.5)	521 (3.8)	883 (6.5)	21.3	12.6
36	13073	100 (0.8)	399 (3.1)	785 (6.0)	11789 (90.2)	499 (3.8)	885 (6.8)	20.0	11.3
42	12574	108 (0.9)	361 (2.9)	815 (6.5)	11290 (89.8)	469 (3.7)	923 (7.3)	23.0	11.7
48	12105	108 (0.9)	379 (3.1)	818 (6.8)	10800 (89.2)	487 (4.0)	926 (7.6)	22.2	11.7
54	11618	133 (1.1)	383 (3.3)	817 (7.0)	10285 (88.5)	516 (4.4)	950 (8.2)	25.8	14.0
60	11102	118 (1.1)	393 (3.5)	815 (7.3)	9776 (88.1)	511 (4.6)	933 (8.4)	23.1	12.6
All	153957	2354 (1.5)	7055 (4.6)	8254 (5.4)	136294 (88.5)	9409 (6.1)	10608 (6.9)	25.0	22.2

[a] Tests performed when the patient was diseased.

Table F.10: Monitoring simulation–results using reference strategy with increased measurement error by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1169 (5.8)	3257 (16.3)	812 (4.1)	14762 (73.8)	4426 (22.1)	1981 (9.9)	26.4	59.0
6	15574	369 (2.4)	1897 (12.2)	621 (4.0)	12687 (81.5)	2266 (14.5)	990 (6.4)	16.3	37.3
12	13308	213 (1.6)	1360 (10.2)	540 (4.1)	11195 (84.1)	1573 (11.8)	753 (5.7)	13.5	28.3
18	11735	141 (1.2)	1130 (9.6)	495 (4.2)	9969 (85.0)	1271 (10.8)	636 (5.4)	11.1	22.2
24	10464	123 (1.2)	881 (8.4)	455 (4.3)	9005 (86.1)	1004 (9.6)	578 (5.5)	12.3	21.3
30	9460	99 (1.0)	753 (8.0)	430 (4.5)	8178 (86.4)	852 (9.0)	529 (5.6)	11.6	18.7
36	8608	93 (1.1)	682 (7.9)	408 (4.7)	7425 (86.3)	775 (9.0)	501 (5.8)	12.0	18.6
42	7833	88 (1.1)	534 (6.8)	404 (5.2)	6807 (86.9)	622 (7.9)	492 (6.3)	14.1	17.9
48	7211	86 (1.2)	514 (7.1)	391 (5.4)	6220 (86.3)	600 (8.3)	477 (6.6)	14.3	18.0
54	6611	78 (1.2)	454 (6.9)	388 (5.9)	5691 (86.1)	532 (8.0)	466 (7.0)	14.7	16.7
60	6079	77 (1.3)	409 (6.7)	375 (6.2)	5218 (85.8)	486 (8.0)	452 (7.4)	15.8	17.0
All	116883	2536 (2.2)	11871 (10.2)	5319 (4.6)	97157 (83.1)	14407 (12.3)	7855 (6.7)	17.6	32.3

[a] Tests performed when the patient was diseased.

Table F.11: Monitoring simulation–results using reference strategy with increased measurement error by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	848 (4.2)	1562 (7.8)	1133 (5.7)	16457 (82.3)	2410 (12.0)	1981 (9.9)	35.2	42.8
6	17590	368 (2.1)	1081 (6.1)	1001 (5.7)	15140 (86.1)	1449 (8.2)	1369 (7.8)	25.4	26.9
12	16141	274 (1.7)	827 (5.1)	942 (5.8)	14098 (87.3)	1101 (6.8)	1216 (7.5)	24.9	22.5
18	15040	189 (1.3)	726 (4.8)	913 (6.1)	13212 (87.8)	915 (6.1)	1102 (7.3)	20.7	17.2
24	14125	160 (1.1)	583 (4.1)	889 (6.3)	12493 (88.4)	743 (5.3)	1049 (7.4)	21.5	15.3
30	13382	139 (1.0)	580 (4.3)	881 (6.6)	11782 (88.0)	719 (5.4)	1020 (7.6)	19.3	13.6
36	12663	130 (1.0)	596 (4.7)	861 (6.8)	11076 (87.5)	726 (5.7)	991 (7.8)	17.9	13.1
42	11937	121 (1.0)	473 (4.0)	883 (7.4)	10460 (87.6)	594 (5.0)	1004 (8.4)	20.4	12.1
48	11343	129 (1.1)	497 (4.4)	877 (7.7)	9840 (86.7)	626 (5.5)	1006 (8.9)	20.6	12.8
54	10717	119 (1.1)	441 (4.1)	883 (8.2)	9274 (86.5)	560 (5.2)	1002 (9.3)	21.2	11.9
60	10157	125 (1.2)	433 (4.3)	882 (8.7)	8717 (85.8)	558 (5.5)	1007 (9.9)	22.4	12.4
All	153095	2602 (1.7)	7799 (5.1)	10145 (6.6)	132549 (86.6)	10401 (6.8)	12747 (8.3)	25.0	20.4

[a] Tests performed when the patient was diseased.

Table F.12: Monitoring simulation–results using reference strategy with decreased between-individual variability by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1171 (5.9)	963 (4.8)	639 (3.2)	17227 (86.1)	2134 (10.7)	1810 (9.0)	54.9	64.7
6	17866	290 (1.6)	548 (3.1)	514 (2.9)	16514 (92.4)	838 (4.7)	804 (4.5)	34.6	36.1
12	17028	185 (1.1)	421 (2.5)	500 (2.9)	15922 (93.5)	606 (3.6)	685 (4.0)	30.5	27.0
18	16422	181 (1.1)	406 (2.5)	464 (2.8)	15371 (93.6)	587 (3.6)	645 (3.9)	30.8	28.1
24	15835	128 (0.8)	391 (2.5)	447 (2.8)	14869 (93.9)	519 (3.3)	575 (3.6)	24.7	22.3
30	15316	117 (0.8)	373 (2.4)	449 (2.9)	14377 (93.9)	490 (3.2)	566 (3.7)	23.9	20.7
36	14826	127 (0.9)	445 (3.0)	434 (2.9)	13820 (93.2)	572 (3.9)	561 (3.8)	22.2	22.6
42	14254	117 (0.8)	424 (3.0)	436 (3.1)	13277 (93.1)	541 (3.8)	553 (3.9)	21.6	21.2
48	13713	127 (0.9)	471 (3.4)	431 (3.1)	12684 (92.5)	598 (4.4)	558 (4.1)	21.2	22.8
54	13115	136 (1.0)	424 (3.2)	435 (3.3)	12120 (92.4)	560 (4.3)	571 (4.4)	24.3	23.8
60	12555	124 (1.0)	474 (3.8)	442 (3.5)	11515 (91.7)	598 (4.8)	566 (4.5)	20.7	21.9
All	170930	2703 (1.6)	5340 (3.1)	5191 (3.0)	157696 (92.3)	8043 (4.7)	7894 (4.6)	33.6	34.2

[a] Tests performed when the patient was diseased.

Table F.13: Monitoring simulation–results using reference strategy with decreased between-individual variability by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1333 (6.7)	1517 (7.6)	477 (2.4)	16673 (83.4)	2850 (14.2)	1810 (9.0)	46.8	73.6
6	17150	259 (1.5)	826 (4.8)	333 (1.9)	15732 (91.7)	1085 (6.3)	592 (3.5)	23.9	43.8
12	16065	156 (1.0)	655 (4.1)	298 (1.9)	14956 (93.1)	811 (5.0)	454 (2.8)	19.2	34.4
18	15254	131 (0.9)	558 (3.7)	271 (1.8)	14294 (93.7)	689 (4.5)	402 (2.6)	19.0	32.6
24	14565	86 (0.6)	562 (3.9)	253 (1.7)	13664 (93.8)	648 (4.4)	339 (2.3)	13.3	25.4
30	13917	79 (0.6)	552 (4.0)	256 (1.8)	13030 (93.6)	631 (4.5)	335 (2.4)	12.5	23.6
36	13286	96 (0.7)	576 (4.3)	241 (1.8)	12373 (93.1)	672 (5.1)	337 (2.5)	14.3	28.5
42	12614	72 (0.6)	529 (4.2)	238 (1.9)	11775 (93.3)	601 (4.8)	310 (2.5)	12.0	23.2
48	12013	85 (0.7)	575 (4.8)	250 (2.1)	11103 (92.4)	660 (5.5)	335 (2.8)	12.9	25.4
54	11353	107 (0.9)	565 (5.0)	236 (2.1)	10445 (92.0)	672 (5.9)	343 (3.0)	15.9	31.2
60	10681	84 (0.8)	535 (5.0)	236 (2.2)	9826 (92.0)	619 (5.8)	320 (3.0)	13.6	26.2
All	156898	2488 (1.6)	7450 (4.7)	3089 (2.0)	143871 (91.7)	9938 (6.3)	5577 (3.6)	25.0	44.6

[a] Tests performed when the patient was diseased.

Table F.14: Monitoring simulation–results using reference strategy with increased between-individual variability by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1279 (6.4)	5806 (29.0)	656 (3.3)	12259 (61.3)	7085 (35.4)	1935 (9.7)	18.1	66.1
6	12915	104 (0.8)	1193 (9.2)	685 (5.3)	10933 (84.7)	1297 (10.0)	789 (6.1)	8.0	13.2
12	11618	78 (0.7)	768 (6.6)	707 (6.1)	10065 (86.6)	846 (7.3)	785 (6.8)	9.2	9.9
18	10772	81 (0.8)	589 (5.5)	734 (6.8)	9368 (87.0)	670 (6.2)	815 (7.6)	12.1	9.9
24	10102	85 (0.8)	468 (4.6)	757 (7.5)	8792 (87.0)	553 (5.5)	842 (8.3)	15.4	10.1
30	9549	84 (0.9)	392 (4.1)	745 (7.8)	8328 (87.2)	476 (5.0)	829 (8.7)	17.6	10.1
36	9073	78 (0.9)	368 (4.1)	758 (8.4)	7869 (86.7)	446 (4.9)	836 (9.2)	17.5	9.3
42	8627	85 (1.0)	356 (4.1)	774 (9.0)	7412 (85.9)	441 (5.1)	859 (10.0)	19.3	9.9
48	8186	90 (1.1)	302 (3.7)	768 (9.4)	7026 (85.8)	392 (4.8)	858 (10.5)	23.0	10.5
54	7794	99 (1.3)	285 (3.7)	779 (10.0)	6631 (85.1)	384 (4.9)	878 (11.3)	25.8	11.3
60	7410	79 (1.1)	305 (4.1)	794 (10.7)	6232 (84.1)	384 (5.2)	873 (11.8)	20.6	9.0
All	116046	2142 (1.8)	10832 (9.3)	8157 (7.0)	94915 (81.8)	12974 (11.2)	10299 (8.9)	16.5	20.8

[a] Tests performed when the patient was diseased.

Table F.15: Monitoring simulation–results using reference strategy with increased between-individual variability by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	890 (4.4)	2368 (11.8)	1045 (5.2)	15697 (78.5)	3258 (16.3)	1935 (9.7)	27.3	46.0
6	16742	117 (0.7)	610 (3.6)	1127 (6.7)	14888 (88.9)	727 (4.3)	1244 (7.4)	16.1	9.4
12	16015	98 (0.6)	413 (2.6)	1208 (7.5)	14296 (89.3)	511 (3.2)	1306 (8.2)	19.2	7.5
18	15504	102 (0.7)	339 (2.2)	1286 (8.3)	13777 (88.9)	441 (2.8)	1388 (9.0)	23.1	7.3
24	15063	81 (0.5)	332 (2.2)	1399 (9.3)	13251 (88.0)	413 (2.7)	1480 (9.8)	19.6	5.5
30	14650	78 (0.5)	308 (2.1)	1490 (10.2)	12774 (87.2)	386 (2.6)	1568 (10.7)	20.2	5.0
36	14264	114 (0.8)	295 (2.1)	1535 (10.8)	12320 (86.4)	409 (2.9)	1649 (11.6)	27.9	6.9
42	13855	103 (0.7)	276 (2.0)	1606 (11.6)	11870 (85.7)	379 (2.7)	1709 (12.3)	27.2	6.0
48	13476	122 (0.9)	318 (2.4)	1641 (12.2)	11395 (84.6)	440 (3.3)	1763 (13.1)	27.7	6.9
54	13036	103 (0.8)	271 (2.1)	1742 (13.4)	10920 (83.8)	374 (2.9)	1845 (14.2)	27.5	5.6
60	12662	130 (1.0)	286 (2.3)	1817 (14.4)	10429 (82.4)	416 (3.3)	1947 (15.4)	31.2	6.7
All	165267	1938 (1.2)	5816 (3.5)	15896 (9.6)	141617 (85.7)	7754 (4.7)	17834 (10.8)	25.0	10.9

[a] Tests performed when the patient was diseased.

Table F.16: Monitoring simulation–results using reference strategy with decreased fibrosis progression rate by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1038 (5.2)	2122 (10.6)	698 (3.5)	16142 (80.7)	3160 (15.8)	1736 (8.7)	32.8	59.8
6	16840	172 (1.0)	848 (5.0)	631 (3.7)	15189 (90.2)	1020 (6.1)	803 (4.8)	16.9	21.4
12	15820	110 (0.7)	603 (3.8)	617 (3.9)	14490 (91.6)	713 (4.5)	727 (4.6)	15.4	15.1
18	15107	97 (0.6)	580 (3.8)	624 (4.1)	13806 (91.4)	677 (4.5)	721 (4.8)	14.3	13.5
24	14430	86 (0.6)	448 (3.1)	648 (4.5)	13248 (91.8)	534 (3.7)	734 (5.1)	16.1	11.7
30	13896	84 (0.6)	402 (2.9)	655 (4.7)	12755 (91.8)	486 (3.5)	739 (5.3)	17.3	11.4
36	13410	82 (0.6)	459 (3.4)	677 (5.0)	12192 (90.9)	541 (4.0)	759 (5.7)	15.2	10.8
42	12869	102 (0.8)	413 (3.2)	667 (5.2)	11687 (90.8)	515 (4.0)	769 (6.0)	19.8	13.3
48	12354	108 (0.9)	416 (3.4)	646 (5.2)	11184 (90.5)	524 (4.2)	754 (6.1)	20.6	14.3
54	11830	84 (0.7)	371 (3.1)	663 (5.6)	10712 (90.5)	455 (3.8)	747 (6.3)	18.5	11.2
60	11375	96 (0.8)	351 (3.1)	678 (6.0)	10250 (90.1)	447 (3.9)	774 (6.8)	21.5	12.4
All	157931	2059 (1.3)	7013 (4.4)	7204 (4.6)	141655 (89.7)	9072 (5.7)	9263 (5.9)	22.7	22.2

[a] Tests performed when the patient was diseased.

