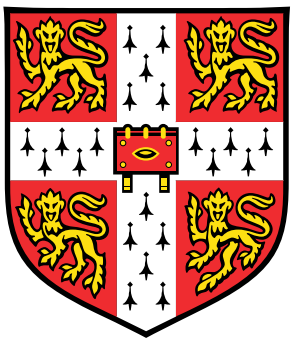
Statistical Methods to Improve Efficiency in Composite Endpoint Analysis



Martina McMenamin

MRC Biostatistics Unit
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Magdalene College August 2019

This thesis is dedicated to my beautiful nephew Oscar Knox, who lost his battle with neuroblastoma at five years of age. Through thoughtful science and principled collaboration may we save other families from this heartbreak.

Acknowledgements

I would like to acknowledge my supervisor Professor James Wason for his invaluable insight and guidance throughout the duration of my PhD, without which I could not have succeeded. I would also like to acknowledge my secondary supervisors Dr. Jessica Barrett and Dr. Anna Berglind for the helpful input and feedback they provided during the project. I would like to thank AstraZeneca for sharing their clinical trial data and wish to acknowledge and commend the bravery and selflessness of all the patients involved in these studies.

The work presented in Section 1.3 was undertaken as a collaboration with Professor James Wason and Dr. Susanna Dodd, for which James is the first author. I was responsible for contributing to the search strategies, performing one third of the review and providing feedback on subsequent drafts.

Finally, I would like to acknowledge and sincerely thank my family and friends, in particular my partner David for his unwavering support and encouragement.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Acknowledgements and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Acknowledgements and specified in the text.

Martina McMenamin
August 2019

Statistical Methods to Improve Efficiency in Composite Endpoint Analysis

Martina McMenamin

Composite endpoints combine a number of outcomes to assess the efficacy of a treatment. They are used in situations where it is difficult to identify a single relevant endpoint, such as in complex multisystem diseases. Our focus in this thesis is on composite responder endpoints, which allocate patients as either 'responders' or 'non-responders' based on whether they cross predefined thresholds in the individual outcomes. These composites are often combinations of continuous and discrete measures and are typically collapsed into a single binary endpoint and analysed using logistic regression. However, this is at the expense of losing information on how close each patient was to the responder threshold. As well as being inefficient the analysis is sensitive to misclassification due to measurement error. The augmented binary method was introduced to improve the analysis of composite responder endpoints comprised of a single continuous and binary endpoint, by making use of the continuous information.

In this thesis we build on this work to address some of the existing limitations. We implement small sample corrections for the standard binary and augmented binary methods and assess the performance for application in rare disease trials, where the gains are most needed. We find that employing the small sample corrected augmented binary method results in a reduction of required sample size of 32%. Motivated by systemic lupus erythematosus (SLE), we consider the case where the composite has multiple continuous, ordinal and binary components. We adapt latent variable models for application to these endpoints and assess the performance in simulated data and phase IIb trial data in SLE. Our findings show reductions in required sample size of at least 60%, however the magnitude of the gains depends on which components drive response. Finally, we develop a method for sample size estimation so that the model may be used as a primary analysis method in clinical trials. We assess the impact of correlation structure and drivers of response on the sample size required.

Table of contents

List of figures **xiii**

List of tables **xvii**

Nomenclature **xxiii**

1 Introduction **1**

1.1 Composite Endpoints . . . 2

1.1.1 Events in at Least One Component . . . 2

1.1.2 Events in All Components . . . 4

1.1.3 Opportunities and Limitations . . . 4

1.1.4 Recommendations for Use . . . 5

1.2 Existing Methods . . . 7

1.2.1 Notation . . . 7

1.2.2 Standard Binary Method . . . 7

1.2.3 Suissa Method . . . 9

1.2.4 Augmented Binary Method . . . 10

1.3 Scope for Application . . . 13

1.3.1 Motivation . . . 13

1.3.2 Methods . . . 13

1.3.3 Findings . . . 14

1.4 Thesis Aims . . . 14

2 Composite Endpoints in Rare Disease Trials **19**

2.1 Introduction . . . 19

2.1.1 Motivation . . . 19

2.1.2 Aims . . . 21

2.2 Small Sample Adjustments . . . 23

2.2.1 Binary Component Adjustment . . . 23
2.2.2 Continuous Component Adjustment for GEE . . . 24
2.3 Assessing Properties: Re-sampling . . . 26
2.3.1 Data . . . 26
2.3.2 Model Notation . . . 28
2.3.3 Re-sampling . . . 28
2.3.4 OSKIRA-1 data . . . 29
2.3.5 Results . . . 30
2.3.5.1 ACR20 . . . 30
2.3.5.2 ACR50 . . . 38
2.3.5.3 ACR70 . . . 41
2.4 Assessing Properties: Simulated Example . . . 44
2.4.1 Data Generating Model . . . 44
2.4.2 Results . . . 45
2.5 Discussion . . . 46

3 Complex Composite Structures 51
3.1 Motivation . . . 51
3.1.1 Application: Systemic Lupus Erythematosus . . . 53
3.2 Background . . . 56
3.2.1 Copulas . . . 56
3.2.2 Factorisation . . . 57
3.2.3 Latent Variable Models . . . 57
3.2.4 Extensions to Multivariate Probit Models . . . 58
3.3 Latent Variable Model . . . 59
3.3.1 Notation . . . 59
3.3.2 Model . . . 60
3.3.3 Estimation . . . 63
3.3.4 Inference . . . 64
3.3.5 Pragmatic Considerations . . . 65
3.4 Models for Comparison . . . 70
3.4.1 Augmented Binary Method . . . 70
3.4.2 Standard Binary Method . . . 71
3.5 Simulation Study . . . 71
3.5.1 Data Generating Models . . . 71

3.5.2 Performance Measures . . . 71
3.5.3 Findings . . . 74
3.5.3.1 Varying Treatment Effect . . . 74
3.5.3.2 Varying η_1 . . . 77
3.5.3.3 Components Contributing to Response . . . 78
3.5.3.4 Probability of Response in Each Arm . . . 81
3.6 Sensitivity Analysis . . . 91
3.6.1 Multivariate Skew-Normal Distribution . . . 91
3.6.2 Results . . . 91
3.7 Case Study . . . 97
3.7.1 Trial Data . . . 97
3.7.2 MUSE Primary Analysis . . . 97
3.7.3 Exploratory Data Analysis . . . 97
3.7.4 MUSE Trial Re-Analysis . . . 108
3.7.5 Model Fit . . . 109
3.8 Bias Correction Using the Bootstrap . . . 109
3.8.1 Bootstrap Method . . . 111
3.8.2 Application in the One-Sample Multivariate Case . . . 111
3.9 Discussion . . . 112

4 Sample Size Estimation using the Latent Variable Model 117
4.1 Motivation . . . 117
4.2 Literature Review . . . 118
4.3 Aims . . . 119
4.4 Model . . . 119
4.5 Mixed Outcome Co-Primary Endpoints . . . 121
4.5.1 Hypothesis Testing . . . 121
4.5.2 Overall Power . . . 122
4.5.3 Sample Size Calculation . . . 123
4.5.4 Application to MUSE Trial . . . 123
4.6 Mixed Outcome Composite Endpoints . . . 125
4.6.1 Hypothesis Testing . . . 127
4.6.2 Obtaining Required Quantities . . . 130
4.6.3 Critical Value . . . 130
4.6.4 Power . . . 130

4.6.5 Sample Size Estimation . . . 132
4.7 Empirical Comparisons . . . 133
4.7.1 One Continuous, One Ordinal, One Binary . . . 133
4.7.2 Two Continuous, One Ordinal, One Binary . . . 141
4.8 Application: MUSE Trial . . . 148
4.9 Discussion . . . 149

5 Discussion **153**
5.1 Summary . . . 153
5.2 Limitations . . . 155
5.3 Recommendations . . . 156
5.4 Future Work . . . 157
5.4.1 Multiple Time Points . . . 158
5.4.2 Estimation Methods . . . 159
5.4.3 Other Outcome Types . . . 160
5.4.4 Further Research Directions . . . 162
5.5 Conclusion . . . 163

Bibliography **165**

Appendix A Preprint: Scope for Application **177**

Appendix B Orphanet Paper **197**

Appendix C Orphanet Paper: Supplementary Materials **207**
C.1 Models and Small Sample Adjustments . . . 207
C.2 Supplementary Results: ACR20 . . . 211
C.3 Supplementary Results: ACR50, ACR70 . . . 214
C.4 Supplementary Results: Simulated Example . . . 218

Appendix D Small Sample Adjusted Methods: R Code **221**
D.1 Augmented Binary: GLS . . . 221
D.2 Augmented Binary: GEE . . . 224
D.3 Standard Binary . . . 231

Appendix E Preprint: Complex Composite Structures **233**

Appendix F Latent Variable Method: R Code **279**

List of figures

1.1 Illustration of the Suissa method for using the Gaussian distribution underpinning a dichotomy 9
1.2 Visual comparison of the steps involved in fitting the standard binary and augmented binary methods 12

2.1 Structure of the composite responder endpoint used in rheumatoid arthritis 27
2.2 Power of the standard binary and augmented binary methods for ACR20 log-odds treatment effect 30
2.3 Type I error rate of the standard binary and augmented binary methods for ACR20 log-odds treatment effect 31
2.4 Power of the standard binary and augmented binary methods for the ACR20 risk difference treatment effect 34
2.5 Type I error rate of the standard binary and augmented binary methods for ACR20 risk difference treatment effect 34
2.6 Type I error rate of the standard binary and augmented binary methods for the ACR50 log-odds treatment effect 39
2.7 Power of the standard binary and augmented binary methods for the ACR50 log-odds treatment effect 39
2.8 Type I error rate of the standard binary and augmented binary methods for the ACR50 risk difference treatment effect 40
2.9 Power of the standard binary and augmented binary methods for the ACR50 risk difference treatment effect 41
2.10 Type I error rate of the standard binary and augmented binary methods for the ACR70 log-odds treatment effect 42
2.11 Power of the standard binary and augmented binary methods for the ACR70 log-odds treatment effect 43

2.12 Type I error rate of the standard binary and augmented binary methods for the ACR70 risk difference treatment effect 43
2.13 Power of the standard binary and augmented binary methods for the ACR70 risk difference treatment effect 44

3.1 Structure of the composite endpoint used in systemic lupus erythematosus 53
3.2 Bivariate probability space for the discrete components given the continuous components 62
3.3 Bias in the treatment effect estimate reported from the latent variable, augmented binary and standard binary methods 74
3.4 Coverage probability reported from the latent variable, augmented binary and standard binary methods 75
3.5 Bias-corrected coverage probability reported from the latent variable, augmented binary and standard binary methods 75
3.6 Statistical power reported from the latent variable, augmented binary and standard binary methods 76
3.7 Relative precision of the latent variable, augmented binary and standard binary methods 77
3.8 Mean squared error reported from the latent variable, augmented binary and standard binary methods 78
3.9 Relative precision of the latent variable, augmented binary and standard binary methods as the responder threshold varies 79
3.10 Statistical power of the latent variable, augmented binary and standard binary methods as the responder threshold varies 79
3.11 Relative precision gains from the latent variable, augmented binary and standard binary methods when different components drive response . . 80
3.12 Estimated probability of response in each arm from the latent variable, augmented binary and standard binary methods 82
3.13 Estimated probability of response in each arm from the latent variable, augmented binary and standard binary methods as the treatment effect varies 82
3.14 Histogram of univariate skew-normal error terms 92
3.15 Histogram for Physician's Global Assessment measure 101
3.16 Hisogram for Physician's Global Assessment measure by treatment arm 101

3.17 Histogram for Systemic Lupus Erythematosus Disease Activity Index in the MUSE study . . . 103
3.18 Histogram for Systemic Lupus Erythematosus Disease Activity Index by treatment arm in the MUSE study . . . 103
3.19 Barplot showing the British Isles Lupus Assessment Group measure in the MUSE study . . . 104
3.20 Observed response rates from the SLE responder index in the MUSE study . . . 106
3.21 Heatmap showing the correlations between the four components of the systemic lupus erythematosus composite endpoint . . . 106
3.22 Violin plots of the components of the systemic lupus erythematosus composite endpoint . . . 107
3.23 Estimated log-odds treatment effects from the latent variable, augmented binary and standard binary methods in the MUSE study . . . 108
3.24 Modified Pearson residuals from the latent variable model for each patient in the MUSE trial . . . 110
3.25 Histogram of the modified Pearson residuals from the latent variable model in the MUSE trial dataset with the corresponding χ^2_9 density . . 110

4.1 Power of the co-primary endpoints in the MUSE dataset . . . 125
4.2 Power of the co-primary endpoints excluding components . . . 126
4.3 Power of the co-primary endpoints excluding components for different correlations . . . 126
4.4 Stages in analysis and hypothesis testing for composite and co-primary endpoints . . . 128
4.5 Sample size per group as risk difference changes using the latent variable and standard binary methods . . . 134
4.6 Boxplots of the estimated variance from the standard binary and latent variable methods . . . 136
4.7 Boxplots of the estimated sample size per group from the standard binary and latent variable methods . . . 137
4.8 Violin plots of the estimated sample size from the latent variable and standard binary methods for different treatment effect structures . . . 137
4.9 Boxplots of the estimated reduction in required sample size by using the latent variable methods . . . 138

4.10 Boxplots comparing the estimated sample size from the latent variable method for composites containing one and two continuous outcomes . . 142
4.11 Boxplots of the estimated reduction in required sample size by using the latent variable method with two continuous components 143
4.12 Boxplots of the estimated sample size per group from the latent variable method for composites containing two continuous, one ordinal and one binary component 146
4.13 Boxplots of the estimated sample size per group from the standard binary method for composites containing two continuous, one ordinal and one binary component 147
4.14 Power of the latent variable method in the MUSE dataset 150

List of tables

1.1 List of conditions using composite responder endpoints or dichotomised continuous variables to determine efficacy 15
1.2 List of conditions using composite responder endpoints or dichotomised continuous variables to determine efficacy (continued) 16
1.3 List of conditions using composite responder endpoints or dichotomised continuous variables to determine efficacy (continued) 17

2.1 Examples of rare diseases using a composite responder endpoint with continuous and binary components 22
2.2 Unadjusted and small sample adjusted methods to be compared 29
2.3 Average ACR20 log-odds treatment effect estimates in the null case for the standard binary and augmented binary methods 32
2.4 Median confidence interval width for the ACR20 log-odds treatment effect estimates from the standard binary and augmented binary methods 33
2.5 Average ACR20 risk difference treatment effect estimates in the null case for the standard binary and augmented binary methods 35
2.6 Median confidence interval width for the ACR20 risk difference treatment effect estimates from the standard binary and augmented binary methods 36
2.7 Percentage of cases with perfect separation in the ACR20 risk difference treatment effect estimate from the standard binary and augmented binary methods 37
2.8 Percentage reduction in average confidence interval width for the ACR20 treatment effect estimate from the standard binary vs. augmented binary method 37
2.9 Average ACR20 risk difference treatment effect estimate and power in simulated data for the small sample corrected standard binary and augmented binary methods 46

2.10 Average confidence interval width for the ACR20 risk difference treatment effect in simulated data from the small sample corrected standard binary and augmented binary methods 47
2.11 Average ACR20 risk difference estimate and type I error rate in simulated data for the small sample corrected standard binary and augmented binary methods 47
2.12 Average confidence interval width for the ACR20 risk difference null effect in the simulated data from the small sample corrected standard binary and augmented binary methods 48

3.1 Examples of diseases using complex composite endpoints with multiple discrete and continuous components 52
3.2 Response definition in Systemic Lupus Erythematosus Responder Index (SRI) 54
3.3 Grading system in the British Isles Lupus Assessment Group index . . 55
3.4 Average execution time for the latent variable, augmented binary and standard binary methods 66
3.5 Benchmarked time for each process required to fit the latent variable model to the systemic lupus erythematosus endpoint 67
3.6 Parameter values for the simulated scenarios which investigate the effect of varying the responder threshold η_1, changing the components driving response and differing treatment effects on the performance of the latent variable, augmented binary and standard binary methods for the systemic lupus erythematosus composite endpoint 72
3.7 Performance measures used to assess the behaviour of the latent variable, augmented binary and binary methods 73
3.8 Bias and coverage estimates from the latent variable, augmented binary and standard binary methods 83
3.9 Bias-corrected coverage and power estimates from the latent variable, augmented binary and standard binary methods 84
3.10 Mean squared error estimates from the latent variable, augmented binary and standard binary methods 85
3.11 Empirical standard error and model standard error estimates from the latent variable, augmented binary and standard binary methods 86

3.12 Estimated probability of response from the latent variable, augmented binary and standard binary methods 87
3.13 Estimated odds ratios from the latent variable, augmented binary and standard binary methods 88
3.14 Relative precision estimates with the 10th centile and 90th centile values from the latent variable, augmented binary and standard binary methods 89
3.15 Median confidence interval width for log-odds treatment effects reported from the latent variable, augmented binary and standard binary methods 90
3.16 Simulation scenarios to investigate deviations from joint normality based on the multivariate skew-normal distribution 92
3.17 Operating characteristics of the latent variable, augmented binary and standard binary methods when data is drawn from a multivariate skew-normal distribution 94
3.18 Mean squared error, empirical standard error and model standard error estimates from the latent variable, augmented binary and standard binary methods when the data are drawn from a multivariate skew-normal distribution 95
3.19 Estimated probability of response in each arm from the latent variable, augmented binary and standard binary methods when the data are drawn from a multivariate skew-normal 95
3.20 Estimated odds ratio treatment effect from the latent variable, augmented binary and standard binary methods when the data are drawn from a multivariate skew-normal distribution 96
3.21 Relative precision estimates from the latent variable, augmented binary and standard binary methods when the data are drawn from a multivariate skew normal distribution 96
3.22 Median confidence interval width of the treatment effect reported from the latent variable, augmented binary and standard binary methods when the data are drawn from a multivariate skew-normal distribution 96
3.23 Demographic and baseline clinical characteristics of patients enrolled in the MUSE study 98
3.24 Summary of efficacy results from the MUSE trial 99
3.25 Observed response rates in each component of the SRI endpoint in the MUSE study 102

3.26 Estimated probability of response in each arm from the latent variable, augmented binary and standard binary methods in the MUSE study . . 108
3.27 Relative precision estimates from the latent variable, augmented binary and standard binary methods in the MUSE study 109
3.28 Log-odds treatment effect estimates from the latent variable, augmented binary and standard binary methods in the MUSE trial dataset and the bootstrap sample 112

4.1 Sample sizes for co-primary endpoints in the MUSE trial data 124
4.2 Methods for determining the target difference in a sample size calculation 131
4.3 Median sample sizes for one continuous, one ordinal and one binary measure using the latent variable method 139
4.4 Median sample sizes for one continuous, one ordinal and one binary measure using the latent variable method 140
4.5 Empirical power (%) when employing the latent variable method 141
4.6 Median sample sizes per group for two continuous, one ordinal and one binary measure using the latent variable method 144
4.7 Median sample sizes per group for two continuous, one ordinal and one binary measure using the standard binary method 145
4.8 Sample sizes from the latent variable method for the MUSE trial data . 149

5.1 Summary of the analysis methods recommended in a range of scenarios with different structures of composite endpoints 157

C.1 Median width of confidence intervals of the standard binary and augmented binary methods for the log-odds treatment effect 211
C.2 Median width of confidence intervals of the standard binary and augmented binary methods for the difference in response probabilities treatment effect 211
C.3 Average treatment effect in subsamples using the standard binary and augmented binary methods for the log-odds treatment effect 212
C.4 Average treatment effect in subsamples using the standard binary and augmented binary methods for the difference in response probabilities treatment effect 212
C.5 Average treatment effect in permuted subsamples using the standard binary and augmented binary methods for the log-odds treatment effect 213

C.6 Average treatment effect in permuted subsamples using the standard binary and augmented binary methods for the difference in response probabilities treatment effect 213
C.7 Type I error of the log-odds ACR50 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 214
C.8 Power of the log-odds ACR50 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 214
C.9 Type I error of the ACR50 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 215
C.10 Power of the ACR50 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 215
C.11 Type I error of the log-odds ACR70 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 216
C.12 Power of the log-odds ACR70 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 216
C.13 Type I error of the ACR70 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 217

C.14 Power of the ACR70 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment 217
C.15 Power and average confidence interval width in ACR20 response in the small sample adjusted standard binary and augmented binary methods in 5000 simulations 218
C.16 Type I error rate and average confidence interval width in ACR20 response in the small sample adjusted standard binary and augmented binary methods in 5000 simulations 219

Nomenclature

Acronyms

ACR American College of Rheumatology

AIC Akaike Information Criterion

ALP Alkaline Phosphatase

AugBin Augmented Binary Method

BILAG British Isles Lupus Assessment Group

CGCM Conditional Grouped Continuous Model

CGD Conditional Gaussian Distribution

CI Confidence Interval

COMET Core Outcome Measures in Effectiveness Trials

COS Core Outcome Sets

cv Critical Value

eGFR Estimated Glomerular Filtration Rate

EM Expectation Maximisation

EMA European Medicines Agency

EmpSE Empirical Standard Error

FDG-PET 18-Fluorodeoxyglucose Positron Emission Tomography

GEE Generalised Estimating Equations

GLS Generalised Least Squares

IFN Interferon

LatVar Latent Variable Method

MCMC Markov Chain Monte Carlo

MIBG Metaiodobenzylguanidine

MLE Maximum Likelihood Estimation

ModSE Model Standard Error

MSE Mean Squared Error

PFS Progression-Free Survival

PGA Physician's Global Assessment

PML Penalised Maximum Likelihood

RA Rheumatoid Arthritis

RCT Randomised Controlled Trial

RECIST Response Evaluation In Solid Tumours

SD Standard Deviation

SE Standard Error

SLE Systemic Lupus Erythematosus

SLEDAI Systemic Lupus Erythematosus Disease Activity Index

SRI Systemic Lupus Responder Index

ULN Upper Limits of Normal

UPC Urinary Protein to Creatinine

VAS Visual Analogue Scale

VGPR Very Good Partial Response

Notation

$''\delta$	Partial derivatives of δ with respect to each of the parameter estimates
α_F	Parameters for augmented binary continuous model
β_F	Parameters for augmented binary logistic regression at time 1
χ^2	Chi-squared distribution
δ	Treatment effect
η	Responder threshold
γ_F	Parameters for augmented binary logistic regression at time 2
κ	n_C/n_T
$\mu_{k\|l}$	conditional mean of outcome k given outcome l
μ_k	mean of outcome k
$\Phi_k(;\mu,\Sigma)$	k dimensional distribution function with mean μ and covariance matrix Σ
ψ_F	Parameters for standard binary logistic regression
ρ	Correlation parameter
σ	Variance parameter
τ	Thresholds for the latent space of a discrete outcome
θ	Model parameters
υ	Transformed parameters to be estimated
ε	Error term for observed outcomes
ε^*	Error term for latent outcomes
F	Binary failure indicator
$f_Y(y;)$	Probability density function of Y
H_0	Null hypothesis

H_1	Alternative hypothesis
I	Indicator function
k	Level in the binary variable
n	Number of patients per arm
n_{boot}	Number of bootstrap samples
n_C	Number of patients in the control group
N_{rep}	Number of re-sampled datasets
n_{sim}	Number of simulated datasets
n_T	Number of patients in the treatment group
p_C	Probability of response on control arm
p_T	Probability of response on treatment arm
r^p	Modified Pearson residuals
S	Overall response indicator
T	Treatment indicator
w	Level in the ordinal variable
Y^*	Latent outcome
Y_{cts}	Vector of observed continuous outcomes
Y_{dis}	Vector of observed discrete outcomes
Y_{ijk}	Outcome k for patient i, at time point j
z_α	$(1-\alpha)100^{th}$ standard normal percentile
N	Total number of patients

Chapter 1

Introduction

Clinical trials are studies which assess the impact of an intervention on the general health of a population of interest. Often this intervention is in the form of a drug treatment and is assessed by assigning eligible patients to receive either the experimental treatment or a control treatment, which may be a placebo or the current standard of care. The objective of a trial may be to assess the effectiveness of the proposed treatment, which tests that it provides benefit overall. Alternatively, it may be concerned with the efficacy which identifies the benefits in some identifiable subpopulation, such as those who adhere to the treatment protocol. Interventions tested in clinical trials proceed through four phases of testing. Phase I is concerned with collecting safety data by administering the treatment in a small number of volunteers to support further testing. The treatment is administered in a larger sample of patients in phase II, at a dose that was judged to be safe in phase I. The aim of the phase II study is to seek safety data and preliminary evidence of efficacy [1]. Phase III trials are large confirmatory studies which assess the effectiveness of the treatment in order to gain regulatory approval. The sample size chosen in the phase III study will be based on controlling the probability with which a real effect can be identified as statistically significant and estimating the treatment effect with high statistical precision [2]. Phase IV trials take place after marketing to collect data on rare but serious effects of the treatment that may not have been discovered in phase I-III.