Table F.17: Monitoring simulation–results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	960 (4.8)	1707 (8.5)	776 (3.9)	16557 (82.8)	2667 (13.3)	1736 (8.7)	36.0	55.3
6	17333	184 (1.1)	730 (4.2)	707 (4.1)	15712 (90.6)	914 (5.3)	891 (5.1)	20.1	20.7
12	16419	125 (0.8)	516 (3.1)	692 (4.2)	15086 (91.9)	641 (3.9)	817 (5.0)	19.5	15.3
18	15778	101 (0.6)	466 (3.0)	706 (4.5)	14505 (91.9)	567 (3.6)	807 (5.1)	17.8	12.5
24	15211	96 (0.6)	395 (2.6)	735 (4.8)	13985 (91.9)	491 (3.2)	831 (5.5)	19.6	11.6
30	14720	77 (0.5)	393 (2.7)	760 (5.2)	13490 (91.6)	470 (3.2)	837 (5.7)	16.4	9.2
36	14250	83 (0.6)	427 (3.0)	792 (5.6)	12948 (90.9)	510 (3.6)	875 (6.1)	16.3	9.5
42	13740	101 (0.7)	374 (2.7)	793 (5.8)	12472 (90.8)	475 (3.5)	894 (6.5)	21.3	11.3
48	13265	110 (0.8)	381 (2.9)	783 (5.9)	11991 (90.4)	491 (3.7)	893 (6.7)	22.4	12.3
54	12774	100 (0.8)	379 (3.0)	795 (6.2)	11500 (90.0)	479 (3.7)	895 (7.0)	20.9	11.2
60	12295	105 (0.9)	342 (2.8)	813 (6.6)	11035 (89.8)	447 (3.6)	918 (7.5)	23.5	11.4
All	165785	2042 (1.2)	6110 (3.7)	8352 (5.0)	149281 (90.0)	8152 (4.9)	10394 (6.3)	25.0	19.6

[a] Tests performed when the patient was diseased.

Table F.18: Monitoring simulation–results using reference strategy with increased fibrosis progression rate by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1535 (7.7)	2428 (12.1)	947 (4.7)	15090 (75.4)	3963 (19.8)	2482 (12.4)	38.7	61.8
6	16037	269 (1.7)	990 (6.2)	909 (5.7)	13869 (86.5)	1259 (7.9)	1178 (7.3)	21.4	22.8
12	14778	200 (1.4)	772 (5.2)	930 (6.3)	12876 (87.1)	972 (6.6)	1130 (7.6)	20.6	17.7
18	13806	186 (1.3)	649 (4.7)	945 (6.8)	12026 (87.1)	835 (6.0)	1131 (8.2)	22.3	16.4
24	12971	163 (1.3)	607 (4.7)	938 (7.2)	11263 (86.8)	770 (5.9)	1101 (8.5)	21.2	14.8
30	12201	165 (1.4)	557 (4.6)	975 (8.0)	10504 (86.1)	722 (5.9)	1140 (9.3)	22.9	14.5
36	11479	196 (1.7)	596 (5.2)	969 (8.4)	9718 (84.7)	792 (6.9)	1165 (10.1)	24.7	16.8
42	10687	171 (1.6)	506 (4.7)	974 (9.1)	9036 (84.6)	677 (6.3)	1145 (10.7)	25.3	14.9
48	10010	204 (2.0)	512 (5.1)	979 (9.8)	8315 (83.1)	716 (7.2)	1183 (11.8)	28.5	17.2
54	9294	197 (2.1)	440 (4.7)	1009 (10.9)	7648 (82.3)	637 (6.9)	1206 (13.0)	30.9	16.3
60	8657	205 (2.4)	413 (4.8)	1009 (11.7)	7030 (81.2)	618 (7.1)	1214 (14.0)	33.2	16.9
All	139920	3491 (2.5)	8470 (6.1)	10584 (7.6)	117375 (83.9)	11961 (8.5)	14075 (10.1)	29.2	24.8

[a] Tests performed when the patient was diseased.

Table F.19: Monitoring simulation–results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%.

Observation month	**Tests (n)**	**TP results n (%)**	**FP results n (%)**	**FN results n (%)**	**TN results n (%)**	**Positive results n (%)**	**Diseased**[a] **n (%)**	**PPV (%)**	**Sensitivity (%)**
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	271 (1.8)	1308 (8.7)	709 (4.7)	12676 (84.7)	1579 (10.6)	980 (6.5)	17.2	27.7
12	13385	183 (1.4)	979 (7.3)	699 (5.2)	11524 (86.1)	1162 (8.7)	882 (6.6)	15.7	20.7
18	12223	174 (1.4)	743 (6.1)	705 (5.8)	10601 (86.7)	917 (7.5)	879 (7.2)	19.0	19.8
24	11306	147 (1.3)	691 (6.1)	680 (6.0)	9788 (86.6)	838 (7.4)	827 (7.3)	17.5	17.8
30	10468	139 (1.3)	566 (5.4)	693 (6.6)	9070 (86.6)	705 (6.7)	832 (7.9)	19.7	16.7
36	9763	139 (1.4)	591 (6.1)	699 (7.2)	8334 (85.4)	730 (7.5)	838 (8.6)	19.0	16.6
42	9033	146 (1.6)	553 (6.1)	684 (7.6)	7650 (84.7)	699 (7.7)	830 (9.2)	20.9	17.6
48	8334	159 (1.9)	472 (5.7)	689 (8.3)	7014 (84.2)	631 (7.6)	848 (10.2)	25.2	18.8
54	7703	162 (2.1)	465 (6.0)	706 (9.2)	6370 (82.7)	627 (8.1)	868 (11.3)	25.8	18.7
60	7076	163 (2.3)	450 (6.4)	712 (10.1)	5751 (81.3)	613 (8.7)	875 (12.4)	26.6	18.6
All	124255	3387 (2.7)	10150 (8.2)	7754 (6.2)	102964 (82.9)	13537 (10.9)	11141 (9.0)	25.0	30.4

[a] Tests performed when the patient was diseased.

Table F.20: Monitoring simulation–results of strategies A-H for adjusted fibrosis progression estimate data.

	Monitoring strategy components					**PPV**	**Tests**[a]			**Delay to diagnosis**[c]			**Test performance**			
	Decision rule	Threshold value	Interval (months)	Retest	Initial threshold	%	Total	Mean pp[b]	Median (Q1, Q3)	N	% of all[d]	% of stage 4[e]	TP pp[f]	FP pp[g]	Positive n (%)	Sensitivity (%)
A	Simple threshold	10.460	6	FALSE	-	25	124255	6.21	6 (1,11)	1126	5.63	14.64	0.17	0.51	13537 (10.89)	30.40
B	Simple threshold	10.325	6	TRUE	-	25	192590†	9.63	11 (2,15)	1194	5.97	15.53	0.16	0.49	13048 (10.43)	29.88
C	Simple threshold	10.265	12	FALSE	-	25	72001	3.60	4 (1,6)	1128	5.64	14.67	0.17	0.52	13814 (19.19)	47.02
D	Absolute increase from initial value	1.138	6	FALSE	10.46	25	125699	6.28	7 (1,11)	1415	7.07	18.40	0.18	0.54	14509 (11.54)	28.57
E	Absolute increase from last value	1.245	6	FALSE	10.46	25	142849	7.14	10 (1,11)	2220	11.10	28.87	0.13	0.40	10587 (7.41)	13.40
F	Relative increase from initial value	1.122	6	FALSE	10.46	25	124237	6.21	6 (1,11)	1450	7.25	18.86	0.18	0.55	14565 (11.72)	27.98
G	Relative increase from last value	1.154	6	FALSE	10.46	25	151266	7.56	11 (1,11)	2547	12.73	33.13	0.11	0.33	8673 (5.73)	9.60
H	Linear regression	10.235	6	FALSE	10.46	25	121214	6.06	5 (1,11)	1039	5.20	13.51	0.16	0.48	12930 (10.67)	30.01

[a] Tests over the duration of monitoring.
[b] Mean number of tests per person over the duration of monitoring.
[c] Patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).
[d] % of all patients with delay to diagnosis.
[e] % of patients that would reach cirrhosis within the trial period with delay to diagnosis.
[f] TP pp is the mean number of true positive results per person over the duration of monitoring.
[g] FP pp is the mean number of false positive results per person over the duration of monitoring.
† 192590 tests were carried out to generate 125091 results due to retests being used.

Table F.21: Monitoring simulation–results by observation point for the reference strategy (strategy A) for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	271 (1.8)	1308 (8.7)	709 (4.7)	12676 (84.7)	1579 (10.6)	980 (6.5)	17.2	27.7
12	13385	183 (1.4)	979 (7.3)	699 (5.2)	11524 (86.1)	1162 (8.7)	882 (6.6)	15.7	20.7
18	12223	174 (1.4)	743 (6.1)	705 (5.8)	10601 (86.7)	917 (7.5)	879 (7.2)	19.0	19.8
24	11306	147 (1.3)	691 (6.1)	680 (6.0)	9788 (86.6)	838 (7.4)	827 (7.3)	17.5	17.8
30	10468	139 (1.3)	566 (5.4)	693 (6.6)	9070 (86.6)	705 (6.7)	832 (7.9)	19.7	16.7
36	9763	139 (1.4)	591 (6.1)	699 (7.2)	8334 (85.4)	730 (7.5)	838 (8.6)	19.0	16.6
42	9033	146 (1.6)	553 (6.1)	684 (7.6)	7650 (84.7)	699 (7.7)	830 (9.2)	20.9	17.6
48	8334	159 (1.9)	472 (5.7)	689 (8.3)	7014 (84.2)	631 (7.6)	848 (10.2)	25.2	18.8
54	7703	162 (2.1)	465 (6.0)	706 (9.2)	6370 (82.7)	627 (8.1)	868 (11.3)	25.8	18.7
60	7076	163 (2.3)	450 (6.4)	712 (10.1)	5751 (81.3)	613 (8.7)	875 (12.4)	26.6	18.6
All	124255	3387 (2.7)	10150 (8.2)	7754 (6.2)	102964 (82.9)	13537 (10.9)	11141 (9.0)	25.0	30.4

[a] Tests performed when the patient was diseased.

Table F.22: Monitoring simulation–results using retest monitoring strategy (strategy B) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests[a] (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[b] n (%)	PPV (%)	Sensitivity (%)
0	30364/20000	1805 (9.0)	3694 (18.5)	677 (3.4)	13824 (69.1)	5499 (27.5)	2482 (12.4)	32.8	72.7
6	22078/14501	196 (1.4)	1108 (7.6)	656 (4.5)	12541 (86.5)	1304 (9.0)	852 (5.9)	15.0	23.0
12	20017/13197	141 (1.1)	827 (6.3)	682 (5.2)	11547 (87.5)	968 (7.3)	823 (6.2)	14.6	17.1
18	18707/12229	139 (1.1)	691 (5.7)	709 (5.8)	10690 (87.4)	830 (6.8)	848 (6.9)	16.7	16.4
24	17432/11399	146 (1.3)	585 (5.1)	682 (6.0)	9986 (87.6)	731 (6.4)	828 (7.3)	20.0	17.6
30	16494/10668	132 (1.2)	564 (5.3)	706 (6.6)	9266 (86.9)	696 (6.5)	838 (7.9)	19.0	15.8
36	15442/9972	153 (1.5)	562 (5.6)	698 (7.0)	8559 (85.8)	715 (7.2)	851 (8.5)	21.4	18.0
42	14378/9257	145 (1.6)	537 (5.8)	686 (7.4)	7889 (85.2)	682 (7.4)	831 (9.0)	21.3	17.4
48	13460/8575	161 (1.9)	466 (5.4)	683 (8.0)	7265 (84.7)	627 (7.3)	844 (9.8)	25.7	19.1
54	12558/7948	141 (1.8)	462 (5.8)	703 (8.8)	6642 (83.6)	603 (7.6)	844 (10.6)	23.4	16.7
60	11660/7345	100 (1.4)	293 (4.0)	766 (10.4)	6186 (84.2)	393 (5.4)	866 (11.8)	25.4	11.5
All	192590/125091	3259 (2.6)	9789 (7.8)	7648 (6.1)	104395 (83.5)	13048 (10.4)	10907 (8.7)	25.0	29.9

[a] Tests performed/number of people tests performed on (number of results generated).

[b] Tests performed when the patient was diseased.

Table F.23: Monitoring simulation–results using reduced frequency of monitoring strategy (strategy C) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1818 (9.1)	4227 (21.1)	664 (3.3)	13291 (66.5)	6045 (30.2)	2482 (12.4)	30.1	73.2
12	13955	383 (2.7)	1789 (12.8)	650 (4.7)	11133 (79.8)	2172 (15.6)	1033 (7.4)	17.6	37.1
24	11783	314 (2.7)	1327 (11.3)	634 (5.4)	9508 (80.7)	1641 (13.9)	948 (8.0)	19.1	33.1
36	10142	292 (2.9)	1133 (11.2)	654 (6.4)	8063 (79.5)	1425 (14.1)	946 (9.3)	20.5	30.9
48	8717	318 (3.6)	995 (11.4)	641 (7.4)	6763 (77.6)	1313 (15.1)	959 (11.0)	24.2	33.2
60	7404	331 (4.5)	887 (12.0)	651 (8.8)	5535 (74.8)	1218 (16.5)	982 (13.3)	27.2	33.7
All	72001	3456 (4.8)	10358 (14.4)	3894 (5.4)	54293 (75.4)	13814 (19.2)	7350 (10.2)	25.0	47.0

[a] Tests performed when the patient was diseased.

Table F.24: Monitoring simulation–results using absolute increase from start value monitoring strategy (strategy D) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	90 (0.6)	867 (5.8)	890 (5.9)	13117 (87.7)	957 (6.4)	980 (6.5)	9.4	9.2
12	14007	132 (0.9)	816 (5.8)	968 (6.9)	12091 (86.3)	948 (6.8)	1100 (7.9)	13.9	12.0
18	13059	172 (1.3)	852 (6.5)	1008 (7.7)	11027 (84.4)	1024 (7.8)	1180 (9.0)	16.8	14.6
24	12035	185 (1.5)	754 (6.3)	991 (8.2)	10105 (84.0)	939 (7.8)	1176 (9.8)	19.7	15.7
30	11096	222 (2.0)	780 (7.0)	966 (8.7)	9128 (82.3)	1002 (9.0)	1188 (10.7)	22.2	18.7
36	10094	233 (2.3)	785 (7.8)	882 (8.7)	8194 (81.2)	1018 (10.1)	1115 (11.0)	22.9	20.9
42	9076	231 (2.5)	794 (8.7)	789 (8.7)	7262 (80.0)	1025 (11.3)	1020 (11.2)	22.5	22.6
48	8051	249 (3.1)	732 (9.1)	688 (8.5)	6382 (79.3)	981 (12.2)	937 (11.6)	25.4	26.6
54	7070	224 (3.2)	599 (8.5)	594 (8.4)	5653 (80.0)	823 (11.6)	818 (11.6)	27.2	27.4
60	6247	190 (3.0)	566 (9.1)	527 (8.4)	4964 (79.5)	756 (12.1)	717 (11.5)	25.1	26.5
All	125699	3632 (2.9)	10877 (8.7)	9081 (7.2)	102109 (81.2)	14509 (11.5)	12713 (10.1)	25.0	28.6

[a] Tests performed when the patient was diseased.

Table F.25: Monitoring simulation–results using absolute increase from last value monitoring strategy (strategy E) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	63 (0.4)	629 (4.2)	917 (6.1)	13355 (89.2)	692 (4.6)	980 (6.5)	9.1	6.4
12	14272	61 (0.4)	600 (4.2)	1078 (7.6)	12533 (87.8)	661 (4.6)	1139 (8.0)	9.2	5.4
18	13611	69 (0.5)	534 (3.9)	1231 (9.0)	11777 (86.5)	603 (4.4)	1300 (9.6)	11.4	5.3
24	13008	76 (0.6)	451 (3.5)	1360 (10.5)	11121 (85.5)	527 (4.1)	1436 (11.0)	14.4	5.3
30	12481	89 (0.7)	445 (3.6)	1520 (12.2)	10427 (83.5)	534 (4.3)	1609 (12.9)	16.7	5.5
36	11947	83 (0.7)	425 (3.6)	1687 (14.1)	9752 (81.6)	508 (4.3)	1770 (14.8)	16.3	4.7
42	11439	101 (0.9)	431 (3.8)	1852 (16.2)	9055 (79.2)	532 (4.7)	1953 (17.1)	19.0	5.2
48	10907	135 (1.2)	428 (3.9)	2034 (18.6)	8310 (76.2)	563 (5.2)	2169 (19.9)	24.0	6.2
54	10344	123 (1.2)	345 (3.3)	2256 (21.8)	7620 (73.7)	468 (4.5)	2379 (23.0)	26.3	5.2
60	9876	147 (1.5)	316 (3.2)	2423 (24.5)	6990 (70.8)	463 (4.7)	2570 (26.0)	31.7	5.7
All	142849	2651 (1.9)	7936 (5.6)	17136 (12.0)	115126 (80.6)	10587 (7.4)	19787 (13.9)	25.0	13.4

[a] Tests performed when the patient was diseased.

Table F.26: Monitoring simulation–results using relative increase from start value monitoring strategy (strategy F) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	87 (0.6)	1020 (6.8)	893 (6.0)	12964 (86.6)	1107 (7.4)	980 (6.5)	7.9	8.9
12	13857	133 (1.0)	877 (6.3)	975 (7.0)	11872 (85.7)	1010 (7.3)	1108 (8.0)	13.2	12.0
18	12847	159 (1.2)	862 (6.7)	1028 (8.0)	10798 (84.1)	1021 (7.9)	1187 (9.2)	15.6	13.4
24	11826	182 (1.5)	784 (6.6)	1019 (8.6)	9841 (83.2)	966 (8.2)	1201 (10.2)	18.8	15.2
30	10860	227 (2.1)	737 (6.8)	994 (9.2)	8902 (82.0)	964 (8.9)	1221 (11.2)	23.5	18.6
36	9896	243 (2.5)	759 (7.7)	911 (9.2)	7983 (80.7)	1002 (10.1)	1154 (11.7)	24.3	21.1
42	8894	215 (2.4)	731 (8.2)	843 (9.5)	7105 (79.9)	946 (10.6)	1058 (11.9)	22.7	20.3
48	7948	255 (3.2)	701 (8.8)	737 (9.3)	6255 (78.7)	956 (12.0)	992 (12.5)	26.7	25.7
54	6992	241 (3.4)	598 (8.6)	637 (9.1)	5516 (78.9)	839 (12.0)	878 (12.6)	28.7	27.4
60	6153	196 (3.2)	522 (8.5)	558 (9.1)	4877 (79.3)	718 (11.7)	754 (12.3)	27.3	26.0
All	124237	3642 (2.9)	10923 (8.8)	9373 (7.5)	100299 (80.7)	14565 (11.7)	13015 (10.5)	25.0	28.0

[a] Tests performed when the patient was diseased.

Table F.27: Monitoring simulation–results using relative increase from last value monitoring strategy (strategy G) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	33 (0.2)	455 (3.0)	947 (6.3)	13529 (90.4)	488 (3.3)	980 (6.5)	6.8	3.4
12	14476	41 (0.3)	429 (3.0)	1135 (7.8)	12871 (88.9)	470 (3.2)	1176 (8.1)	8.7	3.5
18	14006	40 (0.3)	379 (2.7)	1321 (9.4)	12266 (87.6)	419 (3.0)	1361 (9.7)	9.5	2.9
24	13587	44 (0.3)	327 (2.4)	1508 (11.1)	11708 (86.2)	371 (2.7)	1552 (11.4)	11.9	2.8
30	13216	45 (0.3)	311 (2.4)	1741 (13.2)	11119 (84.1)	356 (2.7)	1786 (13.5)	12.6	2.5
36	12860	44 (0.3)	290 (2.3)	1978 (15.4)	10548 (82.0)	334 (2.6)	2022 (15.7)	13.2	2.2
42	12526	47 (0.4)	282 (2.3)	2254 (18.0)	9943 (79.4)	329 (2.6)	2301 (18.4)	14.3	2.0
48	12197	63 (0.5)	283 (2.3)	2565 (21.0)	9286 (76.1)	346 (2.8)	2628 (21.5)	18.2	2.4
54	11851	56 (0.5)	212 (1.8)	2923 (24.7)	8660 (73.1)	268 (2.3)	2979 (25.1)	20.9	1.9
60	11583	52 (0.4)	204 (1.8)	3267 (28.2)	8060 (69.6)	256 (2.2)	3319 (28.7)	20.3	1.6
All	151266	2169 (1.4)	6504 (4.3)	20417 (13.5)	122176 (80.8)	8673 (5.7)	22586 (14.9)	25.0	9.6

[a] Tests performed when the patient was diseased.

Table F.28: Monitoring simulation–results using linear regression monitoring strategy (strategy H) by observation point for adjusted fibrosis progression estimate data.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1704 (8.5)	3332 (16.7)	778 (3.9)	14186 (70.9)	5036 (25.2)	2482 (12.4)	33.8	68.7
6	14964	367 (2.5)	2070 (13.8)	613 (4.1)	11914 (79.6)	2437 (16.3)	980 (6.5)	15.1	37.4
12	12527	156 (1.2)	1032 (8.2)	607 (4.8)	10732 (85.7)	1188 (9.5)	763 (6.1)	13.1	20.4
18	11339	135 (1.2)	703 (6.2)	619 (5.5)	9882 (87.2)	838 (7.4)	754 (6.6)	16.1	17.9
24	10501	126 (1.2)	513 (4.9)	597 (5.7)	9265 (88.2)	639 (6.1)	723 (6.9)	19.7	17.4
30	9862	99 (1.0)	397 (4.0)	649 (6.6)	8717 (88.4)	496 (5.0)	748 (7.6)	20.0	13.2
36	9366	97 (1.0)	413 (4.4)	684 (7.3)	8172 (87.3)	510 (5.4)	781 (8.3)	19.0	12.4
42	8856	125 (1.4)	362 (4.1)	702 (7.9)	7667 (86.6)	487 (5.5)	827 (9.3)	25.7	15.1
48	8369	130 (1.6)	297 (3.5)	737 (8.8)	7205 (86.1)	427 (5.1)	867 (10.4)	30.4	15.0
54	7942	151 (1.9)	303 (3.8)	764 (9.6)	6724 (84.7)	454 (5.7)	915 (11.5)	33.3	16.5
60	7488	147 (2.0)	271 (3.6)	801 (10.7)	6269 (83.7)	418 (5.6)	948 (12.7)	35.2	15.5
All	121214	3237 (2.7)	9693 (8.0)	7551 (6.2)	100733 (83.1)	12930 (10.7)	10788 (8.9)	25.0	30.0

[a] Tests performed when the patient was diseased.

Table F.29: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results of using the reference strategy when changing estimates required for data simulation *(Difference to reference strategy for original adjusted fibrosis progression estimate simulation data).*

Change in data simulation	**Threshold**	**PPV (%)**	**Number of tests per person**[a]	**Delay**[b] **(%)**	**Develop cirrhosis**[c] **n (%)**
None	10.460	25.0	6.21	5.63	7689 (38.45)
Decreased†measurement error	10.460	29.7	7.04	7.18	7664 (38.32)
		(+4.7)	*(+0.83)*	*(+1.55)*	*(-25 (0.13))*
	10.205	25.0	6.22	5.16	
			(+0.01)	*(-0.47)*	
Increased‡measurement error	10.460	19.9	4.64	3.86	7808 (39.04)
		(-5.1)	*(-1.57)*	*(-1.77)*	*(+119 (0.60))*
	10.970	25.0	6.01	6.60	
			(-0.21)	*(+0.97)*	
Decrease†between-individual variability	10.460	28.5	7.10	2.97	7531 (37.66)
		(+3.5)	*(+0.89)*	*(-2.66)*	*(-158 (0.79))*
	10.350	25.0	6.66	2.23	
			(+0.45)	*(-3.40)*	
Increased‡between-individual variability	10.460	19.7	4.78	5.63	7659 (37.85)
		(-5.3)	*(-1.43)*	*(+0.00)*	*(-120 (0.60))*
	11.320	25.0	6.68	10.08	
			(+0.47)	*(+4.45)*	
Decreased†fibrosis progression rate	10.460	20.7	6.85	4.72	5314 (26.57)
		(-4.3)	*(+0.64)*	*(-0.91)*	*(-2375 (11.88))*
	10.725	25.0	7.64	6.10	
			(+1.43)	*(+0.47)*	
Increased‡fibrosis progression rate	10.460	36.3	5.13	9.29	13967 (69.84)
		(+11.3)	*(-1.08)*	*(+3.66)*	*(+6278 (31.39))*
	9.765	25.0	3.14	2.81	
			(-3.07)	*(-2.82)*	

[a] Number of tests per person over the duration of monitoring.
[b] % of all patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).
[c] Patients that would go on to develop cirrhosis in the monitoring duration if no intervention were received.
† Decrease is halving the estimate used in the original simulation.
‡ Increase is doubling the estimate used in the original simulation.

Table F.30: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased measurement error by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1705 (8.5)	2899 (14.5)	752 (3.8)	14644 (73.2)	4604 (23.0)	2457 (12.3)	37.0	69.4
6	15396	152 (1.0)	773 (5.0)	802 (5.2)	13669 (88.8)	925 (6.0)	954 (6.2)	16.4	15.9
12	14471	157 (1.1)	563 (3.9)	845 (5.8)	12906 (89.2)	720 (5.0)	1002 (6.9)	21.8	15.7
18	13751	150 (1.1)	543 (3.9)	899 (6.5)	12159 (88.4)	693 (5.0)	1049 (7.6)	21.6	14.3
24	13058	161 (1.2)	521 (4.0)	895 (6.9)	11481 (87.9)	682 (5.2)	1056 (8.1)	23.6	15.2
30	12376	168 (1.4)	490 (4.0)	928 (7.5)	10790 (87.2)	658 (5.3)	1096 (8.9)	25.5	15.3
36	11718	165 (1.4)	519 (4.4)	941 (8.0)	10093 (86.1)	684 (5.8)	1106 (9.4)	24.1	14.9
42	11034	166 (1.5)	502 (4.5)	959 (8.7)	9407 (85.3)	668 (6.1)	1125 (10.2)	24.9	14.8
48	10366	218 (2.1)	490 (4.7)	958 (9.2)	8700 (83.9)	708 (6.8)	1176 (11.3)	30.8	18.5
54	9658	191 (2.0)	462 (4.8)	978 (10.1)	8027 (83.1)	653 (6.8)	1169 (12.1)	29.2	16.3
60	9005	230 (2.6)	427 (4.7)	963 (10.7)	7385 (82.0)	657 (7.3)	1193 (13.2)	35.0	19.3
All	140833	3463 (2.5)	8189 (5.8)	9920 (7.0)	119261 (84.7)	11652 (8.3)	13383 (9.5)	29.7	25.9

[a] Tests performed when the patient was diseased.