A growing concern raised by the various stakeholders in trials is the duration and expense of the clinical trial process [3–5]. Studies requiring many patients to detect a given treatment effect have higher costs and take longer to get effective treatments to market than those requiring fewer patients. These additional costs accrued at the development stage result in higher drug prices meaning that society bears the burden

[6]. It is therefore the responsibility of researchers working on all aspects of clinical trials to improve efficiency wherever possible [7].

1.1 Composite Endpoints

Composite endpoints combine a number of individual outcomes in order to assess the effectiveness or efficacy of a treatment. They are typically used in situations where it is difficult to identify a single relevant endpoint to sufficiently capture the change in disease status incited by the treatment, however they may be employed for multiple purposes [8–11].

The construction of the composite endpoint differs depending on the disease. For instance, it is common in randomised trials of cardiovascular conditions to combine a number of binary outcomes such as death, myocardial infarction, stroke or ischemia-driven target vessel revascularization, as in [12]. These composite endpoints are typically analysed using time-to-event methods. Composite endpoints in other diseases combine outcomes on different scales, such as continuous and discrete measures. Our focus in this thesis is on a subset of these outcomes known as composite responder endpoints. These endpoints allocate patients as either 'responders' or 'non-responders' based on whether they cross predefined thresholds in the individual outcomes and are typically treated as a single binary endpoint. It is both theoretically and pragmatically important to make the distinction between composites that require patients to experience an event in all components and those which require patients to have an event in at least one of the components. We introduce the general characteristics of both below.

1.1.1 Events in at Least One Component

To be classed as a responder in some diseases, patients may have to meet one of multiple criteria, which may be defined on different scales. Alternatively, patients may have to respond in a subset of the components in order to be responders overall, such as in rheumatoid arthritis where response in five of seven components equates to response overall. Properties related to composite endpoints requiring events in at least one component have been discussed at length in the literature, e.g. [11]. The considerations in the construction of composite endpoints are summarised as follows.

1. Coherence

   Coherence in this context means that components should measure the same underlying pathophysiologic process, as well as the same disease process.

2. Coincidence

   Although composite endpoints should be coherent, coincidence ensures that components are not so closely related that patients experience all of them. In this case it is considered that the composite endpoint has become redundant and the effects of treatment can be captured in a single component.

3. Therapy homogeneity

   From an investigator's perspective it is important that the composite endpoint is sensitive to the treatment being evaluated and it is desirable that effect sizes are similar on each component.

A desirable aspect of these endpoints is that their application in trials may result in an increase in power, provided everything else remains constant. This is due to an increase in the number of events, where events are defined as any occurrence of response. Moyé [11] frames the possible power gains in terms of probability by assuming a two-dimensional composite endpoint with outcomes A and B, as shown in (1.1).

$$\begin{aligned} P(A \cup B) =& P(A) + P(B) - P(A \cap B) \\ =& P(A) + P(B) - P(A|B)P(B) \\ =& P(A) + P(B)(1 - P(A|B)) \end{aligned} \tag{1.1}$$

From this we can see that the event rate $P(A \cup B)$ is at its maximum when $P(A|B) = 0$, implying that mutual exclusivity of events in the composite is desirable. However, for two components to be mutually exclusive in practice often requires linking together events that physicians are unaccustomed to combining, leaving the interpretation of the endpoint challenging [9, 10]. Furthermore, this is simplistic as the magnitude of power gains may also depend on the treatment effect in each component as well as the occurrence of events, where therapy homogeneity across components is considered optimal in terms of power [11, 13].

1.1.2 Events in All Components

Patient response may otherwise be obtained through meeting specific criteria in all components of the outcome. As before, these endpoints must be both coherent and homogeneous in response to therapy. However, they are not designed to avoid coincidence. Examples of these endpoints arise in solid tumour cancers, where a patient is only classed as a responder if they have experienced a predefined reduction in tumour size and have not developed new lesions [14].
Generally these endpoints do not have the advantage of increasing power in a given study, as an increased number of events are required. Instead, they are useful in multisystem diseases which require interventions to treat a range of symptoms in order for the treatment to be considered truly effective. Considering the probability of the events in the composite occurring, we are now concerned with $P(A \cap B)$ which will be largest when $P(A \cup B)$ is minimised as shown in (1.2). However, as before other considerations such as treatment effect homogeneity are also relevant.

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \tag{1.2}$$

The methods in this thesis will be developed for composite responder endpoints in general, which may be applied where events are required in all components or in at least one component.

1.1.3 Opportunities and Limitations

There are many potential benefits to conducting a clinical trial using composite endpoints. As discussed, in the case of requiring response in at least one component, composite endpoints have the advantage that they increase the number of events in the trial [8, 15, 16]. In the likely case that this leads to a reduction in sample size, trials may be shorter and less expensive resulting in effective drugs being brought to market earlier [17–20]. Composite endpoints are particularly useful when a number of outcomes are equally relevant. In particular, in the case of diseases with large variation in symptoms, employing a composite endpoint will avoid an arbitrary choice of a single outcome [9, 15, 21, 22]. Furthermore, combining equally relevant outcomes and analysing as a composite endpoint negates the requirement for a multiple comparison adjustment [22–25]. In addition, proponents of composite endpoints believe that they are appropriate as they estimate the net clinical benefit of intervention by accounting

for the multiple factors of interest in a given disease [26–29].
However, there are limitations in the application of composite endpoints. In practice, composites may be inconsistently defined and provide opportunities for post-hoc changes [30]. Composite endpoints may be driven by less important or subjective components, meaning that a promising treatment effect may not translate to the expected benefit for patients [9, 22, 31]. If 'quantitative heterogeneity' occurs and treatment effects observed on the components are in different directions, this will make interpreting the overall effect challenging [10, 11, 15]. Furthermore, treatment effects on the overall composite may be diminished or harmful effects may be masked if unresponsive components are included [8, 10, 24]. Although composite endpoints have the capacity to capture multiple aspects of a disease, 'qualitative heterogeneity' means that not all patients will attach similar importance to each component [10, 18, 27–29]. Finally, it may not always be possible to avoid multiple testing corrections as many applications of composite endpoints require that the treatment effects on each individual component should also be reported [11, 15].

1.1.4 Recommendations for Use

When employing composite endpoints, guidance must be followed in order to ensure valid and meaningful implementation in clinical trials [8]. As the overall treatment effect reported on a composite endpoint depends on the correlation between components, the direction of treatment effect in each component and hence the patient responder rates, it is therefore important for interpretation that effects are reported on individual components as secondary results. In order to reduce any ambiguity in application, many sets of guidelines have been issued, including from the European Network for Health Technology Assessment (EUnetHTA) for application in pharmaceuticals [32]. We summarise the recommendations from the literature for construction, reporting and interpretation of composite endpoints below [13, 30].

A. Construction

- Composite endpoints should generally not be used if a suitable single endpoint is available, except when it can be justified to be more suitable (e.g. rare disease/event) [32]

- Composites and components should be clearly prespecified before starting the trial [9, 21]

- Prior evidence should exist for each component to avoid including clinically unimportant outcomes [19, 22]

- Including outcomes that are unlikely to experience an effect of the intervention should be avoided [11, 27]

- A mix of objective and subjective outcomes should be avoided [13, 31, 32]

B. Reporting

- Components should be separately defined as secondary endpoints and effects reported with the primary analysis results to determine if one component has dominated the composite [13, 18, 25]

- Separate components can be reported according to severity level and the 'worst' outcome experienced should be reported according to a predefined ranking system [25, 32]

- Report relevant combinations of the components relating to subgroups or special patient populations at risk [24]

- The number of patients with partially missing values on some components should be reported in detail [32]

C. Interpretation

- Treatment effects should be interpreted based on the composite endpoint (any effect of the components should be interpreted together rather than concluding efficacy of individual components) [9, 13]

- Clinically important components should be checked to ensure that they have not been affected negatively by the treatment [22, 24]

- Basing the overall conclusion on a meta-analysis if comparable composite endpoints are available from several studies should be considered [32]

1.2 Existing Methods

The work in this thesis will focus on methodology for composite responder endpoints with components defined on a mixture of discrete and continuous scales. We introduce the existing methods for this application below and highlight the need for further methods development in this area.

1.2.1 Notation

Let us initially consider the case of a composite endpoint comprised of a single continuous and single binary outcome measured at multiple time points. Suppose we have n patients per arm and $T_i \in \{1, 2\}$ indicates the treatment arm of patient i. Y_{ij} denotes the continuous score at time $j \in \{1, 2\}$ where y_{i0} is the baseline score and η is the continuous responder threshold. Let F_{ij} be an indicator variable taking a value equal to 1 if the patient fails to respond in the binary outcome at time point $j \in \{1, 2\}$, where F_{i2} is equal to 1 if patient i fails to respond any time between the first and second visit. S_i is a binary variable indicating whether or not patient i was a responder overall, where $S_i = 1$ if $Y_{i2} \geq \eta$ and $F_{i1} = F_{i2} = 0$.

1.2.2 Standard Binary Method

The method often employed to analyse these data in trials is a logistic regression on the binary indicator for response, S_i. We refer to this as the standard binary method, which is shown in (1.3).

$$logit\left(P\left(S_i = 1 | T_i, y_{i0}\right)\right) = \psi_{F0} + \psi_{F1} T_i + \psi_{F2} y_{i0} \tag{1.3}$$

This provides maximum likelihood estimates $\hat{\boldsymbol{\theta}} = \left(\hat{\psi}_{F0}, \hat{\psi}_{F1}, \hat{\psi}_{F2}\right)$ and $Cov(\hat{\boldsymbol{\theta}})$ which can be used directly to estimate the odds ratio and its confidence interval. We can also obtain predicted probabilities for each patient as if they were treated p_{iT} and not treated p_{iC}. This allows us to estimate the risk difference, risk ratio and odds ratio as shown in (1.4), (1.5) and (1.6) respectively.

1. Risk difference

$$\delta_1 = \frac{\sum_{i=1}^{N} p_{iT} - \sum_{i=1}^{N} p_{iC}}{N} \tag{1.4}$$

2. Risk ratio

$$\delta_2 = \frac{\sum_{i=1}^{N} p_{iT}}{\sum_{i=1}^{N} p_{iC}} \tag{1.5}$$

3. Odds ratio

$$\delta_3 = \frac{\left(\frac{\sum_{i=1}^{N} p_{iT}}{N - \sum_{i=1}^{N} p_{iT}}\right)}{\left(\frac{\sum_{i=1}^{N} p_{iC}}{N - \sum_{i=1}^{N} p_{iC}}\right)} \tag{1.6}$$

Confidence intervals for these treatment effect estimates can be constructed by obtaining standard error estimates using the delta method. This requires the covariance matrix of the maximum likelihood estimates $\text{Cov}(\hat{\boldsymbol{\theta}})$ and the vector of partial derivatives of δ with respect to each of the parameter estimates, $''\boldsymbol{\delta}$. For example, the variance of δ_1 is obtained as shown in (1.7).

$$Var(\delta_1) = (''\boldsymbol{\delta_1})^T Cov(\hat{\boldsymbol{\theta}})(''\boldsymbol{\delta_1}) \tag{1.7}$$

The advantage of the standard binary method is that the method is extremely straightforward and quick to implement. However, this is at the expense of losing a lot of information detailing how close each patient was to the responder threshold. Therefore, those who narrowly missed being classified as a responder are indistinguishable in the analysis from those who had measurements far from the threshold. As well as being inefficient, a problematic consequence of this is that the standard binary method is sensitive to misclassification due to measurement error. Measurement error in this instance means that a patient's continuous measurements could differ on two readings, hence patients who have measurements close to the response threshold could feasibly be classed as a responder under one reading and a non-responder with the other reading [33]. As the binary method does not distinguish between responders and non-responders of different magnitudes, we may be sceptical about any conclusions drawn if the dataset contains a large number of patients with measurements near the dichotomisation threshold. It is possible to do a sensitivity analysis in this case to determine whether classifying these patients differently would lead to different conclusions.