Table F.31: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased measurement error by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1869 (9.3)	4113 (20.6)	588 (2.9)	13430 (67.2)	5982 (29.9)	2457 (12.3)	31.2	76.1
6	14018	142 (1.0)	926 (6.6)	607 (4.3)	12343 (88.1)	1068 (7.6)	749 (5.3)	13.3	19.0
12	12950	135 (1.0)	675 (5.2)	632 (4.9)	11508 (88.9)	810 (6.3)	767 (5.9)	16.7	17.6
18	12140	133 (1.1)	661 (5.4)	657 (5.4)	10689 (88.0)	794 (6.5)	790 (6.5)	16.8	16.8
24	11346	119 (1.0)	555 (4.9)	655 (5.8)	10017 (88.3)	674 (5.9)	774 (6.8)	17.7	15.4
30	10672	139 (1.3)	598 (5.6)	664 (6.2)	9271 (86.9)	737 (6.9)	803 (7.5)	18.9	17.3
36	9935	138 (1.4)	525 (5.3)	660 (6.6)	8612 (86.7)	663 (6.7)	798 (8.0)	20.8	17.3
42	9272	121 (1.3)	537 (5.8)	673 (7.3)	7941 (85.6)	658 (7.1)	794 (8.6)	18.4	15.2
48	8614	170 (2.0)	451 (5.2)	647 (7.5)	7346 (85.3)	621 (7.2)	817 (9.5)	27.4	20.8
54	7993	165 (2.1)	464 (5.8)	653 (8.2)	6711 (84.0)	629 (7.9)	818 (10.2)	26.2	20.2
60	7364	190 (2.6)	439 (6.0)	624 (8.5)	6111 (83.0)	629 (8.5)	814 (11.1)	30.2	23.3
All	124304	3321 (2.7)	9944 (8.0)	7060 (5.7)	103979 (83.6)	13265 (10.7)	10381 (8.4)	25.0	32.0

[a] Tests performed when the patient was diseased.

Table F.32: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased measurement error by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1790 (8.9)	4412 (22.1)	773 (3.9)	13025 (65.1)	6202 (31.0)	2563 (12.8)	28.9	69.8
6	13798	385 (2.8)	2254 (16.3)	608 (4.4)	10551 (76.5)	2639 (19.1)	993 (7.2)	14.6	38.8
12	11159	243 (2.2)	1572 (14.1)	532 (4.8)	8812 (79.0)	1815 (16.3)	775 (6.9)	13.4	31.4
18	9344	169 (1.8)	1255 (13.4)	484 (5.2)	7436 (79.6)	1424 (15.2)	653 (7.0)	11.9	25.9
24	7920	123 (1.6)	892 (11.3)	452 (5.7)	6453 (81.5)	1015 (12.8)	575 (7.3)	12.1	21.4
30	6905	130 (1.9)	738 (10.7)	432 (6.3)	5605 (81.2)	868 (12.6)	562 (8.1)	15.0	23.1
36	6037	105 (1.7)	636 (10.5)	406 (6.7)	4890 (81.0)	741 (12.3)	511 (8.5)	14.2	20.5
42	5296	95 (1.8)	528 (10.0)	380 (7.2)	4293 (81.1)	623 (11.8)	475 (9.0)	15.2	20.0
48	4673	99 (2.1)	464 (9.9)	372 (8.0)	3738 (80.0)	563 (12.0)	471 (10.1)	17.6	21.0
54	4110	89 (2.2)	378 (9.2)	374 (9.1)	3269 (79.5)	467 (11.4)	463 (11.3)	19.1	19.2
60	3643	101 (2.8)	312 (8.6)	363 (10.0)	2867 (78.7)	413 (11.3)	464 (12.7)	24.5	21.8
All	92885	3329 (3.6)	13441 (14.5)	5176 (5.6)	70939 (76.4)	16770 (18.1)	8505 (9.2)	19.9	39.1

[a] Tests performed when the patient was diseased.

Table F.33: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased measurement error by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1468 (7.3)	2728 (13.6)	1095 (5.5)	14709 (73.5)	4196 (21.0)	2563 (12.8)	35.0	57.3
6	15804	464 (2.9)	1604 (10.1)	916 (5.8)	12820 (81.1)	2068 (13.1)	1380 (8.7)	22.4	33.6
12	13736	278 (2.0)	1252 (9.1)	871 (6.3)	11335 (82.5)	1530 (11.1)	1149 (8.4)	18.2	24.2
18	12206	222 (1.8)	1015 (8.3)	826 (6.8)	10143 (83.1)	1237 (10.1)	1048 (8.6)	17.9	21.2
24	10969	206 (1.9)	811 (7.4)	768 (7.0)	9184 (83.7)	1017 (9.3)	974 (8.9)	20.3	21.1
30	9952	177 (1.8)	750 (7.5)	760 (7.6)	8265 (83.0)	927 (9.3)	937 (9.4)	19.1	18.9
36	9025	168 (1.9)	651 (7.2)	735 (8.1)	7471 (82.8)	819 (9.1)	903 (10.0)	20.5	18.6
42	8206	145 (1.8)	582 (7.1)	723 (8.8)	6756 (82.3)	727 (8.9)	868 (10.6)	19.9	16.7
48	7479	173 (2.3)	543 (7.3)	709 (9.5)	6054 (80.9)	716 (9.6)	882 (11.8)	24.2	19.6
54	6763	159 (2.4)	460 (6.8)	721 (10.7)	5423 (80.2)	619 (9.2)	880 (13.0)	25.7	18.1
60	6144	139 (2.3)	388 (6.3)	731 (11.9)	4886 (79.5)	527 (8.6)	870 (14.2)	26.4	16.0
All	120284	3599 (3.0)	10784 (9.0)	8855 (7.4)	97046 (80.7)	14383 (12.0)	12454 (10.4)	25.0	28.9

[a] Tests performed when the patient was diseased.

Table F.34: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased between-individual variability by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1809 (9.0)	1745 (8.7)	546 (2.7)	15900 (79.5)	3554 (17.8)	2355 (11.8)	50.9	76.8
6	16446	349 (2.1)	925 (5.6)	397 (2.4)	14775 (89.8)	1274 (7.7)	746 (4.5)	27.4	46.8
12	15172	189 (1.2)	771 (5.1)	368 (2.4)	13844 (91.2)	960 (6.3)	557 (3.7)	19.7	33.9
18	14212	143 (1.0)	731 (5.1)	367 (2.6)	12971 (91.3)	874 (6.1)	510 (3.6)	16.4	28.0
24	13338	139 (1.0)	668 (5.0)	335 (2.5)	12196 (91.4)	807 (6.1)	474 (3.6)	17.2	29.3
30	12531	141 (1.1)	660 (5.3)	334 (2.7)	11396 (90.9)	801 (6.4)	475 (3.8)	17.6	29.7
36	11730	140 (1.2)	695 (5.9)	334 (2.8)	10561 (90.0)	835 (7.1)	474 (4.0)	16.8	29.5
42	10895	153 (1.4)	665 (6.1)	304 (2.8)	9773 (89.7)	818 (7.5)	457 (4.2)	18.7	33.5
48	10077	132 (1.3)	698 (6.9)	332 (3.3)	8915 (88.5)	830 (8.2)	464 (4.6)	15.9	28.4
54	9247	153 (1.7)	649 (7.0)	326 (3.5)	8119 (87.8)	802 (8.7)	479 (5.2)	19.1	31.9
60	8445	143 (1.7)	560 (6.6)	341 (4.0)	7401 (87.6)	703 (8.3)	484 (5.7)	20.3	29.5
All	142093	3491 (2.5)	8767 (6.2)	3984 (2.8)	125851 (88.6)	12258 (8.6)	7475 (5.3)	28.5	46.7

[a] Tests performed when the patient was diseased.

Table F.35: Monitoring simulation adjusted fibrosis progression sensitivity analysesresults using reference strategy with decreased between-individual variability by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1917 (9.6)	2120 (10.6)	438 (2.2)	15525 (77.6)	4037 (20.2)	2355 (11.8)	47.5	81.4
6	15963	297 (1.9)	1119 (7.0)	321 (2.0)	14226 (89.1)	1416 (8.9)	618 (3.9)	21.0	48.1
12	14547	161 (1.1)	951 (6.5)	290 (2.0)	13145 (90.4)	1112 (7.6)	451 (3.1)	14.5	35.7
18	13435	127 (0.9)	849 (6.3)	287 (2.1)	12172 (90.6)	976 (7.3)	414 (3.1)	13.0	30.7
24	12459	121 (1.0)	757 (6.1)	247 (2.0)	11334 (91.0)	878 (7.0)	368 (3.0)	13.8	32.9
30	11581	122 (1.1)	722 (6.2)	235 (2.0)	10502 (90.7)	844 (7.3)	357 (3.1)	14.5	34.2
36	10737	119 (1.1)	757 (7.1)	222 (2.1)	9639 (89.8)	876 (8.2)	341 (3.2)	13.6	34.9
42	9861	109 (1.1)	721 (7.3)	211 (2.1)	8820 (89.4)	830 (8.4)	320 (3.2)	13.1	34.1
48	9031	95 (1.1)	696 (7.7)	239 (2.6)	8001 (88.6)	791 (8.8)	334 (3.7)	12.0	28.4
54	8240	126 (1.5)	672 (8.2)	222 (2.7)	7220 (87.6)	798 (9.7)	348 (4.2)	15.8	36.2
60	7442	129 (1.7)	605 (8.1)	221 (3.0)	6487 (87.2)	734 (9.9)	350 (4.7)	17.6	36.9
All	133296	3323 (2.5)	9969 (7.5)	2933 (2.2)	117071 (87.8)	13292 (10.0)	6256 (4.7)	25.0	53.1

[a] Tests performed when the patient was diseased.

Table F.36: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased between-individual variability by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1779 (8.9)	6916 (34.6)	679 (3.4)	10626 (53.1)	8695 (43.5)	2458 (12.3)	20.5	72.4
6	11305	171 (1.5)	1299 (11.5)	683 (6.0)	9152 (81.0)	1470 (13.0)	854 (7.6)	11.6	20.0
12	9835	109 (1.1)	808 (8.2)	741 (7.5)	8177 (83.1)	917 (9.3)	850 (8.6)	11.9	12.8
18	8918	119 (1.3)	631 (7.1)	746 (8.4)	7422 (83.2)	750 (8.4)	865 (9.7)	15.9	13.8
24	8168	99 (1.2)	544 (6.7)	757 (9.3)	6768 (82.9)	643 (7.9)	856 (10.5)	15.4	11.6
30	7525	111 (1.5)	446 (5.9)	765 (10.2)	6203 (82.4)	557 (7.4)	876 (11.6)	19.9	12.7
36	6968	124 (1.8)	390 (5.6)	752 (10.8)	5702 (81.8)	514 (7.4)	876 (12.6)	24.1	14.2
42	6454	129 (2.0)	412 (6.4)	747 (11.6)	5166 (80.0)	541 (8.4)	876 (13.6)	23.8	14.7
48	5913	143 (2.4)	346 (5.9)	718 (12.1)	4706 (79.6)	489 (8.3)	861 (14.6)	29.2	16.6
54	5424	122 (2.2)	305 (5.6)	727 (13.4)	4270 (78.7)	427 (7.9)	849 (15.7)	28.6	14.4
60	4997	130 (2.6)	286 (5.7)	717 (14.3)	3864 (77.3)	416 (8.3)	847 (17.0)	31.2	15.3
All	95507	3036 (3.2)	12383 (13.0)	8032 (8.4)	72056 (75.4)	15419 (16.1)	11068 (11.6)	19.7	27.4

[a] Tests performed when the patient was diseased.

Table F.37: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased between-individual variability by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1406 (7.0)	3876 (19.4)	1052 (5.3)	13666 (68.3)	5282 (26.4)	2458 (12.3)	26.6	57.2
6	14718	190 (1.3)	1002 (6.8)	1109 (7.5)	12417 (84.4)	1192 (8.1)	1299 (8.8)	15.9	14.6
12	13526	141 (1.0)	667 (4.9)	1215 (9.0)	11503 (85.0)	808 (6.0)	1356 (10.0)	17.5	10.4
18	12718	125 (1.0)	523 (4.1)	1272 (10.0)	10798 (84.9)	648 (5.1)	1397 (11.0)	19.3	8.9
24	12070	117 (1.0)	461 (3.8)	1349 (11.2)	10143 (84.0)	578 (4.8)	1466 (12.1)	20.2	8.0
30	11492	132 (1.1)	423 (3.7)	1418 (12.3)	9519 (82.8)	555 (4.8)	1550 (13.5)	23.8	8.5
36	10937	159 (1.5)	441 (4.0)	1437 (13.1)	8900 (81.4)	600 (5.5)	1596 (14.6)	26.5	10.0
42	10337	130 (1.3)	392 (3.8)	1509 (14.6)	8306 (80.4)	522 (5.0)	1639 (15.9)	24.9	7.9
48	9815	161 (1.6)	363 (3.7)	1587 (16.2)	7704 (78.5)	524 (5.3)	1748 (17.8)	30.7	9.2
54	9291	196 (2.1)	342 (3.7)	1627 (17.5)	7126 (76.7)	538 (5.8)	1823 (19.6)	36.4	10.8
60	8753	183 (2.1)	317 (3.6)	1651 (18.9)	6602 (75.4)	500 (5.7)	1834 (21.0)	36.6	10.0
All	133657	2940 (2.2)	8807 (6.6)	15226 (11.4)	106684 (79.8)	11747 (8.8)	18166 (13.6)	25.0	16.2

[a] Tests performed when the patient was diseased.

Table F.38: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased fibrosis progression rate by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1290 (6.4)	3094 (15.5)	643 (3.2)	14973 (74.9)	4384 (21.9)	1933 (9.7)	29.4	66.7
6	15616	174 (1.1)	1204 (7.7)	606 (3.9)	13632 (87.3)	1378 (8.8)	780 (5.0)	12.6	22.3
12	14238	128 (0.9)	833 (5.9)	601 (4.2)	12676 (89.0)	961 (6.7)	729 (5.1)	13.3	17.6
18	13277	116 (0.9)	693 (5.2)	602 (4.5)	11866 (89.4)	809 (6.1)	718 (5.4)	14.3	16.2
24	12468	99 (0.8)	608 (4.9)	597 (4.8)	11164 (89.5)	707 (5.7)	696 (5.6)	14.0	14.2
30	11761	108 (0.9)	532 (4.5)	594 (5.1)	10527 (89.5)	640 (5.4)	702 (6.0)	16.9	15.4
36	11121	98 (0.9)	540 (4.9)	596 (5.4)	9887 (88.9)	638 (5.7)	694 (6.2)	15.4	14.1
42	10483	110 (1.0)	463 (4.4)	593 (5.7)	9317 (88.9)	573 (5.5)	703 (6.7)	19.2	15.6
48	9910	106 (1.1)	460 (4.6)	572 (5.8)	8772 (88.5)	566 (5.7)	678 (6.8)	18.7	15.6
54	9344	90 (1.0)	415 (4.4)	583 (6.2)	8256 (88.4)	505 (5.4)	673 (7.2)	17.8	13.4
60	8839	90 (1.0)	401 (4.5)	587 (6.6)	7761 (87.8)	491 (5.6)	677 (7.7)	18.3	13.3
All	137057	2409 (1.8)	9243 (6.7)	6574 (4.8)	118831 (86.7)	11652 (8.5)	8983 (6.6)	20.7	26.8

[a] Tests performed when the patient was diseased.