1.2.3 Suissa Method

It is also common in medical studies when working with a single continuous outcome of interest Y, that the clinically relevant event is a state of disease characterised by being above or below a given cut-off point η. An example of this is in reflux chest pain syndrome, where the reduction in chest pain must be greater than or equal to 50%. Often in practice, these outcomes are also considered to be binary and analysed using standard binary methods. Suissa [34] introduced an alternative method for estimating the risk of events defined by a sub-domain of continuous outcomes that does not require dichotomisation of the variable. This is based on assuming a Gaussian distribution such that $Y \sim N(\mu_T, \sigma_T^2)$ in the treatment group and $Y \sim N(\mu_C, \sigma_C^2)$ in the control group. If responders are defined as those who have Y values less than some value η then the event of interest is defined as $p_T = P(Y < \eta | T = 1)$ and $p_C = P(Y < \eta | T = 0)$ in the treatment and control arms respectively. The measurements of interest such as the risk difference, risk ratio and odds ratio are defined as (1.4), (1.5) and (1.6), as before.

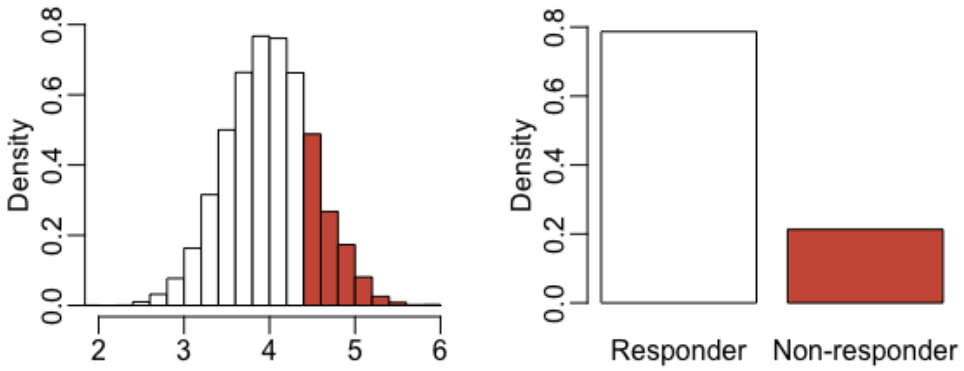


Figure 1.1: Graphical illustration of the Suissa method, which uses the Gaussian distribution underpinning the dichotomy to estimate the binary responder rate, where responders are defined as having a continuous measure less than 4.4

Figure 1.1 shows an illustration of the example used by Suissa, where $Y \sim N(4, 0.5^2)$. The response threshold is set at $\eta = 4.4$ meaning that any patients with Y scores greater than 4.4 are classed as non-responders, which is shown by the dichotomised response data in the bar plot. The findings showed that analysing the dichotomised data using the standard binary method is 33% less efficient than the method utilising the continuous information, as determined by the reduction in confidence interval width. Therefore, if a sample contained 100 patients, the same precision could be obtained in 67 patients by using the Suissa method.

1.2.4 Augmented Binary Method

The Suissa method is shown to improve efficiency and allieviate the problems with measurement error when the endpoint is a dichotomised continuous component. However, when the endpoint of interest is a combination of a continuous and binary outcome, there is no obvious joint distribution with which we can model the components. The augmented binary method is an extension of the Suissa method, proposed by Wason & Seaman [35] for a composite responder endpoint comprised of a dichotomised continuous component and some additional binary information. This improves efficiency by making use of how close patients were to being responders in the continuous component. For a fixed sample size, the method was shown to provide a substantial increase in power over the standard binary method currently in use, whilst still making inference on the outcome of interest to clinicians. This was illustrated in both solid tumour cancer and rheumatoid arthritis data [35, 36].
In the case of two time points, as considered by Wason & Seaman [35], a continuous component Y is measured at baseline and two follow up times and a binary component F is measured at two time points. The augmented binary method models the joint distribution of (Y_1, Y_2, F_1, F_2) by employing factorisation techniques to model each of the components separately, as shown by the equations below.

$$\begin{aligned} Y_{ij} &= \alpha_{F0}+\alpha_{F1}T_iI\{j=1\} + \alpha_{F2}T_iI\{j=2\} + \alpha_{F3}y_{i0} + \alpha_j + \varepsilon_{ij} \\ (\varepsilon_{i1}, \varepsilon_{i2}) &\sim N\left((0,0), \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \end{aligned} \tag{1.8}$$

$$logit\left(P(F_{i1}=1|T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \beta_{F0} + \beta_{F1}T_i + \beta_{F2}y_{i0} \tag{1.9}$$

$$logit\left(P(F_{i2}=1|F_{i1}=0, T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \gamma_{F0} + \gamma_{F1}T_i + \gamma_{F2}Y_{i1} \tag{1.10}$$

Note that in (1.8), α_{F1} and α_{F2} are the treatment effects at time point one and two respectively and (α_1, α_2) are the time effects. Equation (1.9) represents the probability of failure at time point 1 and (1.10) determines the probability of failure at time point 2, given the patient did not fail at time point 1. Repeated measures models can be fit to the continuous component using generalised least squares (GLS), which estimates the variance-covariance matrix using restricted maximum likelihood methods. After fitting these models and obtaining maximum likelihood estimates

$\hat{\boldsymbol{\theta}}_{AB} = \left(\hat{\alpha}_{F0}, \hat{\alpha}_{F1}, \hat{\alpha}_{F2}, \hat{\alpha}_{F3}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_{F0}, \hat{\beta}_{F1}, \hat{\beta}_{F2}, \hat{\gamma}_{F0}, \hat{\gamma}_{F1}, \hat{\gamma}_{F2}\right)$, we can obtain the overall probability of response in each arm. The probability of response in the endpoint which has dichotomisation threshold η_1 is shown below.

$$
\begin{aligned}
&P(Y_2 \geqslant \eta_1, F_1 = F_2 = 0 | T, y_0) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(Y_2 \geqslant \eta_1, F_1 = F_2 = 0 | T, y_0, Y_1 = y_1, Y_2 = y_2) f(y_1, y_2; T, y_0) d\boldsymbol{y} \\
&= \int_{-\infty}^{\infty} \int_{\eta_1}^{\infty} P(F_1 = F_2 = 0 | T, y_0, Y_1 = y_1, Y_2 = y_2) f(y_1, y_2; T, y_0) d\boldsymbol{y} \\
&= \int_{-\infty}^{\infty} \int_{\eta_1}^{\infty} P(F_2 = 0 | F_1 = 0, T, y_0, Y_1 = y_1) P(F_1 = 0 | T, y_0, Y_1 = y_1) f(y_1, y_2; T, y_0) d\boldsymbol{y}
\end{aligned}
$$

We can obtain a fitted probability of response for each patient i as if they were treated with the experimental treatment p_{iT} and the control treatment p_{iC}. Treatment effect estimates and confidence intervals are constructed as before, where $Cov(\hat{\boldsymbol{\theta}}_{AB})$ is as shown below.

$$
Cov(\hat{\boldsymbol{\theta}}_{AB}) = \begin{pmatrix}
Cov(\hat{\alpha}_{F0}, \hat{\alpha}_{F1}, \hat{\alpha}_{F2}, \hat{\alpha}_{F3}, \hat{\alpha}_1, \hat{\alpha}_2) & 0 & 0 \\
0 & Cov(\hat{\beta}_{F0}, \hat{\beta}_{F1}, \hat{\beta}_{F2}) & 0 \\
0 & 0 & Cov(\hat{\gamma}_{F0}, \hat{\gamma}_{F1}, \hat{\gamma}_{F2})
\end{pmatrix}
$$

Figure 1.2 is a schematic showing the stages involved in applying the standard binary and augmented binary methods. This clearly illustrates that the gains in efficiency in the augmented binary method arise from collapsing the observed data after the analysis, rather than before. Importantly, both models provide the same outcome of interest.

Number of Time Points

We have discussed the application of the methods for two time points however this is not a requirement. The augmented binary method may be easily employed to model one time point where the method reduces to one logistic regression model and a linear model for the continuous information. Furthermore, it may be used in situations with more than two time points however it may become too computationally demanding for a large number of follow-up times [37].

To make further improvements to the analysis of composite endpoints it will be important to consider the forms of composite endpoints which are most commonly employed in practice, which we consider in Section 1.3.

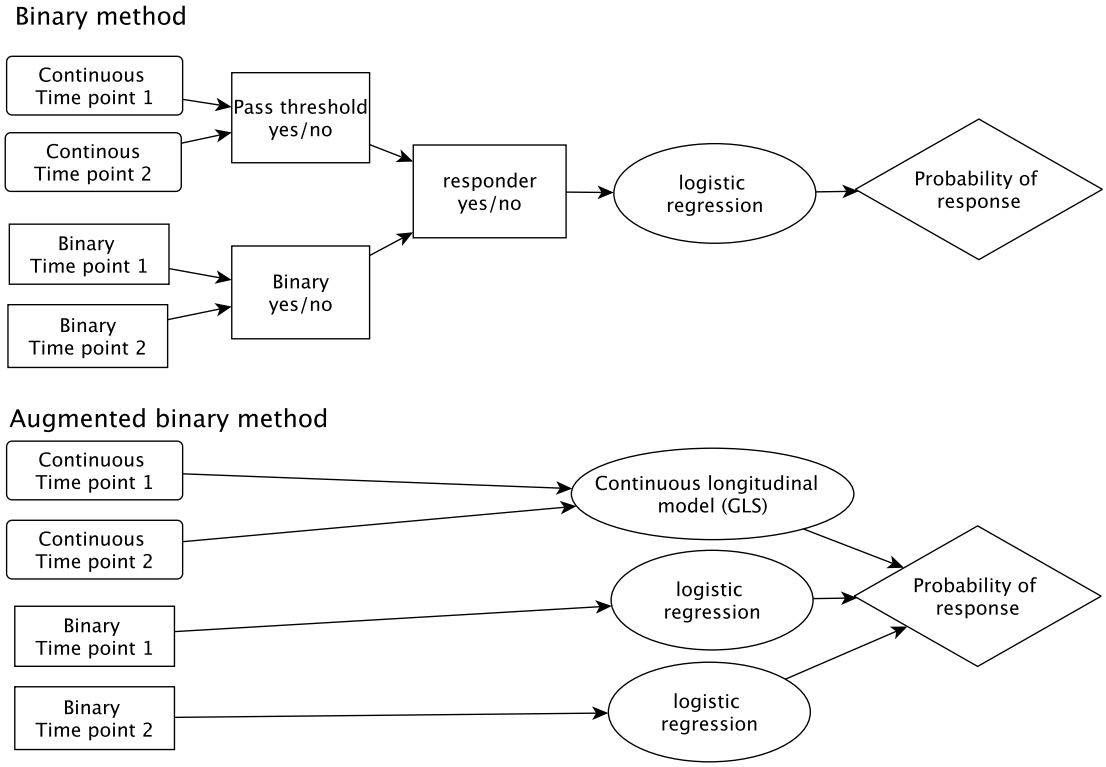


Figure 1.2: Schematic comparing the stages involved in fitting the standard binary and augmented binary methods for a composite endpoint with a continuous score and binary indicator measured at two time points

1.3 Scope for Application

1.3.1 Motivation

We have previously discussed the rationale for employing different forms of composite endpoints and highlighted the opportunities available when they are used correctly. Another important consideration for motivating further methods development in this area is to understand the range of clinical areas that commonly make use of mixed outcome composite endpoints. Previously identified clinical areas include solid tumour oncology and rheumatoid arthritis, both of which use a composite responder endpoint with a single dichotomised continuous component combined with additional binary information. The augmented binary method has previously been applied to data in both of these areas [35, 36]. Another endpoint which we have identified is in systemic lupus erythematosus (SLE), which contains multiple continuous and ordinal components and will be the motivation for Chapter 3. However, for the work to have the maximum potential impact we wish to identify additional disease areas which are making use of composite responder endpoints containing dichotomised continuous components.
By understanding the extent to which mixed outcome composites are being employed in trials and the most common forms that exist for these endpoints, we can ensure that the methods developed in this thesis are closely aligned with what is needed in clinical practice.

1.3.2 Methods

To answer this question, we make use of the COMET (Core Outcome Measures in Effectiveness Trials) database (http://www.comet-initiative.org/resources), which lists the minimum that should be measured and reported in all clinical trials of a specific condition. We reviewed physiological and mortality trial outcomes recorded within all core outcome sets (COS) that were published before 2016 and identified 287 in total. Each core outcome set paper was reviewed to determine if any responder endpoints were recommended for reporting in all clinical trials within that condition. In some cases, a potentially relevant endpoint was not clearly described in the core outcome paper. In this case, we examined randomised controlled trials (RCTs) that had used the endpoint to determine whether it was a suitable responder endpoint.