Table F.39: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	1162 (5.8)	2236 (11.2)	771 (3.9)	15831 (79.2)	3398 (17.0)	1933 (9.7)	34.2	60.1
6	16602	207 (1.2)	920 (5.5)	733 (4.4)	14742 (88.8)	1127 (6.8)	940 (5.7)	18.4	22.0
12	15475	147 (0.9)	661 (4.3)	740 (4.8)	13927 (90.0)	808 (5.2)	887 (5.7)	18.2	16.6
18	14667	102 (0.7)	558 (3.8)	783 (5.3)	13224 (90.2)	660 (4.5)	885 (6.0)	15.5	11.5
24	14007	113 (0.8)	495 (3.5)	786 (5.6)	12613 (90.0)	608 (4.3)	899 (6.4)	18.6	12.6
30	13399	104 (0.8)	446 (3.3)	813 (6.1)	12036 (89.8)	550 (4.1)	917 (6.8)	18.9	11.3
36	12849	127 (1.0)	458 (3.6)	805 (6.3)	11459 (89.2)	585 (4.6)	932 (7.3)	21.7	13.6
42	12264	121 (1.0)	429 (3.5)	828 (6.8)	10886 (88.8)	550 (4.5)	949 (7.7)	22.0	12.8
48	11714	129 (1.1)	449 (3.8)	817 (7.0)	10319 (88.1)	578 (4.9)	946 (8.1)	22.3	13.6
54	11136	135 (1.2)	390 (3.5)	806 (7.2)	9805 (88.0)	525 (4.7)	941 (8.5)	25.7	14.3
60	10611	125 (1.2)	369 (3.5)	814 (7.7)	9303 (87.7)	494 (4.7)	939 (8.8)	25.3	13.3
All	152724	2472 (1.6)	7411 (4.9)	8696 (5.7)	134145 (87.8)	9883 (6.5)	11168 (7.3)	25.0	22.1

[a] Tests performed when the patient was diseased.

Table F.40: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased fibrosis progression rate by observation point.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	2473 (12.4)	3320 (16.6)	1123 (5.6)	13084 (65.4)	5793 (29.0)	3596 (18.0)	42.7	68.8
6	14207	447 (3.1)	1394 (9.8)	1097 (7.7)	11269 (79.3)	1841 (13.0)	1544 (10.9)	24.3	29.0
12	12366	344 (2.8)	1097 (8.9)	1096 (8.9)	9829 (79.5)	1441 (11.7)	1440 (11.6)	23.9	23.9
18	10925	311 (2.8)	933 (8.5)	1088 (10.0)	8593 (78.7)	1244 (11.4)	1399 (12.8)	25.0	22.2
24	9681	335 (3.5)	859 (8.9)	1035 (10.7)	7452 (77.0)	1194 (12.3)	1370 (14.2)	28.1	24.5
30	8487	347 (4.1)	766 (9.0)	990 (11.7)	6384 (75.2)	1113 (13.1)	1337 (15.8)	31.2	26.0
36	7374	350 (4.7)	746 (10.1)	1019 (13.8)	5259 (71.3)	1096 (14.9)	1369 (18.6)	31.9	25.6
42	6278	386 (6.1)	579 (9.2)	1036 (16.5)	4277 (68.1)	965 (15.4)	1422 (22.7)	40.0	27.1
48	5313	395 (7.4)	509 (9.6)	1016 (19.1)	3393 (63.9)	904 (17.0)	1411 (26.6)	43.7	28.0
54	4409	421 (9.5)	389 (8.8)	949 (21.5)	2650 (60.1)	810 (18.4)	1370 (31.1)	52.0	30.7
60	3599	391 (10.9)	277 (7.7)	843 (23.4)	2088 (58.0)	668 (18.6)	1234 (34.3)	58.5	31.7
All	102639	6200 (6.0)	10869 (10.6)	11292 (11.0)	74278 (72.4)	17069 (16.6)	17492 (17.0)	36.3	35.4

[a] Tests performed when the patient was diseased.

Table F.41: Monitoring simulation adjusted fibrosis progression sensitivity analyses–results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%.

Observation month	Tests (n)	TP results n (%)	FP results n (%)	FN results n (%)	TN results n (%)	Positive results n (%)	Diseased[a] n (%)	PPV (%)	Sensitivity (%)
0	20000	3037 (15.2)	6915 (34.6)	559 (2.8)	9489 (47.4)	9952 (49.8)	3596 (18.0)	30.5	84.5
6	10048	328 (3.3)	2124 (21.1)	454 (4.5)	7142 (71.1)	2452 (24.4)	782 (7.8)	13.4	41.9
12	7596	216 (2.8)	1330 (17.5)	403 (5.3)	5647 (74.3)	1546 (20.4)	619 (8.1)	14.0	34.9
18	6050	207 (3.4)	990 (16.4)	339 (5.6)	4514 (74.6)	1197 (19.8)	546 (9.0)	17.3	37.9
24	4853	153 (3.2)	789 (16.3)	295 (6.1)	3616 (74.5)	942 (19.4)	448 (9.2)	16.2	34.2
30	3911	139 (3.6)	635 (16.2)	277 (7.1)	2860 (73.1)	774 (19.8)	416 (10.6)	18.0	33.4
36	3137	135 (4.3)	493 (15.7)	284 (9.1)	2225 (70.9)	628 (20.0)	419 (13.4)	21.5	32.2
42	2509	158 (6.3)	358 (14.3)	262 (10.4)	1731 (69.0)	516 (20.6)	420 (16.7)	30.6	37.6
48	1993	151 (7.6)	290 (14.6)	247 (12.4)	1305 (65.5)	441 (22.1)	398 (20.0)	34.2	37.9
54	1552	126 (8.1)	219 (14.1)	234 (15.1)	973 (62.7)	345 (22.2)	360 (23.2)	36.5	35.0
60	1207	111 (9.2)	155 (12.8)	213 (17.6)	728 (60.3)	266 (22.0)	324 (26.8)	41.7	34.3
All	62856	4761 (7.6)	14298 (22.7)	3567 (5.7)	40230 (64.0)	19059 (30.3)	8328 (13.2)	25.0	57.2

[a] Tests performed when the patient was diseased.

References

1. Dinnes, J, Hewison, J, Altman, DG, and Deeks, JJ. The basis for monitoring strategies in clinical guidelines: a case study of prostate-specific antigen for monitoring in prostate cancer. Canadian Medical Association Journal 2012;184:169–77.
2. Glasziou, PP and Aronson, JK. An introduction to monitoring therapeutic interventions in clinical practice. In: *Evidence-based Medical Monitoring: From Principles to Practice*. Ed. by Glasziou, PP, Irwig, L, and Aronson, JK. Oxford: Blackwell Publishing, 2008:3–14.
3. Glasziou, P, Irwig, L, and Mant, D. Monitoring in chronic disease: a rational approach. British Medical Journal 2005;330:644–8.
4. Selby, PJ, Banks, RE, Gregory, W, et al. A Multi-Centre Programme into the Evaluation of Biomarkers Suitable for Use in Patients with Kidney and Liver Diseases: Chapter 4: Has the randomised controlled trial design been successfully used to evaluate strategies for monitoring disease progression or recurrence? An assessment of experience to date. Programme Grants for Applied Research 2018.
5. MRC. A framework for development and evaluation of RCTs for complex interventions to improve health. Report. 2000.
6. Oxford University Press. "monitoring". URL: `http://oxforddictionaries.com/definition/monitor`.

7. Selby, PJ, Banks, RE, Gregory, W, et al. A Multi-Centre Programme into the Evaluation of Biomarkers Suitable for Use in Patients with Kidney and Liver Diseases: Chapter 6: How can monitoring impact on patient outcomes? Programme Grants for Applied Research 2018.
8. Mostafid, H, Bryan, RT, and Rees, J. Diagnosis and treatment of non-muscle-invasive bladder cancer. Trends in Urology & Men's Health 2015;6:23–27.
9. Buclin, T, Telenti, A, Perera, R, et al. Development and validation of decision rules to guide frequency of monitoring CD4 cell count in HIV-1 infection before starting antiretroviral therapy. PLoS One 2011;6:e18578–e18578.
10. Keenan, K, Hayen, A, Neal, BC, and Irwig, L. Long term monitoring in patients receiving treatment to lower blood pressure: analysis of data from placebo controlled randomised controlled trial. British Medical Journal 2009;338:b1492–b1492.
11. Oke, JL, Stevens, RJ, Gaitskell, K, and Farmer, AJ. Establishing an evidence base for frequency of monitoring glycated haemoglobin levels in patients with Type 2 diabetes: projections of effectiveness from a regression model. Diabetic Medicine 2012;29:266–71.
12. Stevens, RJ, Oke, J, and Perera, R. Statistical models for the control phase of clinical monitoring. Statistical Methods in Medical Research 2010;19:394–414.
13. Aarsand Aasne, K, Røraas, T, and Sandberg, S. Biological variation–reliable data is essential. Clinical Chemistry and Laboratory Medicine 2015;53:153.
14. Glasziou, PP, Irwig, L, Heritier, S, Simes, RJ, and Tonkin, A. Monitoring cholesterol levels: measurement error or true change? Annals of Internal Medicine 2008;148:656–61.
15. Segen's Medical Dictionary. "biological variability". 2011. URL: `https://medical-dictionary.thefreedictionary.com/biological+variability`.
16. Fraser, CG. Biological Variation: From Principles to Practice. AACC Press, 2001.
17. Bargnoux, AS, Servel, AC, Pieroni, L, et al. Accuracy of GFR predictive equations in renal transplantation: validation of a new turbidimetric cystatin C assay on Architect c8000. Clinical Biochemistry 2012;45:151–3.

18. Perich, C, Minchinela, J, Ricós, C, et al. Biological variation database: structure and criteria used for generation and update. Clinical Chemistry and Laboratory Medicine 2015;53:299–305.

19. Lamb, EJ, Brettell, EA, Cockwell, P, et al. The eGFR-C study: accuracy of glomerular filtration rate (GFR) estimation using creatinine and cystatin C and albuminuria for monitoring disease progression in patients with stage 3 chronic kidney disease - prospective longitudinal study in a multiethnic population. BMC Nephrology 2014;15:1–11.

20. Lippi, G, Mattiuzzi, C, and Favaloro, EJ. Pre-analytical variability and quality of diagnostic testing. Looking at the moon and gazing beyond the finger. New Zealand Journal of Medical Laboratory Science 2015.

21. Castilla, JA, Alvarez, C, Aguilar, J, González-Varea, C, Gonzalvo, MC, and Martínez, L. Influence of analytical and biological variation on the clinical interpretation of seminal parameters. Human Reproduction 2006;21:847–851.

22. Simundic, AM, Kackov, S, Miler, M, Fraser, CG, and Petersen, PH. Terms and symbols used in studies on biological variation: the need for harmonization. Clinical Chemistry 2015;61:438–9.

23. Streiner, D, Norman, G, and Cairney, J. Health Measurement Scales: A practical guide to their development and use. Oxford University Press, 2014.

24. Khullar, V, Salvatore, S, Cardozo, L, Bourne, TH, Abbott, D, and Kelleher, C. A novel technique for measuring bladder wall thickness in women using transvaginal ultrasound. Ultrasound in Obstetrics and Gynecology 1994;4:220–3.

25. Reed, GF, Lynn, F, and Meade, BD. Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. Clinical and Diagnostic Laboratory Immunology 2002;9:1235–1239.

26. Mockus, L, Peterson, JJ, Lainez, JM, and Reklaitis, GV. Batch-to-Batch Variation: A Key Component for Modeling Chemical Manufacturing Processes. Organic Process Research & Development 2015;19:908–914.

27. Barroso, G, Mercan, R, Ozgur, K, et al. Intra- and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation,

staining techniques and manual versus computerized analysis. Human Reproduction 1999;14:2036–40.

28. Sölétormos, G and Schiøler, V. Description of a computer program to assess cancer antigen 15.3, carcinoembryonic antigen, and tissue polypeptide antigen information during monitoring of metastatic breast cancer. Clinical Chemistry 2000;46:1106–13.
29. Bellera, CA, Hanley, JA, Joseph, L, and Albertsen, PC. Detecting trends in noisy data series: application to biomarker series. American Journal of Epidemiology 2008;167:1130–9.
30. Pepe, MS, Zheng, Y, Jin, Y, Huang, Y, Parikh, CR, and Levy, WC. Evaluating the ROC performance of markers for future events. Lifetime Data Analysis 2008;14:86–113.
31. Li, H and Gatsonis, C. Dynamic optimal strategy for monitoring disease recurrence. Science China Mathematics 2012;55:1565–1582.
32. Ferrante di Ruffano, L, Hyde, CJ, McCaffery, KJ, Bossuyt, PM, and Deeks, JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. British Medical Journal 2012;344:e686.
33. Adriaensen, WJ, Matheï, C, Buntinx, FJ, and Arbyn, M. A framework provided an outline toward the proper evaluation of potential screening strategies. Journal of Clinical Epidemiology 2013;66:639–647.
34. Burdick, R and Graybill, F. Confidence Intervals on Variance Components. Taylor & Francis, 1992.
35. Fraser, CG and Harris, EK. Generation and application of data on biological variation in clinical chemistry. Critical Reviews in Clinical Laboratory Sciences 1989;27:409–37.
36. Greenhalgh, T and Peacock, R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. British Medical Journal 2005;331:1064–1065.
37. Young, DS, Harris, EK, and Cotlove, E. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subject. Clinical Chemistry 1971.
38. Sandberg, S, Fraser, CG, Horvath, AR, et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European

Federation of Clinical Chemistry and Laboratory Medicine. Clinical Chemistry and Laboratory Medicine;53:833–5.

39. Fraser, CG. The 1999 Stockholm Consensus Conference on quality specifications in laboratory medicine. Clinical Chemistry and Laboratory Medicine 2015;53:837–40.
40. Kottner, J, Audige, L, Brorson, S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Journal of Clinical Epidemiology 2011;64:96–106.
41. Koski, JM, Saarakkala, S, Helle, M, et al. Assessing the intra- and inter-reader reliability of dynamic ultrasound images in power Doppler ultrasonography. Annals of the Rheumatic Diseases 2006;65:1658–1660.
42. To, T, Estrabillo, E, Wang, C, and Cicutto, L. Examining intra-rater and inter-rater response agreement: A medical chart abstraction study of a community-based asthma care program. BMC Medical Research Methodology 2008;8:29–29.
43. Røraas, T, Petersen, PH, and Sandberg, S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. Clinical Chemistry 2012;58:1306–13.
44. de Vet, H, Terwee, C, Mokkink, L, and Knol, D. Measurement in Medicine: A Practical Guide. Cambridge University Press, 2011.
45. Giraudeau, B and Mary, JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Statistics in Medicine 2001;20:3205–3214.
46. Braga, F and Panteghini, M. Generation of data on within-subject biological variation in laboratory medicine: An update. Critical Reviews in Clinical Laboratory Sciences 2016;53:313–325.
47. Bland, JM and Altman, DG. Statistics Notes: Measurement error proportional to the mean. British Medical Journal 1996;313:106.
48. Shapiro, SS and Wilk, MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591–611.

49. Razali, NM and Yap, BW. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics 2011.

50. McNeish, DM and Stapleton, LM. The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. Educational Psychology Review 2016;28:295–314.

51. Røraas, T, Støve, B, Petersen, PH, and Sandberg, S. Biological Variation: The Effect of Different Distributions on Estimated Within-Person Variation and Reference Change Values. Clinical Chemistry 2016;62:725–736.

52. Bland, JM and Altman, DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999;8:135–60.

53. Bartlett, JW and Frost, C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology 2008;31:466–475.

54. McHugh, ML. Interrater reliability: the kappa statistic. Biochemia Medica 2012;22:276–282.

55. Burdick, R, Borror, C, and Montgomery, D. Design and Analysis of Gauge R and R Studies: Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models. Society for Industrial and Applied Mathematics, 2005.

56. Fraser, CG, Peterson, PH, and Larsen, ML. Setting analytical goals for random analytical error in specific clinical monitoring situations. Clinical Chemistry 1990;36:1625–1628.

57. Smellie, WSA. What is a significant difference between sequential laboratory results? Journal of Clinical Pathology 2008;61:419–25.

58. Alexander, KS, Kazmierczak, SC, Snyder, CK, Oberdorf, JA, and Farrell, DH. Prognostic utility of biochemical markers of cardiovascular risk: impact of biological variability. Clinical Chemistry and Laboratory Medicine 2013;51:1875–82.

59. Cole, TJ. Sympercents: symmetric percentage differences on the 100 log(e) scale simplify the presentation of log transformed data. Statistics in Medicine 2000;19:3109–25.

60. Koopmans, LH, Owen, DB, and Rosenblatt, JI. Confidence intervals for the coefficient of variation for the normal and log normal distributions. Biometrika 1964;51:25–32.

61. Kirkwood, TBL. Geometric Means and Measures of Dispersion. Biometrics 1979;35:908–909.

62. Frankenstein, L, Wu, AH, Hallermayer, K, Wians F. H., J, Giannitsis, E, and Katus, HA. Biological variation and reference change value of high-sensitivity troponin T in healthy individuals during short and intermediate follow-up periods. Clinical Chemistry 2011;57:1068–71.

63. Fokkema, MR, Herrmann, Z, Muskiet, FA, and Moecks, J. Reference change values for brain natriuretic peptides revisited. Clinical Chemistry 2006;52:1602–3.

64. Henderson, AR. Chemistry with confidence: should Clinical Chemistry require confidence intervals for analytical and other data? Clinical Chemistry 1993;39:929–35.

65. Harris, EK. On P values and confidence intervals (why can't we P with more confidence?) Clinical Chemistry 1993;39:927–8.

66. AACC Clincal Chemistry Information for Authors. 2016. URL: `http://www.clinchem.org/site/info_ar/info_authors.xhtml`.

67. Røraas, T, Støve, B, Petersen, PH, and Sandberg, S. Biological variation: Evaluation of methods for constructing confidence intervals for estimates of within-person biological variation for different distributions of the within-person effect. Clinica Chimica Acta 2017;468:166–173.

68. Verrill, S. Confidence Bounds for Normal and Lognormal Distribution Coefficients of Variation. United States Department of Agriculture 2003.

69. Bartlett, WA, Braga, F, Carobene, A, et al. A checklist for critical appraisal of studies of biological variation. Clinical Chemistry and Laboratory Medicine 2015;53:879–85.

70. Biological Variation Working Group, European Federation of Clinical Chemistry and Laboratory Medicine. Definition of a minimun data set to accompany indicies of biological variation. In: *IFCC WorldLab Istanbul*. 2014.

71. LONIC. 2017. URL: `https://loinc.org/`.

72. SNOWMED. 2017. URL: `https://www.snomed.org/snomed-ct`.

73. IFCC. 2017. URL: http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-npu/.

74. Bartlett, W. Biological variation data: the need for appraisal of the evidence base. In: *Advances in clinical chemistry and laboratory medicine.* Ed. by Renz H., TR. Berlin, Boston, Beijing: De Gruyter, 2012.

75. Bartlett, W, Braga, F, Carobene, A, Coskun, A, Prusa, R, and P, FC. Identification of key metadata to enable safe accurate and effective transferability of biological variation data. In: *American Association of Clinical Chemistry Annual Meeting AACC Chicago, Illinois*. 2015.

76. Rifai, N, Annesley, T, Berg, J, et al. An appeal to medical journal editors: The need for a full description of laboratory methods and specimen handling in clinical study reports. Statement by the consortium of laboratory medicine journal editors. Annals of Clinical Biochemistry 2012;49:105–107.

77. Simundic, AM, Bartlett, WA, and Fraser, CG. Biological variation: a still evolving facet of laboratory medicine. Annals of Clinical Biochemistry 2015;52:189–90.

78. Westgard, J. Westgard QC database. 2014. URL: https://www.westgard.com/biodatabase1.htm.

79. Ricós, C, Alvarez, V, Cava, F, et al. Current databases on biological variation: pros, cons and progress. Scandinavian Journal of Clinical and Laboratory Investigation 1999;59:491–500.

80. Ricós, C, Iglesias, N, Garcia-Lario, JV, et al. Within-subject biological variation in disease: collated data and clinical consequences. Annals of Clinical Biochemistry 2007;44:343–52.

81. Moher, D, Liberati, A, Tetzlaff, J, Altman, DG, and The Prisma Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine 2009;6:e1000097.

82. Bailey, D, Bevilacqua, V, Colantonio, DA, et al. Pediatric within-day biological variation and quality specifications for 38 biochemical markers in the CALIPER cohort. Clinical Chemistry 2014;60:518–29.

83. Chen, R, Song, Y, Jiang, L, Hong, X, and Ye, P. The assessment of voluntary pelvic floor muscle contraction by three-dimensional transperineal ultrasonography. Archives of Gynecology & Obstetrics 2011;284:931–6.

84. Oliveira, E, Castro, RA, Takano, CC, et al. Ultrasonographic and Doppler velocimetric evaluation of the levator ani muscle in premenopausal women with and without urinary stress incontinence. European Journal of Obstetrics, Gynecology, & Reproductive Biology 2007;133:213–7.

85. Aakre, KM, Røraas, T, Petersen, PH, et al. Weekly and 90-minute biological variations in cardiac troponin T and cardiac troponin I in hemodialysis patients and healthy controls. Clinical Chemistry 2014;60:838–47.

86. Donati, OF, Chong, D, Nanz, D, et al. Diffusion-weighted MR imaging of upper abdominal organs: field strength and intervendor variability of apparent diffusion coefficients. Radiology 2014;270:454–63.

87. Jimmerson, LC, Zheng, JH, Bushman, LR, MacBrayne, CE, Anderson, PL, and Kiser, JJ. Development and validation of a dried blood spot assay for the quantification of ribavirin using liquid chromatography coupled to mass spectrometry. Journal of Chromatography B Analytical Technologies in the Biomedical and Life Sciences 2014;944:18–24.

88. Martinez-Morillo, E, Diamandis, A, and Diamandis, EP. Reference intervals and biological variation for kallikrein 6: influence of age and renal failure. Clinical Chemistry and Laboratory Medicine 2012;50:931–4.

89. Herpel, LB, Kanner, RE, Lee, SM, et al. Variability of spirometry in chronic obstructive pulmonary disease: results from two clinical trials. American Journal of Respiratory & Critical Care Medicine 2006;173:1106–13.

90. Simpson, AJ, Potter, JM, Koerbin, G, et al. Use of observed within-person variation of cardiac troponin in emergency department patients for determination of biological variation and percentage and absolute reference change values. Clinical Chemistry 2014;60:848–54.

91. Alvarez, C, Castilla, JA, Martínez, L, Ramírez, JP, Vergara, F, and Gaforio, JJ. Biological variation of seminal parameters in healthy subjects. Human Reproduction 2003;18:2082–2088.

92. Bandaranayake, N, Ankrah-Tetteh, T, Wijeratne, S, and Swaminathan, R. Intra-individual variation in creatinine and cystatin C. Clinical Chemistry and Laboratory Medicine 2007;45:1237–9.

93. Dednam, M, Vorster, BC, and Ubbink, JB. Biological variation of myeloperoxidase. Clinical Chemistry 2008;54:223–5.

94. Cembrowski, GS, Tran, DV, and Higgins, TN. The use of serial patient blood gas, electrolyte and glucose results to derive biologic variation: a new tool to assess the acceptability of intensive care unit testing. Clinical Chemistry and Laboratory Medicine 2010;48:1447–54.

95. Carlsen, S, Petersen, PH, Skeie, S, Skadberg, O, and Sandberg, S. Within-subject biological variation of glucose and HbA(1c) in healthy persons and in type 1 diabetes patients. Clinical Chemistry and Laboratory Medicine 2011;49:1501–7.

96. Levey, AS, Coresh, J, Greene, T, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Annals of Internal Medicine 2006;145:247–54.

97. Levey, AS, Stevens, LA, Schmid, CH, et al. A new equation to estimate glomerular filtration rate. Annals of Internal Medicine 2009;150:604–12.

98. Inker, LA, Schmid, CH, Tighiouart, H, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. New England Journal of Medicine 2012;367:20–9.

99. 't Lam, RUE. Scrutiny of variance results for outliers: Cochran's test optimized. Analytica Chimica Acta 2010;659:68–84.

100. Ghasemi, A and Zahediasl, S. Normality tests for statistical analysis: a guide for non-statisticians. International Journal Endocrinology and Metabolism 2012;10:486–9.

101. Kreft, C. Are multilevel techniques necessary? An overview, including simulation studies. California State University, Los Angeles 1996.

102. Snijders, TAB and Bosker, RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. 2nd. London: Sage, 2012.

103. Snijders, T and Bosker, R. Standard errors and sample sizes for two-level research. Journal of Educational Statistics 1993;18:237–259.

104. J.J. Hox, JJ. Multilevel modeling: when and why. In: *Classification, Data Analysis, and Data Highways*. Ed. by Balderjahn, RM and Schader, M. Berlin: Springer., 1998:147–154.

105. Mood, AM, Graybill, FA, and Boes, DC. Introduction to the Theory of Statistics. Third. New York: McGraw-Hill, 1974:540–541.

106. Burton, A, Altman, DG, Royston, P, and Holder, RL. The design of simulation studies in medical statistics. Statistics in Medicine 2006;25:4279–92.

107. McNeish, D. Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. Multivariate Behavioral Research 2017;52:661–670.

108. Cousineau, D. Outlier detection and treatment: a review. International Journal of Psychological Research 2010.

109. Aguinis, H, Gottfredson, RK, and Joo, H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods 2013;16:270–301.

110. Leys, C, Ley, C, Klein, O, Bernard, P, and Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology 2013;49:764–766.

111. Tukey, J. Exploratory Data Analysis. Addison-Wesley, 1970. URL: `https://books.google.co.uk/books?id=HCsSLgEACAAJ`.

112. Dean, RB and Dixon, WJ. Simplified Statistics for Small Numbers of Observations. Analytical Chemistry 1951;23:636–638.

113. Grubbs, FE. Sample Criteria for Testing Outlying Observations. Annals of Mathematical Statistics 1950;21:27–58.

114. Birmingham Quality group at the National External Quality Assessment Scheme (NEQAS). personal communication. 2019.

115. Taylor-Phillips, S. personal communication. 2018.

116. Selby, PJ, Banks, RE, Gregory, W, et al. A Multi-Centre Programme into the Evaluation of Biomarkers Suitable for Use in Patients with Kidney and Liver Diseases: Chapter 5: A review of monitoring-related methodology literature. Programme Grants for Applied Research 2018.

117. Takahashi, O, Glasziou, PP, Perera, R, Shimbo, T, and Fukui, T. Blood pressure re-screening for healthy adults: what is the best measure and interval? Journal of Human Hypertension 2012;26:540–6.

118. Takahashi, O, Glasziou, PP, Perera, R, et al. Lipid re-screening: what is the best measure and interval? Heart 2010;96:448–52.

119. Bellera, C, Hanley, J, Joseph, L, and Albertsen, P. A statistical evaluation of rules for biochemical failure after radiotherapy in men treated for prostate cancer. International Journal of Radiation Oncology, Biology, Physics 2009;75:1357–63.

120. Sölétormos, G, Hyltoft Petersen, P, and Dombernowsky, P. Progression criteria for cancer antigen 15.3 and carcinoembryonic antigen in metastatic breast cancer compared by computer simulation of marker data. Clinical Chemistry 2000;46:939–49.

121. Glasziou, P, Chalmers, I, Rawlins, M, and McCulloch, P. When are randomised trials unnecessary? Picking signal from noise. British Medical Journal 2007;334:349–51.

122. Bell, KJL, Irwig, L, Craig, JC, and Macaskill, P. Use of randomised trials to decide when to monitor response to new treatment. British Medical Journal 2008;336:361–5.

123. Bell, KJL, Hayen, A, Macaskill, P, et al. Value of routine monitoring of bone mineral density after starting bisphosphonate treatment: secondary analysis of trial data. British Medical Journal 2009;338:b2266–b2266.