1.3.3 Findings

Through this process we identified 45 clinical areas (additional to solid tumour oncology, rheumatoid arthritis and SLE) where the augmented binary method could be utilised to gain efficiency [38]. An additional 23 clinical areas had used responder endpoints formed from a single categorised or dichotomised continuous variable, which could improve efficiency by making use of the Suissa method. These are shown in the Tables 1.1-1.3 below and a preprint of the work is included in Appendix A. From the results, we can see that composite responder endpoints with dichotomised continuous measures are used across a diverse set of conditions and thus the potential impact of further methods development is substantial. As we identified the diseases using the COMET database, this was not a systematic review and thus it is likely that our estimate of diseases using these endpoints is conservative.

1.4 Thesis Aims

We must note that objections to the formation and application of composite endpoints in general do exist based on the limitations presented. However, the focus of this thesis is on improving the analysis of existing and validated endpoints based on the information already collected and so the general drawbacks of composite endpoints will not be discussed further. The augmented binary method has improved the analysis of composite responder endpoints through reducing the required sample size by at least 35% for the same power using the available observed data. However, we have identified some limitations in the existing work that we will investigate in this thesis. We expect that our developments will have the potential for substantial impact due to the number and range of disease areas using these composite responder endpoints, as identified in our review.

Given that the augmented binary method offers large efficiency gains we are interested in the performance of the method in the rare disease trial setting, where these gains are most needed. As the method uses more parameters, some evidence has suggested that it may not perform as well in small samples. Chapter 2 focuses on the small sample performance of both the standard binary and augmented binary methods and evaluates the behaviour of the methods with small sample corrections. The aim of this work is to modify the method so that it can be applied in rare disease trials either to improve the precision of treatment effect estimates or to reduce the required sample size to something more achievable in a rare disease population.

Table 1.1: List of diseases using composite responder endpoints with at least one dichotomised continuous measure which can utilise the augmented binary method to improve efficiency. * denotes a single dichotomised variable which could use the Suissa method to improve efficiency

Disease category	Condition	Endpoint
Bleeding and transfusion	• Haemophilia	Stroke or new or enlarged cerebral infarct
	• Immune thrombocytopenic purpura	Complete response
	• Transfusion	1 hour CCL<7.5*
Cancer †	• Acute Myeloid Leukaemia	Response, time to relapse
	• Breast cancer related lymphedema	≥50% reduction in excess arm volume*
	• Dyspnea or Breathlessness in Palliative Care	Severe breathlessness (numerical rating scale ≥6)*
	• Fever and neutropenia	Fever*
	• Hodgkin's disease and lymphoma	Complete response
	• Malignant lymphoma	Progression-free survival (PFS)
	• Myeloma (Newly diagnosed)	PFS and Very Good Partial Response (VGPR)
	• Myeloma (Refractory)	Complete response
	• Myeloma (Relapsed)	PFS
	• Systematic light-chain amyloidosis	Haematological response
Dentistry and vision	• Intermittent exotropia	Alignment, Deterioration
	• Edentulous	Implant success
	• Periodontal disease	Mobility grade 1*
Gastroenterology	• Hepatic encephalopathy	NAS improvement of 2 without worsening of fibrosis
	• Inflammatory Bowel Disease	Clinical remission
	• Non-alcoholic steatohepatitis	Resolution of Steatohepatitis without fibrosis
	• Reflux chest pain syndrome	Reduction in chest pain ≥50%*
	• Reflux oesophagitis syndrome	Troublesome regurgitation*

† Excluding solid tumour oncology

Table 1.2: List of diseases using composite responder endpoints with at least one dichotomised continuous measure which can utilise the augmented binary method to improve efficiency (continued). * denotes a single dichotomised variable which could use the Suissa method to improve efficiency

Disease category	Condition	Endpoint
Cardiovascular	• Aortic dissection	Procedural success
	• Aortic valve implantation	Clinical efficacy
	• Aortic valve stenosis	Device success
	• Atrial fibrillation	AF control*
	• Chronic leg admea	Normal range of motion*
	• Deep venous thrombosis and pulmonary embolism	Major bleeding
	• Head and neck lymphatic malformation	Response*
	• Mitrial regurgitation	Mitrial regurgitation
Infectious disease	• Influenza	Resolution of fever
	• Intra-abdominal infection	Recovery
	• Pneumonia	Clinical stability
Urological	• Acute kidney injury	Proportion with acute kidney injury
	• Acute renal failure	Proportion with acute renal failure
	• Male sexual dysfunction	Severe dysfunction*
Lungs and airways	• Connective tissue disease associated lung disease	Decline in forced vital capacity*
	• Idiopathic pulmonary fibrosis	Decline in forced vital capacity*
Neurology	• Cerebral Palsy	Modified Teacher's Drooling scale*
	• Chronic Demyelinating Polyradiculoneuropath	Impairment
	• Headache	>50% improvement in HA index
	• Hypoxic-ischemic brain injury	Moderate-severe disability
	• Intracranial cerebral atherosclerosis	Acute and subacute arterial occlusions
	• Multifocal Motor Neuropathy	INCAT Sensory Sum Score
	• Multiple Sclerosis	Progression of disability
	• Pain	>30% reduction in pain scale*
	• Traumatic brain injury	Severe disability rating*

Table 1.3: List of diseases using composite responder endpoints with at least one dichotomised continuous measure which can utilise the augmented binary method to improve efficiency (continued). * denotes a single dichotomised variable which could use the Suissa method to improve efficiency

Disease category	Condition	Endpoint
Mental health and addiction	• Alcohol abuse	Proportion heavy drinking days
	• Bipolar disorder	Children's depression rating*
	• Major depressive disorder	Response
	• Nicotine abuse	Abstinence
Orthopaedics and trauma	• ACL injury	Knee function
	• Burns	Response
	• Dupuytren's disease	Contracture recurrence
	• Low back pain	Severe disability*
Rheumatology[†]	• Acute gout	Patients with sUA level <6.0mg*
	• Ankylosing spondylitis	ASAS20 response
	• Idiopathic arthritis-associated uveitis	Best corrected visual acuity and no light perception
	• Juvenile arthritis	Response
	• Juvenile dermatomyositis	Responder index
	• Prevention of fracture in high risk populations	Response
	• Proliferative and membranous lupus renal disease	Urinary protein levels within normal range*
	• Sarcopenia prevention	Occurrence of sarcopenia
	• Sjogren's syndrome	>30% reduction in three analog scales
	• Systemic Sclerosis	SCP in normal range, no renal crisis
	• Vasculitis disorders	Response/partial improvement*
Other	• Endometriosis-related pain	>30% reduction in symptom score without use of rescue analgesics
	• Gestational diabetes mellitus	Gestational hypertension
	• Neurofibromatosis	Severe pain*

[†] Excluding rheumatoid arthritis and systemic lupus erythematosus

The augmented binary method performs well in the case of one continuous and one binary component. It can also be applied to more complex composite endpoints that contain more components by combining any remaining information with the binary indicator. However, in the case that the composite contains multiple continuous or ordinal components, this will still result in a loss of information due to the continuous and ordinal measures being treated as binary. We hypothesise that if we could retain the information in multiple continuous and ordinal components then we may have even larger efficiency gains. The aim of Chapter 3 is to develop methodology for more complex composite endpoints and assess the performance through simulation and application to a phase IIb trial in SLE.

Although a joint modelling approach to composite endpoint analysis has proved promising, one restriction for employing the methods in the primary analysis is the absence of a technique to calculate the sample size. The work in Chapter 4 aims to develop a method for sample size estimation and to investigate how the structure of the data, such as the correlation between components and the treatment effect structure within the components, affects the sample size required.

Finally, Chapter 5 will focus on a discussion of the work along with the limitations and areas for future work to further improve the analysis of composite endpoints in practice.

Chapter 2

Composite Endpoints in Rare Disease Trials

2.1 Introduction

2.1.1 Motivation

For stakeholders in rare disease communities, it is imperative to keep in mind that rare diseases are far from 'rare' for those whose lives they consume, with the patient experience for individuals suffering from a rare condition being extremely challenging at every stage. The diagnosis process is typically much longer than for patients suffering from more common diseases due to doctors being unfamiliar with the illness and its manifestations. Once the disease is identified, patients are often given the choice of no treatment options or extremely toxic experimental treatments. Patients may therefore have to enrol in clinical trials to receive any treatment at all for their condition [39–41]. The last few decades have seen a societal shift which recognises some of these issues and has resulted in a much greater focus on rare disease research. Some of these shortfalls have been addressed by a surge in patient advocacy groups, that aim to increase awareness of the disease and lobby the government to increase funding of rare disease research. Furthermore, they undertake activities such as facilitating patient registries and disease natural histories in order to 'de-risk' the drug development process and increase access to treatments for rare disease patients [42]. The targeted agenda of these organisations alongside advances in technologies, which improve international communication between rare disease experts and patients, has resulted in an improvement in education surrounding these diseases and hence the abilty to act

quicker in the diagnosis and treatment phases [39]. However, the reality is that there is still a lot of progress to be made.

A crucial factor inhibiting progress is that pharmaceutical companies have historically neglected the rare disease sector, due to the large costs involved in product development for what will ultimately be used by a relatively small group of patients. To address this issue the European Medicines Agency (EMA) introduced incentives for drug companies that aim to make the rare disease research market more attractive and lucrative. One aspect of this initiative is market exclusivity for orphan drugs, offering ten years of protection from market competition, allowing companies to recover some of the large development costs. Other incentives include protocol assistance, fee reductions and grants [43].

Another restrictive characteristic of rare disease trials is slow recruitment due to a small available population. Standard methods to alleviate these problems involve running multi-centre, international clinical trials. However, this is challenging due to there being no formal standard of care for many rare conditions and the often varying definitions of disease in different countries meaning the control arm in an international trial may be highly heterogeneous. Consequently, achieving the desired sample size for rare disease trials requires alternative approaches. The issues arising from high variability in disease definition and care are exacerbated by the fact that rare diseases typically have large variation in disease manifestation across patients, making running a coherent trial with appropriate power extremely challenging. Composite endpoints are often recommended to address some of these issues. If the components are appropriately chosen, endpoints that require an event in only one of the components may have the ability to improve the power to show a given treatment effect due to the increased number of events [8–10].

These endpoints frequently feature in rare autoimmune diseases and rare cancers. Examples of these are presented in Table 2.1, one of which is the chronic inflammatory disorder Behçet disease. A review of the research performed in this area concludes that evidence continues to be based on anecdotal case reports rather than randomised trials [44]. As well as those shown in Table 2.1, any rare cancers using RECIST criteria (Response Evaluation Criteria In Solid Tumors) to define responders and non-responders use endpoints that assume this structure [14]. As discussed in the introduction, methods currently employed to analyse these endpoints are inefficient and waste a lot of valuable patient information. Due to the additional variation typically

present in rare disease trials, novel statistical design and analysis methods are especially necessary to make the best use of available information [40, 41].

Rare disease trials employing composite responder endpoints with continuous and binary components, such as those shown in Table 2.1, may make use of the augmented binary method, which was shown in more common diseases to reduce the required sample size by 35% for a fixed power. Clearly, a precision gain of this magnitude would be hugely beneficial in a rare disease trial however the method uses more parameters than the standard binary method. Some evidence has suggested that it may not be suitable for trials with small samples, perhaps due to issues with asymptotics [35]. The objective of the work in this chapter is to evaluate the performance of the augmented binary method in small sample settings, modify it using small sample corrections if necessary and explore the feasibility of using this to reduce the required sample size needed for enrolment in a rare disease trial.

If the gains provided by the augmented binary method in common diseases can be realised in smaller samples, we envisage that impact may be obtained in multiple ways. Firstly, it may allow us to gain information from randomised trials that would otherwise not be possible due to the perceived infeasibility of reaching the target sample size. Secondly, trials that already take place may report the treatment effect more precisely, meaning that fewer rare disease trials will conclude with a confidence interval so large that it leaves us uncertain about the utility of the treatment. Otherwise, the efficiency gains could be used to recruit fewer patients resulting in shorter trials and reduced costs for pharmaceutical companies and hence increasing innovation in this area. Furthermore, speeding up the drug development process can benefit patients as efficacious drugs may be brought to market and thus accessed sooner.

The initiatives and incentives currently in place have already dramatically improved how research is conducted in rare diseases. If statisticians can join other stakeholders in using their skills to tackle specific challenges in the rare disease sector, the research landscape in this field may be further improved for previously neglected rare disease populations.