124. When To Start Consortium. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. Lancet 2009;373:1352–63.

125. Ahdieh-Grant, L. When to Initiate Highly Active Antiretroviral Therapy: A Cohort Approach. American Journal of Epidemiology 2003;157:738–746.

126. Thompson, SG and Pocock, SJ. The variability of serum cholesterol measurements: implications for screening and monitoring. Journal of Clinical Epidemiology 1990;43:783–789.

127. Bell, KJL, Kirby, A, Hayen, A, Irwig, L, and Glasziou, P. Monitoring adherence to drug treatment by using change in cholesterol concentration: secondary analysis of trial data. British Medical Journal 2011;342:d12–d12.

128. Bell, KJL, Hayen, A, Macaskill, P, Craig, JC, Neal, BC, and Irwig, L. Mixed models showed no need for initial response monitoring after starting antihypertensive therapy. Journal of Clinical Epidemiology 2009;62:650–9.

129. Powers, BJ, Olsen, MK, Smith, VA, Woolson, RF, Bosworth, HB, and Oddone, EZ. Measuring blood pressure for decision making and quality reporting: where and how many measures? Annals of Internal Medicine 2011;154:781–8, 781–8.

130. Thiébaut, R, Chûne, G, Jacqmin-Gadda, H, et al. Time-updated CD4+ T lymphocyte count and HIV RNA as major markers of disease progression in naive HIV-1-infected patients treated with a highly active antiretroviral therapy: the Aquitaine cohort, 1996-2001. Journal of Acquired Immune Deficiency Syndromes 2003;33:380–6.

131. Wolbers, M, Babiker, A, Sabin, C, et al. Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy–the CASCADE collaboration: a collaboration of 23 cohort studies. PLoS Medicine 2010;7:e1000239–e1000239.

132. Proust-Lima, C, Séne, M, Taylor, JMG, and Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: A review. Statistical Methods in Medical Research 2014;23:74–90.

133. Proust-Lima, C and Taylor, JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics 2009;10:535–49.

134. Slate, EH and Turnbull, BW. Statistical models for longitudinal biomarkers of disease onset. Statistics in Medicine 2000;19:617–37.

135. Bellera, C, Hanley, J, Joseph, L, and Albertsen, P. Hierarchical changepoint models for biochemical markers illustrated by tracking postradiotherapy prostate-specific antigen series in men with prostate cancer. Annals of Epidemiology 2008;18:270–82.

136. Inoue, LYT, Etzioni, R, Slate, EH, Morrell, C, and Penson, DF. Combining longitudinal studies of PSA. Biostatistics 2004;5:483–500.

137. Subtil, F and Rabilloud, M. Robust non-linear mixed modelling of longitudinal PSA levels after prostate cancer treatment. Statistics in Medicine 2010;29:573–87.

138. Taylor, JMG, Yu, M, and Sandler, HM. Individualized predictions of disease progression following radiation therapy for prostate cancer. Journal of Clinical Oncology 2007;23:816–25.

139. de Long, ER, Vernon, WB, and Bollinger, RR. Sensitivity and specificity of a monitoring test. Biometrics 1985;41:947–958.

140. Cole, SR, Li, R, Anastos, K, et al. Accounting for leadtime in cohort studies: evaluating when to initiate HIV therapies. Statistics in Medicine 2004;23:3351–63.

141. Walter, SD and Day, NE. Estimation of the duration of a pre-clinical disease state using screening data. American Journal of Epidemiology 1983;118:865–86.

142. Day, NE and Walter, SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. Biometrics 1984;40:1–14.

143. Etzioni, R and Shen, Y. Estimating asymptomatic duration in cancer: the AIDS connection. Statistics in Medicine 1997;16:627–44.

144. Zelen, M. Optimal Scheduling of Examinations for the Early Detection of Disease. Biometrika 1993;80:279–279.

145. Lee, SJ and Zelen, M. Scheduling periodic examinations for the early detection of disease: applications to breast cancer. Journal of the American Statistical Association 1998;93:1271–1281.

146. Frame, PS and Frame, JS. Determinants of cancer screening frequency: the example of screening for cervical cancer. Journal of the American Board of Family Practice 1998.

147. Lee, S, Huang, H, and Zelen, M. Early detection of disease and scheduling of screening examinations. Statistical Methods in Medical Research 2004;13:443–456.

148. McIntosh, MW, Urban, N, and Karlan, B. Generating longitudinal screening algorithms using novel biomarkers for disease. Cancer Epidemiology, Biomarkers & Prevention 2002;11:159–66.

149. McIntosh, MW and Urban, N. A parametric empirical Bayes method for cancer screening using longitudinal observations of a biomarker. Biostatistics 2003;4:27–40.

150. Cai, T, Pepe, MS, Zheng, Y, Lumley, T, and Jenny, NS. The sensitivity and specificity of markers for event times. Biostatistics 2006;7:182–97.

151. Zheng, Y and Heagerty, PJ. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. Biostatistics 2004;5:615–32.

152. Subtil, F, Pouteil-Noble, C, Toussaint, S, Villar, E, and Rabilloud, M. A Simple Modeling-free Method Provides Accurate Estimates of Sensitivity and Specificity of Longitudinal Disease Biomarkers. Methods of Information in Medicine 2009;48:299–305.

153. Parker, CB and DeLong, ER. ROC methodology within a monitoring framework. Statistics in Medicine 2003;22:3473–88.

154. Etzioni, R, Pepe, M, Longton, G, and Goodman, G. Incorporating the Time Dimension in Receiver Operating Characteristic Curves: A Case Study of Prostate Cancer. Medical Decision Making 1999;19:242–251.

155. Macaskill, P. Control charts and control limits in long-term monitoring. In: *Evidence-based Medical Monitoring: From Principles to Practice*. Ed. by Glasziou, PP, Irwig, L, and Aronson, JK. 2008:90–102.

156. Petersen, PH. Making the most of a patient's laboratory data: optimisation of signal-to-noise ratio. Clinical Biochemist Reviews 2005;26:91–6.

157. Petersen, PH, Jensen, EA, and Brandslund, I. Analytical performance, reference values and decision limits. A need to differentiate between reference intervals and decision limits and to define analytical quality specifications. Clinical Chemistry and Laboratory Medicine 2012;50:819–31.

158. Klee, GG. Establishment of outcome-related analytic performance goals. Clinical Chemistry 2010;56:714–22.

159. Tennant, RC, Mohammed, MA, Coleman, JJ, and Martin, U. Monitoring patients using control charts: a systematic review. International Journal for Quality in Health Care : Journal of the International Society for Quality in Health Care 2007;19:187–94.

160. Thor, J, Lundberg, J, Ask, J, et al. Application of statistical process control in healthcare improvement: systematic review. Quality & Safety in Health Care 2007;16:387–99.

161. Gavit, P, Baddour, Y, and Tholmer, R. Use of change-point analysis for process monitoring and control. BioPharm International 2009.

162. Omar, F, van der Watt, GF, and Pillay, TS. Reference change values: how useful are they? Journal of Clinical Pathology 2008;61:426–7.

163. Biosca, C, Ricós, C, Lauzurica, R, and Petersen, PH. Biological variation at long-term renal post-transplantation. Clinica Chimica Acta 2006;368:188–91.

164. Clerico, A and Emdin, M. Diagnostic accuracy and prognostic relevance of the measurement of cardiac natriuretic peptides: a review. Clinical Chemistry 2004;50:33–50.

165. Karnon, J, Goyder, E, Tappenden, P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technology Assessment 2007;11:iii–iv, ix–xi, 1–145.

166. Baker, RD. Use of a mathematical model to evaluate breast cancer screening policy. Health Care Management Science 1998;1:103–13.

167. Parmigiani, G. Timing medical examinations via intensity functions. Biometrika 1997;84:803–816.

168. Sutton, AJ, Cooper, NJ, Goodacre, S, and Stevenson, M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Medical Decision Making 2008;28:650–67.

169. Palmer, S and Smith, PC. Incorporating option values into the economic evaluation of health care technologies. Journal of Health Economics 2000;19:755–66.

170. Driffield, T and Smith, PC. A real options approach to watchful waiting: theory and an illustration. Medical Decision Making 2007;27:178–88.

171. Meyer, E and Rees, R. Watchfully waiting: medical intervention as an optimal investment decision. Journal of Health Economics 2012;31:349–58.

172. Shechter, SM, Alagoz, O, and Roberts, MS. Irreversible treatment decisions under consideration of the research and development pipeline for new therapies. IIE Transactions 2010;42:632–642.

173. Whynes, DK. Optimal times of transfer between therapies: a mathematical framework. Journal of Health Economics 1995;14:477–90.

174. Lasserre, P, Moatti, JP, and Soubeyran, A. Early initiation of highly active antiretroviral therapies for AIDS: dynamic choice with endogenous and exogenous learning. Journal of Health Economics 2006;25:579–98.

175. Glasziou, P. How much monitoring? British Journal of General Practice 2007:350–351.

176. Selby, PJ, Banks, RE, Gregory, W, et al. A Multi-Centre Programme into the Evaluation of Biomarkers Suitable for Use in Patients with Kidney and Liver Diseases: Chapter 3: How is evidence being used to make recommendations about monitoring: the example of prostate specific antigen (PSA)? Programme Grants for Applied Research 2018.

177. ISRCTN. ISRCTN Register. URL: http://www.controlled-trials.com/ISRCTN74815110/elucidate.

178. Dancygier, H. Clinical Hepatology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. URL: http://link.springer.com/10.1007/978-3-540-93842-2.

179. Poynard, T, Bedossa, P, and Opolon, P. Natural history of liver fibrosis progression in patients with chronic hepatitis C. Lancet 1997;349:825–832.

180. Rosenberg, WMC, Voelker, M, Thiel, R, et al. Serum markers detect the presence of liver fibrosis: A cohort study. Gastroenterology 2004;127:1704–1713.

181. Siemens. ELF Test.

182. Parkes, J. Longitudinal data set. 2010.

183. Parkes, J. personal communication. 2011.

184. Curtis, KA, Ambrose, KM, Kennedy, MS, and Owen, SM. Evaluation of dried blood spots with a multiplex assay for measuring recent HIV-1 infection. PLoS One 2014;9:e107153.

185. Gabriele, A, Marco, V, Gatto, L, et al. Reproducibility of the Carpet View system: a novel technical solution for display and off line analysis of OCT images. International Journal of Cardiovascular Imaging 2014;30:1225–33.

186. Manley, SE, Hikin, LJ, Round, RA, et al. Comparison of IFCC-calibrated HbA(1c) from laboratory and point of care testing systems. Diabetes Research and Clinical Practice 2014;105:364–72.

187. Saez-Benito Godino, A, Vergara Chozas, JM, Marquez Ronchel, A, et al. Multicentre evaluation of glycated haemoglobin (HbA1c) of Roche Diagnostics in Andalusia. Clinical Biochemistry 2014;47:1108–11.

188. Wu, AH, Yang, HS, and Thoren, K. Biological variation of the osmolality and the osmolal gap. Clinical Biochemistry 2014;47:130–1.

189. Alizai, H, Virayavanich, W, Joseph, GB, et al. Cartilage lesion score: comparison of a quantitative assessment score with established semiquantitative MR scoring systems. Radiology 2014;271:479–87.

190. Frings, V, van Velden, FH, Velasquez, LM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. Radiology 2014;273:539–48.

191. Giles, SL, Messiou, C, Collins, DJ, et al. Whole-body diffusion-weighted MR imaging for assessment of treatment response in myeloma. Radiology 2014;271:785–94.

192. Knobloch, V, Binter, C, Kurtcuoglu, V, and Kozerke, S. Arterial, venous, and cerebrospinal fluid flow: simultaneous assessment with Bayesian multipoint velocity-encoded MR imaging. Radiology 2014;270:566–73.

193. Roujol, S, Weingartner, S, Foppa, M, et al. Accuracy, precision, and reproducibility of four T1 mapping sequences: a head-to-head comparison of MOLLI, ShMOLLI, SASHA, and SAPPHIRE. Radiology 2014;272:683–9.

194. Suh, CH, Kim, HS, Lee, SS, et al. Atypical imaging features of primary central nervous system lymphoma that mimics glioblastoma: utility of intravoxel incoherent motion MR imaging. Radiology 2014;272:504–13.

195. Thevenot, J, Hirvasniemi, J, Pulkkinen, P, et al. Assessment of risk of femoral neck fracture with radiographic texture parameters: a retrospective study. Radiology 2014;272:184–91.

196. Karon, BS, Tolan, NV, Koch, CD, et al. Precision and reliability of 5 platelet function tests in healthy volunteers and donors on daily antiplatelet agent therapy. Clinical Chemistry 2014;60:1524–31.

197. Noceti, OM, Woillard, JB, Boumediene, A, et al. Tacrolimus pharmacodynamics and pharmacogenetics along the calcineurin pathway in human lymphocytes. Clinical Chemistry 2014;60:1336–45.

198. Beco, J, Leonard, D, and Leonard, F. Study of the female urethra's submucous vascular plexus by color Doppler. World Journal of Urology 1998;16:224–8.

199. Heit, M. Intraurethral sonography and the test-retest reliability of urethral sphincter measurements in women. Journal of Clinical Ultrasound 2002;30:349–55.

200. Oelke, M, Mamoulakis, C, Ubbink, DT, de la Rosette, JJ, and Wijkstra, H. Manual versus automatic bladder wall thickness measurements: a method comparison study. World Journal of Urology 2009;27:747–53.

201. Otcenasek, M, Halaska, M, Krcmar, M, Maresova, D, and Halaska, MG. New approach to the urogynecological ultrasound examination. European Journal of Obstetrics, Gynecology, & Reproductive Biology 2002;103:72–4.

202. Naresh, CN, Hayen, A, Weening, A, Craig, JC, and Chadban, SJ. Day-to-day variability in spot urine albumin-creatinine ratio. American Journal of Kidney Diseases 2013;62:1095–101.

203. Ristiniemi, N, Savage, C, Bruun, L, Pettersson, K, Lilja, H, and Christensson, A. Evaluation of a new immunoassay for cystatin C, based on a double monoclonal principle, in men with normal and impaired renal function. Nephrology Dialysis Transplantation 2012;27:682–7.