2.1.2 Aims

To achieve the goals and impact discussed, we set a number of specific aims for the work in this chapter. As discussed previously, the motivation is based entirely on application in rare disease trials using composite responder endpoints and is therefore

Table 2.1: Examples of rare disease clinical trials which use a composite responder endpoint comprising continuous and binary measures to determine the effectiveness of a treatment, where patients must respond in all components by meeting a predefined responder threshold in order to be classed as a responder overall

Disease	Example responder endpoint
Primary biliary cholongitis (PBC)	• ALP$<1.67\times$ULN • Total bilirubin $<$ ULN • ALP decrease $\geq 15\%$
Behçets disease	• Length of principal intestinal ulcer compared to size at baseline (%) • No new lesions
Lupus nephritis	• eGFR no more than 10% below preflare value • Proteinuria UPC ratio < 0.5 • Urine sediment: Inactive • No rescue therapy
Neuroblastoma	• $<$10mm residual soft tissue at primary site • Complete resolution of MIBG of FDG-PET uptake (for MIBG non avid tumours) at primary site
Advanced hepatocellular carcinoma	• $<$20% increase in the sum of the longest diameters of target lesions • No new lesions

ALP alkaline phosphatase, ULN upper limits of normal, eGFR estimated glomerular filtration rate, UPC urinary protein to creatinine, MIBG metaiodobenzylguanidine, FDG-PET 18-fluorodeoxyglucose positron emission tomography

focused on determining pragmatic approaches to data analysis using the information available. Therefore, the objectives are to:

- Understand the performance of the standard binary and augmented binary methods in small sample sizes
- Determine the differences in performance when using generalised least squares (GLS) and generalised estimating equations (GEE) for modelling the continuous component in the augmented binary method
- Identify the most appropriate way to express the treatment effect estimate based on performance characteristics (risk difference vs. odds ratio)
- Identify and implement appropriate small sample corrections and compare the performance with the uncorrected methods
- Determine the most efficient analysis methods to reduce the required sample size in a rare disease trial or to report the treatment effect more precisely
- Make analysis recommendations for trials in rare diseases using composite endpoints

The chapter proceeds as follows. We introduce the methods and data that we will use to investigate the performance through re-sampling. We show the behaviour of the methods for varying sample size and response thresholds. We include a simulation study to verify the findings of the re-sampling and conclude with a discussion and recommendations for analysing rare disease trials using these endpoints.

2.2 Small Sample Adjustments

2.2.1 Binary Component Adjustment

Albert and Anderson show when fitting a logistic regression model to small samples that, although the likelihood converges, at least one parameter estimate may be theoretically infinite [45]. This phenomenon is commonly termed 'perfect separation' and occurs if the model can perfectly predict the response or if there are more parameters in the model than can be estimated because the data are sparse [46]. Firth provides an alternative to maximum likelihood estimation (MLE) in these circumstances [47].

This involves using penalised maximum likelihood (PML) to correct the mechanism producing the estimate, namely the score equation, rather than the estimate itself. More generally, penalised maximum likelihood can be thought of as a technique to introduce a small amount of bias in the parameter estimates in order to circumvent problems in the stability of parameter estimates that arise when the likelihood is relatively flat [48].
The penalised likelihood is shown below in equation (2.1), where $L(\boldsymbol{\theta})$ is the usual likelihood function for a logit model and $I(\boldsymbol{\theta})$ is the information matrix.

$$L^*(\boldsymbol{\theta}) = L(\boldsymbol{\theta})|I(\boldsymbol{\theta})|^{\frac{1}{2}} \tag{2.1}$$

As maximum likelihood estimates are always biased away from zero for logisitic regression in small samples, bias correction therefore involves some degree of shrinkage of the estimate towards this point [47]. This results in the method also reducing the variance, so that bias reduction does not necessarily lead to a substantial loss in power. This is an important and attractive property of the Firth correction, as a trade-off between bias and variance usually exists.
We can intuitively understand why the correction results in variance reduction. In this setting, the corrected estimate is always closer to zero, therefore it must always have a reduced variance due to being bounded between zero and the uncorrected estimate. The source of the bias is curvature in the score function $s(\boldsymbol{y}, \boldsymbol{\theta})$, meaning that if the score function is decreasing and curved in the area around the true parameter θ_{true} then a high miss $s(\boldsymbol{y}, \boldsymbol{\theta}) > 0$ implies an estimate well above the true value, so that $\theta_{mle} >> \theta_{true}$ and a low miss $s(\boldsymbol{y}, \boldsymbol{\theta}) < 0$ implies an estimate only slightly below the true value, so that $\theta_{mle} < \theta_{true}$. This implies that low misses and high misses do not cancel and that the MLE is too large on average. This is discussed further in [49].
We will use the Firth correction in both the standard binary method and the multiple logistic regression models in the augmented binary method. The modified estimator can be easily implemented in R using the brglm package [50], which provides the penalised likelihood estimates.

2.2.2 Continuous Component Adjustment for GEE

For continuous longitudinal data, there are a number of estimators providing estimates $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. The generalised least squares (GLS) estimator is typically used in linear mixed models and as it is likelihood based is a consistent and asymptotically unbiased

estimator of $\boldsymbol{\theta}$. The validity of the inference from the GLS estimator is dependent on correctly specifying the subject mean μ_{ij} and variance $\boldsymbol{V_i}$. The mixed model allows both marginal and subject specific inference. Generalised estimating equations (GEE) allow only for marginal inference. However, valid inference is possible from $\hat{\boldsymbol{\theta}}$ in the GEE method if μ_{ij} is correctly specified, even if $\boldsymbol{V_i}$ is misspecified [51]. We are interested in determining whether the estimator used for the continuous component has a substantial effect on model performance. In larger samples we may expect the differences between GLS and GEE to be negligible, however in small samples the estimation method may be more influential.

GEE is typically considered to be a more robust method for model misspecification, particularly as the variance estimator that is commonly used provides robust standard error estimates. This robustness property could be desirable in this setting, where model misspecification may be more problematic for the variance estimates than in a larger sample. However when using these methods where the number of patients is small, the robust standard error estimates are subject to downward bias leading to inflated type I errors [52]. The standard robust sandwich covariance estimator is shown in equation (2.2).

$$V_{sand} = (\textstyle\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i)^{-1} (\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \widehat{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i) (\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i)^{-1} \quad (2.2)$$

where:

$\boldsymbol{D}_i = \frac{\partial \mu_i}{\partial \beta}$

$\boldsymbol{\mu}_i$ is the vector of mean responses

$\boldsymbol{\beta}$ is the parameter vector

$\boldsymbol{V}_i$ is the working variance-covariance matrix for $\boldsymbol{Y}_i$

$\widehat{Cov}(\boldsymbol{Y}_i) = (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)'$

To address the limitations of this estimator in small samples, Morel, Bokossa and Neerchal [53] propose a correction which inflates the variance estimate. The small sample adjusted variance estimator V_{MBN} is shown below.

$$V_{MBN} =$$
$$\left(\textstyle\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \left(k\widehat{Cov}(\boldsymbol{Y}_i) + \delta_m \xi \boldsymbol{V}_i\right) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right) \left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}$$

where:

$k = \frac{N-1}{N-p}\frac{n}{n-1}$

p is the number of parameters

N is the total number of observations

n is the number of patients

$$\delta_m = \begin{cases} \frac{p}{n-p}, & \text{if n} > 3\text{p} \\ \frac{1}{2}, & \text{otherwise} \end{cases}$$

$$\xi = max\left(1, \frac{trace\left(\left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} Cov(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)\right)}{p}\right)$$

An appealing property of this estimator is that as the sample size increases, $k \to 1$ and $\delta_m \to 0$, so that $V_{MBN} \to V$. Note in ξ that the sum of the eigenvalues is used. These eignevalues may also be referred to as 'generalised design effects' [54]. Alternative corrections may include a different function of the eigenvalues, such as the maximum or the product, which corresponds to the determinant of $\left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} Cov(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)$. However in this case we employ the trace, as demonstrated by Morel et al. [53]. We implement this variance correction in R using a modification of the code provided in the geesmv package [55] when using the GEE estimator for the continuous component.

2.3 Assessing Properties: Re-sampling

2.3.1 Data

In order to determine the performance of the methods we will use data from the OSKIRA-1 trial [56]. The trial was a phase III, multi-centre, randomised, double-blind, placebo-controlled, parallel-group study investigating the use of fostamatinib in patients with active rheumatoid arthritis. A common responder endpoint used in rheumatoid arthritis is the ACR20, in which patients demonstrate clinical response if they achieve a 20% improvement from baseline, as measured by a continuous ACR (American College of Rheumatology) score. It is worth noting that the ACR score is a percentage change from baseline which is itself a composite combining seven components. In what follows we will treat this as a single measure, as is the case in practice.

A benefit of responder analyses is that we can easily incorporate additional information in the response definition. In the case of rheumatoid arthritis it is common to assign

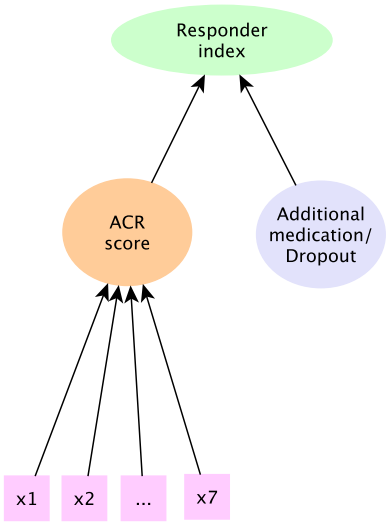


Figure 2.1: Structure of the composite responder endpoint used in rheumatoid arthritis which combines a continuous ACR score which is dichotomised at a predefined threshold and a binary indicator for additional medication use. These components form the overall responder index, where patients must respond in both components to be considered a responder overall

patients to being non-responders in the ACR20 endpoint if they require medications restricted by the protocol, or withdraw from the study. Therefore, in order to be a responder to treatment a patient needs to tolerate treatment, must not receive restricted medications and must demonstrate clinical response. This non-responder imputation allows discontinuations of treatment for lack of efficacy or for adverse events to provide meaningful information on the drug effect and translates to estimating the effect of a combination of continuous and binary components. The structure of the composite responder index is shown in Figure 2.1.

Other endpoints of interest in rheumatoid arthritis are the ACR50 and ACR70 which dichotomise the ACR score at 50% and 70% respectively. Although ACR20 was the primary endpoint in the trial and is the endpoint that is generally used to formally evaluate benefit in the regulatory setting, results for both the ACR50 and ACR70 endpoints are also discussed. These endpoints further characterise the benefit of a treatment by considering different levels of improvement from baseline. Furthermore, these endpoints will demonstrate how the methods perform with differing response rates which we anticipate could substantially alter model performance, particularly in small samples.

2.3.2 Model Notation

We apply the standard binary and augmented binary models defined in Chapter 1. The baseline ACR score for patient i is y_{i0}, with Y_{i1}, Y_{i2} denoting the continuous ACR scores at the week 12 visit and week 24 visit respectively. F_{i1} is an indicator variable taking a value equal to 1 if the patient discontinues treatment or requires rescue medication before the week 12 visit. F_{i2} is the corresponding indicator for the period between the week 12 and week 24 visit. S_i is a binary variable indicating whether or not patient i was a responder. For the ACR20 endpoint, $S_i = 1$ if $Y_{i2} \geqslant 20$ and $F_{i1} = F_{i2} = 0$.

2.3.3 Re-sampling

As an alternative to simulating data from a specified data generating model, we re-sample from the OSKIRA-1 trial. The re-sampling technique involves randomly sampling N observations from a dataset without replacement in order to create a new dataset. This process is repeated N_{rep} times to obtain the desired number of replications. Note that N_{rep} represents the number of re-sampled datasets and so is analogous to the number of simulated datasets N_{sim} in a simulation study. For the purpose of investigating the small sample properties of the methods, we will make use of two of the three arms in the trial, namely the fostamatinib 100mg bid for 52 wks arm and the placebo arm. Furthermore, we obtain samples from these arms rather than using the full sample size to mimic a rare disease scenario.

Acquiring data in this way in order to compare method performance is desirable for a number of reasons. Characteristics of real data that are not known, such as correlation structure and missing data, are present in the replicated datasets meaning that conclusions drawn from the findings may be more applicable to real diseases than those from a simulation model. Furthermore, if we find that the augmented binary method performs well under re-sampling where the true data generating model is unknown, then it could indicate robustness to model misspecification.

To assess and compare the model's performance, we re-sample 5000 replicates, which gives a Monte Carlo standard error of 0.3%, for each total sample size between 30 and 80 in increments of 10. To ensure balance we randomly sample half of the total sample size we are interested in from the placebo arm and the other half from the 100mg arm of the trial. We apply all methods to each sub-sample and record the treatment effect and 95% confidence interval. We do this for both the risk difference and log-odds estimates of the treatment effect. An estimate of the power is the

Table 2.2: Unadjusted and small sample adjusted methods to be compared through re-sampling to assess suitability for application in rare disease trials using composite responder endpoints with one continuous and one binary component

Methods	Unadjusted	Adjusted	
		Firth	MBN
Standard binary	X		
Augmented binary GLS	X		
Augmented binary GEE	X		
Standard binary adj		X	
Augmented binary GLS adj		X	
Augmented binary GEE adj		X	X

proportion of confidence intervals from the 5000 sub-samples that do not contain zero. By re-sampling, rather than simulating from a known distribution, thinking of this quantity as power implicitly assumes the treatment effect in the trial to be the true treatment effect in the population. This is similar to assuming that the data generating model has the true structure when using simulation methods. To determine an estimate for the type I error rate, we permute the treatment labels in each of the sub-samples in order to remove the association between treatment and outcome. An estimate of the type I error rate is the proportion of confidence intervals that do not contain zero. The coverage is estimated as the complement of the type I error rate. The median width of the confidence intervals and the average treatment effect for both methods are also presented.

Table 2.2 details the methods to be compared along with the corrections that will be implemented. Note that the adjusted augmented binary GEE method has two small sample corrections whereas the other methods have only one, so we may expect to see the largest modification in performance for this method.

2.3.4 OSKIRA-1 data

The OSKIRA-1 study was a multicentre, randomised, double-blind, placebo-controlled (for 24 weeks) parallel-group study to investigate the efficacy and safety of fostamatinib in patients with rheumatoid arthritis [56]. The study involved 141 centres in 17 countries, where 918 patients were randomised (1:1:1) to receive fostamatinib 100 mg twice daily, fostamatinib 100 mg twice daily for 4 weeks and then 150 mg once daily, or placebo, on a background of MTX treatment. The trial was blinded for 52 weeks,

with placebo patients switching to fostamatinib treatment at week 24 or at week 12 if requiring early rescue. For the purposes of this work, we consider only the period up to the primary end point at week 24, as was the case in [36].
For the ACR20 endpoint the risk difference was 0.13 with a 95% confidence interval of (0.05, 0.21). The corresponding values for the ACR50 and ACR70 endpoints were 0.15 (0.09, 0.20) and 0.08 (0.04, 0.11) respectively.

2.3.5 Results

The results for the ACR20, ACR50 and ACR70 endpoints on both the odds ratio and risk difference scales are detailed below.

2.3.5.1 ACR20

Odds Ratio

The power for the unadjusted and adjusted methods for the log-odds treatment effect are shown in Figure 2.2. The unadjusted augmented binary method provides higher power than the standard binary method for all sample sizes. The highest gains in power from the augmented binary method are achieved when the total sample size is 80, with power approximately equal to 50% for the augmented method and 23% for the standard analysis. In terms of power, the performance of GEE and GLS are very similar.

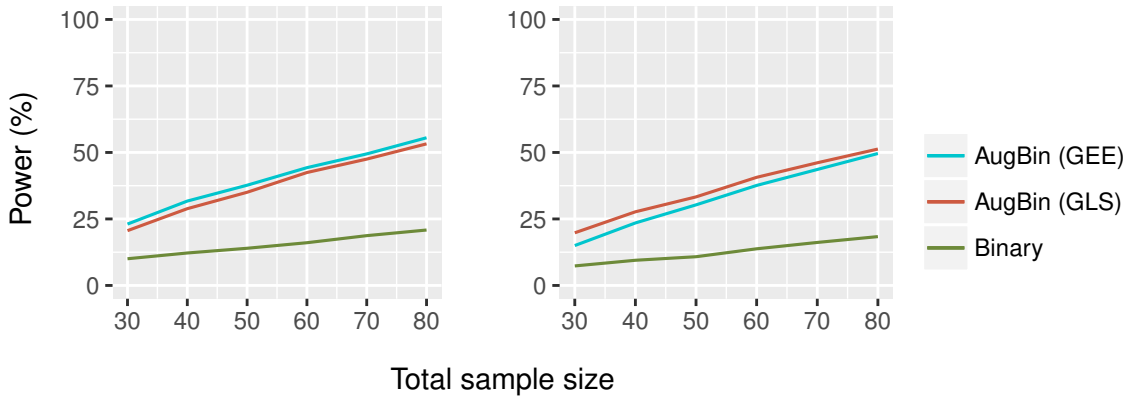


Figure 2.2: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the log-odds treatment effect estimate

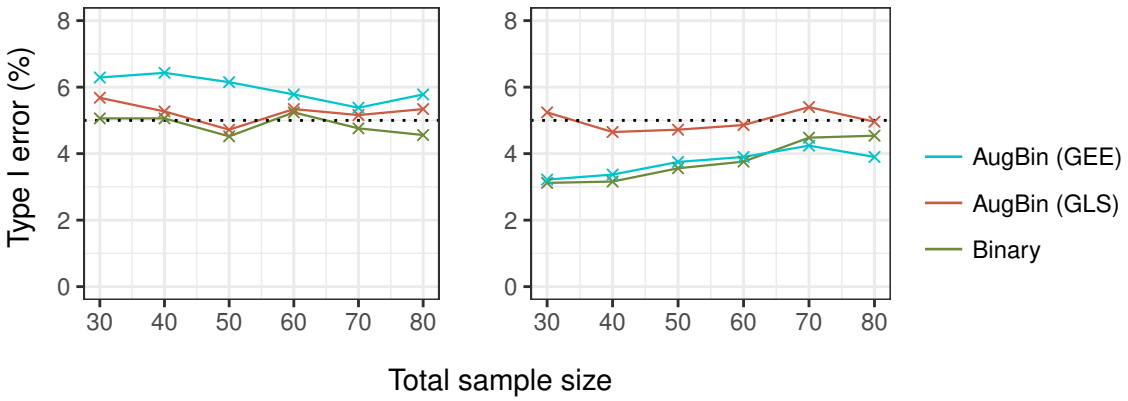


Figure 2.3: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the log-odds treatment effect estimate

The type I error rate of the unadjusted standard binary and augmented binary (GLS) methods are approximately 5% when reporting the log-odds treatment effect, as shown in Figure 2.3. The type I error rate of the augmented binary method (GEE) is 6%. Implementing the Firth adjustment in the augmented binary method with GLS makes negligible difference to the type I error rate. In the adjusted augmented binary method with GEE, the type I error rate drops to 3-4%. Differences between the GLS and GEE estimators diminish with increasing sample size. The small sample adjusted standard method also has type I error rate of 3-4%.

Table 2.3 shows the average treatment effect estimates from the unadjusted and adjusted methods in the null case. The Firth adjustment is useful for correcting the treatment effect estimate in the binary case when the total sample size is less than 50 and reduces the variance of the treatment effect estimates for all three methods. Table 2.3 also shows the estimated average treatment effect from the methods when the intervention has an effect. The findings are similar to the null case with a reduction in variance of the reported treatment effect from all methods. Note that the robust standard errors for the GEE method are larger than the conventional standard errors for the GLS method, as we would expect. However, when the GEE standard errors are adjusted for small samples they are reduced, despite being subject to downward bias in this setting. This is likely due to the additional Firth correction in this case which reduces the variance across all methods.

Table 2.4 contains the median confidence interval width and standard deviation for

Table 2.3: Average log-odds treatment effect estimates and standard deviation (SD) when the intervention does and does not have an effect from the unadjusted and small sample adjusted standard binary and augmented binary methods

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
Null						
30	0.011	0.005	0.001	0.001	-0.002	-0.002
	(1.171)	(0.742)	(0.665)	(0.641)	(0.673)	(0.646)
40	0.012	0.006	-0.005	0.000	0.014	0.002
	(0.755)	(0.645)	(0.559)	(0.537)	(0.583)	(0.555)
50	0.000	0.004	0.007	-0.003	-0.003	0.012
	(0.608)	(0.575)	(0.487)	(0.475)	(0.512)	(0.492)
60	-0.005	0.004	-0.003	0.000	-0.004	-0.005
	(0.557)	(0.527)	(0.449)	(0.437)	(0.461)	(0.452)
70	-0.004	0.001	0.004	0.000	-0.004	-0.007
	(0.507)	(0.411)	(0.413)	(0.487)	(0.424)	(0.417)
80	0.005	-0.009	0.005	-0.002	0.000	-0.001
	(0.467)	(0.457)	(0.390)	(0.354)	(0.401)	(0.386)
Effect						
30	0.606	0.526	0.783	0.747	0.812	0.773
	(1.096)	(0.737)	(0.649)	(0.619)	(0.678)	(0.646)
40	0.595	0.543	0.794	0.765	0.828	0.796
	(0.745)	(0.634)	(0.544)	(0.525)	(0.565)	(0.545)
50	0.572	0.536	0.790	0.767	0.821	0.795
	(0.587)	(0.547)	(0.478)	(0.465)	(0.495)	(0.480)
60	0.570	0.540	0.788	0.770	0.816	0.795
	(0.541)	(0.510)	(0.435)	(0.425)	(0.449)	(0.438)
70	0.577	0.551	0.794	0.902	0.821	0.802
	(0.476)	(0.453)	(0.394)	(0.456)	(0.406)	(0.398)
80	0.568	0.546	0.790	0.707	0.817	0.801
	(0.455)	(0.436)	(0.367)	(0.333)	(0.377)	(0.370)

Table 2.4: Median width of confidence interval with standard deviation in the parenthesis (SD) for the log-odds treatment effect from the unadjusted and small sample adjusted standard binary and augmented binary methods

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	3.019	3.000	2.509	2.477	2.458	2.739
	(431.8)	(0.236)	(1390.3)	(0.212)	(1413.6)	(0.384)
40	2.602	2.592	2.155	2.145	2.134	2.325
	(136.7)	(0.163)	(170.4)	(0.150)	(183.2)	(0.175)
50	2.320	2.314	1.924	1.919	1.916	2.053
	(0.135)	(0.112)	(61.746)	(0.117)	(81.195)	(0.123)
60	2.117	2.112	1.755	1.753	1.755	1.861
	(0.103)	(0.089)	(16.656)	(0.096)	(22.069)	(0.099)
70	1.959	1.954	1.624	1.862	1.630	1.712
	(0.081)	(0.072)	(0.588)	(0.218)	(1.318)	(0.081)
80	1.832	1.828	1.521	1.378	1.531	1.598
	(0.069)	(0.063)	(0.071)	(0.129)	(0.066)	(0.069)

the treatment effects from the unadjusted and adjusted methods. The confidence interval widths are substantially smaller for the augmented binary method however the variation in these confidence interval widths is much larger for the unadjusted augmented binary method than the unadjusted standard binary method. It is clear from this that small sample adjustments should be implemented to avoid scenarios where the treatment effect is reported with an extremely large confidence interval.

Risk Difference

Figure 2.4 shows the power for the risk difference, which is similar to the log-odds case. Figure 2.5 shows the type I error rate of the unadjusted and small sample adjusted methods. Both methods experience an inflation in type I error rate. Implementing the correction in the augmented binary GLS method results in a small improvement in the type I error rate with no power lost. GEE adjustments result in an average reduction in type I error of approximately 2.5% however the power drops to below that of the adjusted method using GLS. Again, differences in GLS and GEE diminish as the sample size increases. The adjustment for the standard binary method reduces the type I error rate from 7% to approximately 5%. For all methods the adjustment results in the type I error rate being close to nominal and so should always be implemented when using the risk difference treatment effect estimate.