204. Rule, AD, Bailey, KR, Lieske, JC, Peyser, PA, and Turner, ST. Estimating the glomerular filtration rate from serum creatinine is better than from cystatin C for evaluating risk factors associated with chronic kidney disease. Kidney International 2013;83:1169–76.

205. Sjostrom, PA, Jones, IL, and Tidman, MA. Cystatin C as a filtration marker–haemodialysis patients expose its strengths and limitations. Scandinavian Journal of Clinical and Laboratory Investigation 2009;69:65–72.

206. Walser, M, Drew, HH, and Guldan, JL. Prediction of glomerular filtration rate from serum creatinine concentration in advanced chronic renal failure. Kidney International 1993;44:1145–8.

207. Beeh, KM, Beier, J, Kornmann, O, Mander, A, and Buhl, R. Long-term repeatability of induced sputum cells and inflammatory markers in stable, moderately severe COPD. Chest 2003;123:778–83.

208. Liistro, G, van Welde, C, Vincken, W, et al. Technical and functional assessment of 10 office spirometers: A multicenter comparative study. Chest 2006;130:657–65.

209. Madsen, F, Ulrik, CS, Dirksen, A, et al. Patient-administered sequential spirometry in healthy volunteers and patients with alpha 1-antitrypsin deficiency. Respiratory Medicine 1996;90:131–8.

210. McCarley, C, Hanneman, SK, Padhye, N, and Smolensky, MH. A pilot home study of temporal variations of symptoms in chronic obstructive lung disease. Biological Research for Nursing 2007;9:8–20.

211. Timmins, SC, Coatsworth, N, Palnitkar, G, et al. Day-to-day variability of oscillatory impedance and spirometry in asthma and COPD. Respiratory Physiology & Neurobiology 2013;185:416–24.

212. Alvarez, L, Ricós, C, Peris, P, et al. Components of biological variation of biochemical markers of bone turnover in Paget's bone disease. Bone 2000;26.

213. Andersen, TB, Erlandsen, EJ, Frokiaer, J, Eskild-Jensen, A, and Brochner-Mortensen, J. Comparison of within- and between-subject variation of serum cystatin C and serum creatinine in children aged 2-13 years. Scandinavian Journal of Clinical and Laboratory Investigation 2010;70:54–9.

214. Andersson, AM, Carlsen, E, Petersen, JH, and Skakkebaek, NE. Variation in levels of serum inhibin B, testosterone, estradiol, luteinizing hormone, follicle-stimulating hormone, and sex hormone-binding globulin in monthly samples from healthy men during a 17-month period: possible effects of seasons. Journal of Clinical Endocrinology and Metabolism 2003;88:932–7.

215. Ankrah-Tetteh, T, Wijeratne, S, and Swaminathan, R. Intraindividual variation in serum thyroid hormones, parathyroid hormone and insulin-like growth factor-1. Annals of Clinical Biochemistry 2008;45:167–9.

216. Braga, F, Dolci, A, Montagnana, M, et al. Revaluation of biological variation of glycated hemoglobin (HbA(1c)) using an accurately designed protocol and an assay traceable to the IFCC reference system. Clinica Chimica Acta 2011;412:1412–6.

217. Brown, LF and Fraser, CG. Assay validation and biological variation of serum receptor for advanced glycation end-products. Annals of Clinical Biochemistry 2008;45:518–9.

218. Browne, RW, Koury, ST, Marion, S, Wilding, G, Muti, P, and Trevisan, M. Accuracy and biological variation of human serum paraoxonase 1 activity and polymorphism (Q192R) by kinetic enzyme assay. Clinical Chemistry 2007;53:310–7.

219. Cheuvront, SN, Ely, BR, Kenefick, RW, and Sawka, MN. Biological variation and diagnostic accuracy of dehydration assessment markers. American Journal of Clinical Nutrition 2010;92:565–73.

220. Cho, LW, Jayagopal, V, Kilpatrick, ES, and Atkin, SL. The biological variation of C-reactive protein in polycystic ovarian syndrome. Clinical Chemistry 2005;51:1905–7.

221. Corte, Z and Venta, R. Biological variation of free plasma amino acids in healthy individuals. Clinical Chemistry and Laboratory Medicine 2010;48:99–104.

222. Desmeules, P, Cousineau, J, and Allard, P. Biological variation of glycated haemoglobin in a paediatric population and its application to calculation of significant change between results. Annals of Clinical Biochemistry 2010;47:35–8.

223. Dittadi, R, Meo, S, and Gion, M. Biological variation of plasma chromogranin A. Clinical Chemistry and Laboratory Medicine 2004;42:109.

224. Dittadi, R, Gelisio, P, Rossi, L, Frigato, F, and Gion, M. Biological variability evaluation and comparison of three different methods for C-peptide measurement. Clinical Chemistry and Laboratory Medicine 2008;46:1480–2.

225. Dittadi, R, Peloso, L, and Gion, M. Within-subject biological variation in disease: the case of tumour markers. Annals of Clinical Biochemistry 2008;45:226–7.

226. Garde, AH, Hansen, AM, Skovgaard, LT, and Christensen, JM. Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A(1c), IgA, prolactin, and free testosterone in healthy women. Clinical Chemistry 2000;46:551–9.

227. González, C, Cava, F, Ayllón, A, Guevara, P, Navajo José, A, and González-Buitrago José, M. Biological Variation of Interleukin-beta, Interleukin-8 and Tumor Necrosis Factor-alpha in Serum of Healthy Individuals. Clinical Chemistry and Laboratory Medicine 2001;39:836.

228. Jensen, E, Petersen, PH, Blaabjerg, O, and Hegedus, L. Biological variation of thyroid autoantibodies and thyroglobulin. Clinical Chemistry and Laboratory Medicine 2007;45:1058–64.

229. Kristoffersen, AH, Petersen, PH, and Sandberg, S. A model for calculating the within-subject biological variation and likelihood ratios for analytes with a time-dependent change in concentrations; exemplified with the use of D-dimer in suspected venous thromboembolism in healthy pregnant women. Annals of Clinical Biochemistry 2012;49:561–9.

230. Lara-Riegos, J, Brambila, E, Ake-Ku, A, et al. Short-term estimation and application of biological variation of small dense low-density lipoproteins in healthy individuals. Clinical Chemistry and Laboratory Medicine 2013;51:2167–72.

231. McKinley, MC, Strain, JJ, McPartlin, J, Scott, JM, and McNulty, H. Plasma homocysteine is not subject to seasonal variation. Clinical Chemistry 2001;47:1430–6.

232. Melzi d'Eril, G, Anesi, A, Maggiore, M, and Leoni, V. Biological variation of serum amyloid A in healthy subjects. Clinical Chemistry 2001;47:1498–9.

233. Melzi d'Eril, G, Tagnochetti, T, Nauti, A, et al. Biological variation of N-terminal pro-brain natriuretic peptide in healthy individuals. Clinical Chemistry 2003;49:1554–5.

234. Meo, S, Dittadi, R, Gion, M, and Italian Committee for Quality Control in the Oncology Laboratory and Italian Network for Quality Assessment of Tumor Biomarkers. Biological variation of vascular endothelial growth factor. Clinical Chemistry and Laboratory Medicine 2005;43:342–3.

235. Moller, HJ, Petersen, PH, Rejnmark, L, and Moestrup, SK. Biological variation of soluble CD163. Scandinavian Journal of Clinical and Laboratory Investigation 2003;63:15–21.

236. Mosca, A, Paleari, R, and Wild, B. Analytical goals for the determination of HbA(2). Clinical Chemistry and Laboratory Medicine 2013;51:937–41.

237. Nguyen, TV, Nelson, AE, Howe, CJ, et al. Within-subject variability and analytic imprecision of insulinlike growth factor axis and collagen markers: implications for clinical diagnosis and doping tests. Clinical Chemistry 2008;54:1268–76.

238. Pagani, F and Panteghini, M. Biological variation in serum activities of three hepatic enzymes. Clinical Chemistry 2001;47:355–6.

239. Pineda-Tenor, D, Laserna-Mendieta, EJ, Timon-Zapata, J, et al. Biological variation and reference change values of common clinical chemistry and haematologic laboratory analytes in the elderly population. Clinical Chemistry and Laboratory Medicine 2013;51:851–62.

240. Reclos, GJ, Tanyalcin, T, and Pass, KA. Estimation of the biological variation of glucose-6-phosphate dehydrogenase in dried blood spots. Accreditation and Quality Assurance 2006;11:308–312.

241. Reinhard, M, Erlandsen, EJ, and Randers, E. Biological variation of cystatin C and creatinine. Scandinavian Journal of Clinical and Laboratory Investigation 2009;69:831–6.

242. Rohlfing, C, Wiedmeyer, HM, Little, R, et al. Biological variation of glycohemoglobin. Clinical Chemistry 2002;48:1116–8.

243. Rossi, E, Adams, LA, Ching, HL, Bulsara, M, MacQuillan, GC, and Jeffrey, GP. High biological variation of serum hyaluronic acid and Hepascore, a biochemical marker model for the prediction of liver fibrosis. Clinical Chemistry and Laboratory Medicine 2013;51:1107–14.

244. Serteser, M, Coskun, A, Unsal, I, and Inal, TC. Biological variation in pregnancy-associated plasma protein-A in healthy men and non-pregnant healthy women. Clinical Chemistry and Laboratory Medicine 2012;50:2239–41.

245. Shand, B, Elder, P, Scott, R, Frampton, C, and Willis, J. Biovariability of plasma adiponectin. Clinical Chemistry and Laboratory Medicine 2006;44:1264–8.

246. Talwar, DK, Azharuddin, MK, Williamson, C, Teoh, YP, McMillan, DC, and O'Reilly, D. Biological variation of vitamins in blood of healthy individuals. Clinical Chemistry 2005;51:2145–50.

247. Trapé, J, Aliart, Ml, Brunet, M, Dern, E, Abadal, E, and Queraltó Josep, M. Reference Change Value for HbA1c in Patients with Type 2 Diabetes Mellitus. Clinical Chemistry and Laboratory Medicine 2000;38:1283.

248. Trapé, J, Botargues, JM, Porta, F, et al. Reference change value for alpha-fetoprotein and its application in early detection of hepatocellular carcinoma in patients with hepatic disease. Clinical Chemistry 2003;49:1209–11.

249. Trapé, J, Perez de Olaguer, J, Buxo, J, and Lopez, L. Biological variation of tumor markers and its application in the detection of disease progression in patients with non-small cell lung cancer. Clinical Chemistry 2005;51:219–22.

250. Trapé, J, Franquesa, J, Sala, M, et al. Determination of biological variation of alpha-fetoprotein and choriogonadotropin (beta chain) in disease-free patients with testicular cancer. Clinical Chemistry and Laboratory Medicine 2010;48:1799–801.

251. Valero-Politi, J, Ginard-Salvá, M, and González-Alba, JM. Annual Rhythmic and Non-Rhythmic Biological Variation of Magnesium and Ionized Calcium Concentrations. Clinical Chemistry and Laboratory Medicine 2001;39:45.

252. van der Merwe, D, Ubbink, J, Delport, R, Becker, P, Dhatt, G, and Vermaak, W. Biological variation in sweat sodium chloride conductivity. Annals of Clinical Biochemistry 2002.

253. van Hoydonck, PG, Schouten, EG, and Temme, EH. Reproducibility of blood markers of oxidative status and endothelial function in healthy individuals. Clinical Chemistry 2003;49:963–5.

254. Vasile, VC, Saenger, AK, Kroning, JM, and Jaffe, AS. Biological and analytical variability of a novel high-sensitivity cardiac troponin T assay. Clinical Chemistry 2010;56:1086–90.

255. Vasile, VC, Saenger, AK, Kroning, JM, Klee, GG, and Jaffe, AS. Biologic variation of a novel cardiac troponin I assay. Clinical Chemistry 2011;57:1080–1.

256. Viljoen, A, Singh, DK, Twomey, PJ, and Farrington, K. Analytical quality goals for parathyroid hormone based on biological variation. Clinical Chemistry and Laboratory Medicine 2008;46:1438–42.

257. Wu, AH, Lu, QA, Todd, J, Moecks, J, and Wians, F. Short- and long-term biological variation in cardiac troponin I measured with a high-sensitivity assay: implications for clinical practice. Clinical Chemistry 2009;55:52–8.

258. Wu, AH, Akhigbe, P, and Wians, F. Long-term biological variation in cardiac troponin I. Clinical Biochemistry 2012;45:714–6.

259. Delanaye, P, Cavalier, E, Depas, G, Chapelle, JP, and Krzesinski, JM. New data on the intraindividual variation of cystatin C. Nephron Clinical Practice 2008;108:c246–8.

260. Toffaletti, JG and McDonnell, EH. Variation of serum creatinine, cystatin C, and creatinine clearance tests in persons with normal renal function. Clinica Chimica Acta 2008;395:115–9.

261. Gaspari, F, Perico, N, Matalone, M, et al. Precision of plasma clearance of iohexol for estimation of GFR in patients with renal disease. Journal of the American Society of Nephrology 1998;9:310–3.

262. Gowans, EMS and Fraser, CG. Biological Variation of Serum and Urine Creatinine and Creatinine Clearance: Ramifications for Interpretation of Results and Patient Care. Annals of Clinical Biochemistry 1988;25:259–263.

263. Keevil, BG, Kilpatrick, ES, Nichols, SP, and Maylor, PW. Biological variation of cystatin C: implications for the assessment of glomerular filtration rate. Clinical Chemistry 1998;44:1535–9.

264. Kuo, HC. Measurement of detrusor wall thickness in women with overactive bladder by transvaginal and transabdominal sonography. Internation Urogynecology Journal and Pelvic Floor Dysfunction 2009;20:1293–9.

265. Lekskulchai, O and Dietz, HP. Detrusor wall thickness as a test for detrusor overactivity in women. Ultrasound in Obstetrics and Gynecology 2008;32:535–9.

266. Panayi, DC, Digesu, GA, Tekkis, P, Fernando, R, and Khullar, V. Ultrasound measurement of vaginal wall thickness: a novel and reliable technique. International Urogynecology Journal 2010;21:1265–70.

267. Tubaro, A, Khullar, V, Oelke, M, et al. Intra- and inter-reader variability of transvaginal ultrasound bladder wall thickness measurements: results from the shrink study. Neurourology and Urodynamics 2013;32:711–712.