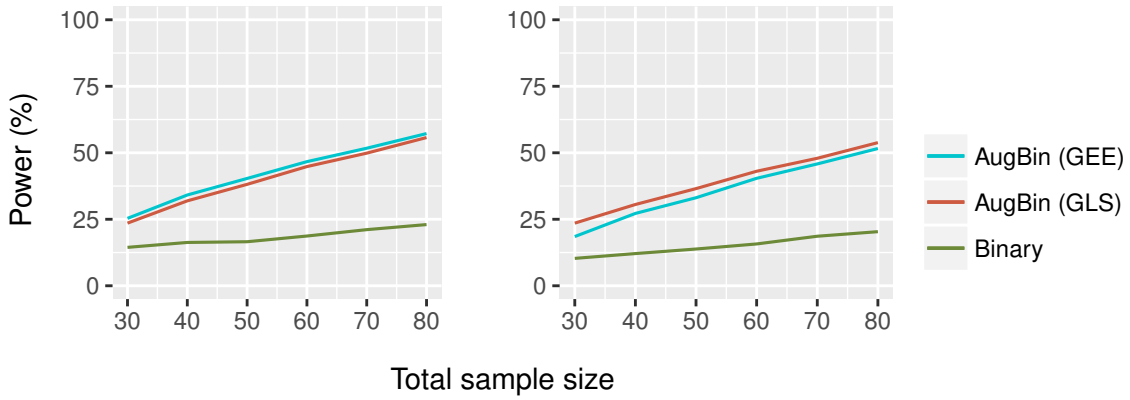


Figure 2.4: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR20 risk difference treatment effect estimate

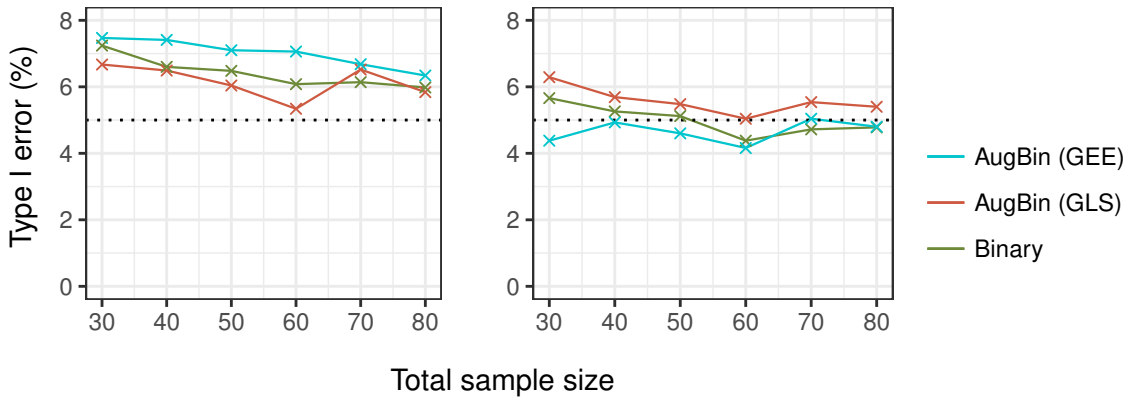


Figure 2.5: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the risk difference treatment effect estimate

Table 2.5: Average risk difference treatment effect estimates and standard deviation (SD) when the intervention does and does not have an effect from the unadjusted and small sample adjusted standard binary and augmented binary methods

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
Null						
30	-0.003	-0.002	0.001	0.000	-0.002	0.000
	(0.182)	(0.166)	(0.151)	(0.144)	(0.154)	(0.145)
40	0.001	0.000	0.001	-0.002	-0.004	0.000
	(0.156)	(0.148)	(0.130)	(0.123)	(0.136)	(0.130)
50	-0.001	-0.002	0.002	0.000	0.001	0.000
	(0.142)	(0.134)	(0.117)	(0.112)	(0.121)	(0.118)
60	0.000	-0.002	0.002	-0.003	-0.002	0.001
	(0.129)	(0.121)	(0.106)	(0.104)	(0.113)	(0.108)
70	-0.004	0.000	0.001	0.001	0.001	-0.001
	(0.120)	(0.114)	(0.101)	(0.098)	(0.104)	(0.101)
80	-0.002	0.000	0.001	0.000	0.000	-0.001
80	(0.112)	(0.108)	(0.094)	(0.092)	(0.097)	(0.094)
Effect						
30	0.129	0.118	0.180	0.170	0.179	0.168
	(0.178)	(0.163)	(0.145)	(0.137)	(0.153)	(0.143)
40	0.133	0.124	0.185	0.171	0.191	0.182
	(0.153)	(0.142)	(0.124)	(0.116)	(0.128)	(0.123)
50	0.131	0.124	0.186	0.195	0.193	0.185
	(0.132)	(0.125)	(0.110)	(0.117)	(0.113)	(0.109)
60	0.131	0.125	0.187	0.181	0.193	0.186
	(0.123)	(0.117)	(0.101)	(0.098)	(0.104)	(0.100)
70	0.134	0.129	0.189	0.184	0.195	0.190
	(0.109)	(0.104)	(0.091)	(0.089)	(0.094)	(0.091)
80	0.133	0.128	0.189	0.184	0.195	0.190
	(0.104)	(0.101)	(0.085)	(0.083)	(0.087)	(0.085)

Table 2.6: Median width of confidence interval with standard deviation in the parenthesis (SD) for the risk difference treatment effect produced from the unadjusted and small sample adjusted standard binary and augmented binary methods

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.675	0.686	0.567	0.554	0.554	0.614
	(0.044)	(0.037)	(248.8)	(0.045)	(245.1)	(0.142)
40	0.592	0.597	0.492	0.464	0.488	0.528
	(0.030)	(0.026)	(36.558)	(0.062)	(39.049)	(0.062)
50	0.536	0.539	0.442	0.470	0.442	0.472
	(0.023)	(0.021)	(14.154)	(0.059)	(18.558)	(0.033)
60	0.490	0.492	0.406	0.403	0.407	0.429
	(0.018)	(0.017)	(3.985)	(0.021)	(5.229)	(0.024)
70	0.455	0.456	0.377	0.375	0.379	0.396
	(0.015)	(0.014)	(0.142)	(0.018)	(0.319)	(0.020)
80	0.426	0.428	0.354	0.351	0.357	0.371
	(0.013)	(0.012)	(0.016)	(0.016)	(0.017)	(0.017)

Table 2.5 shows the average treatment effect in the null case from each of the unadjusted and adjusted methods. The results are similar to the log-odds case; all methods estimate the treatment effect well and have lower variance when small sample corrections are implemented. The table also shows the average effect estimates when there is a treatment effect present. As we do not know the true effect, we cannot quantify bias in the methods however the corrections do modify the treatment effect point estimate and reduce its variance. Table 2.6 shows the median width of the confidence intervals for the risk difference treatment effect along with the standard deviation. The variation in confidence interval size is large for the augmented binary methods however this is corrected for with the small sample adjustments.

To further characterise the benefit of the small sample corrections it is useful to interpret the proportion of cases experiencing perfect separation alongside the average width of the confidence intervals. Table 2.7 shows the percentage of the 5000 sub-samples with confidence intervals for the risk difference that are larger than 1. This is shown for each method at each sample size. From this we can see that perfect separation has occurred in the augmented binary method but not in the standard binary analysis. This is intuitive as fewer events are modelled by each logistic regression model in the augmented binary method, whereas the standard binary method incorporates all events

Table 2.7: Percentage of cases experiencing extremely large variance due to perfect separation from the unadjusted and adjusted standard binary and augmented binary methods on probability scale (confidence interval width for difference >1)

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.00	0.00	10.9	0.00	10.6	0.24
40	0.00	0.00	3.71	0.00	3.81	0.02
50	0.00	0.00	1.11	0.00	1.20	0.00
60	0.00	0.00	0.24	0.00	0.30	0.00
70	0.00	0.00	0.04	0.00	0.08	0.00
80	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.8: Percentage reduction in average confidence interval width and percentage reduction in required sample size for the standard binary method vs. augmented binary method with small sample adjustments on the log-odds and probability scales

Comparison	Reduction in C.I. width (%)	Reduction in sample size (%)
Log-odds		
Stand bin vs. aug bin (GLS)	17.4	31.8
Stand bin vs. aug bin (GEE)	11.2	21.1
Risk difference		
Stand bin vs. aug bin (GLS)	17.6	32.1
Stand bin vs. aug bin (GEE)	12.3	23.1

in one model. The results show that the perfect separation is corrected for using the Firth adjustment and suggest that the corrections are most beneficial when N<60. Table 2.8 shows the average reduction in confidence interval width for the adjusted methods on both scales. We compare the standard binary with both implementations of the augmented binary method and find that the augmented binary method with GLS offers the largest gains in precision. This translates to the adjusted augmented binary method requiring a 32% smaller sample size than what would be required for the adjusted standard binary method.

2.3.5.2 ACR50

Odds Ratio

By investigating the performance of the methods for the ACR50 endpoint, we can understand how lower response rates in both arms affect the behaviour of the methods. The left panel in Figure 2.6 shows the type I error rate of each of the unadjusted methods. The augmented binary method (GLS) has close to nominal type I error rate. The implementation using GEE has a small inflation in type I error rate, which we expect when using robust standard errors in small samples [57]. The type I error rate for the standard binary method is below nominal and close to zero when the total sample size equals 30. The corresponding small sample adjusted type I error rate is shown in the right panel of Figure 2.6. The correction has no effect on the type I error rate for the standard binary or augmented binary (GLS) methods however the type I error rate of the augmented binary (GEE) reduces to below nominal.

Figure 2.7 shows the power of the standard binary and augmented binary methods for the log-odds ACR50 response. The standard binary method has power of between 2% and 36% for total sample sizes between 30 and 80, whereas the augmented binary (GLS) has between 20 and 50%. The small sample adjusted power is shown in the right panel of Figure 2.7. The small sample adjustment does not alter the power for the standard binary and augmented binary (GLS) methods. However, the GEE adjustments reduce the power by 6-8%, meaning that in terms of type I error rate and power, the augmented binary (GLS) method is the best way to model the composite endpoint in trials of small populations and rare diseases for an odds ratio treatment effect.

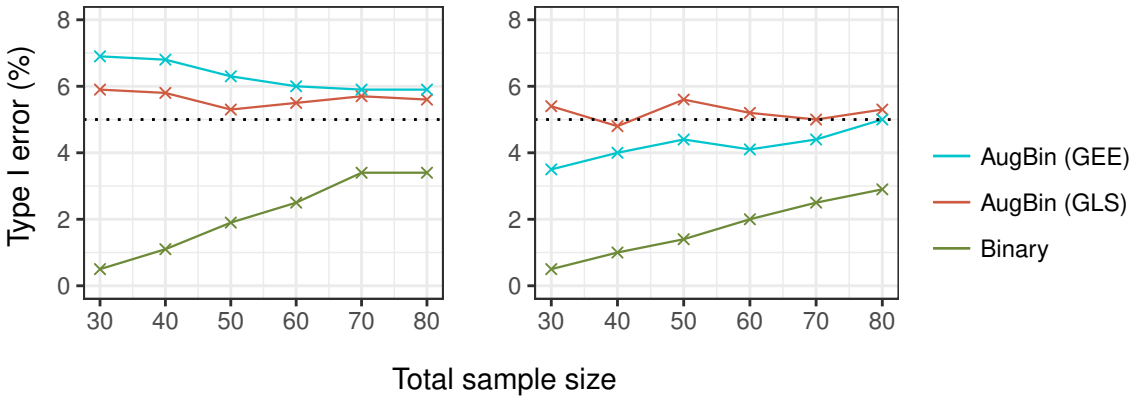


Figure 2.6: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR50 log-odds treatment effect estimate

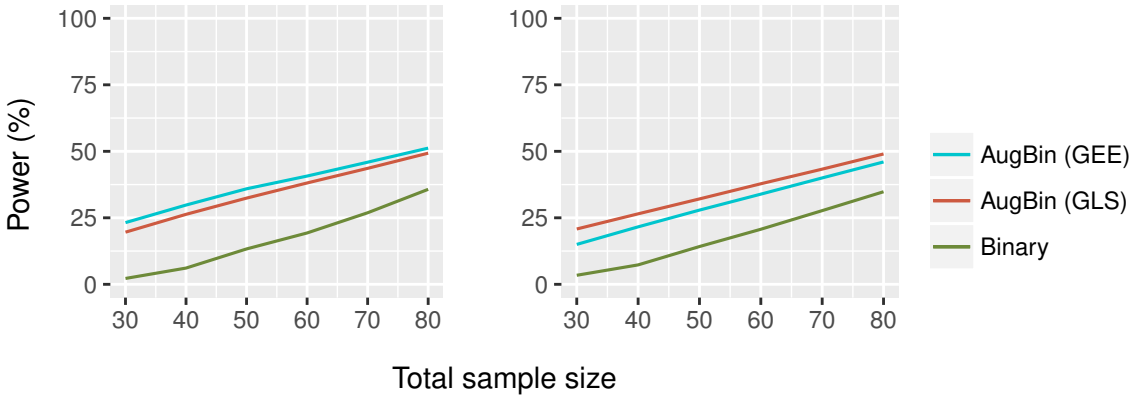


Figure 2.7: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR50 log-odds treatment effect estimate

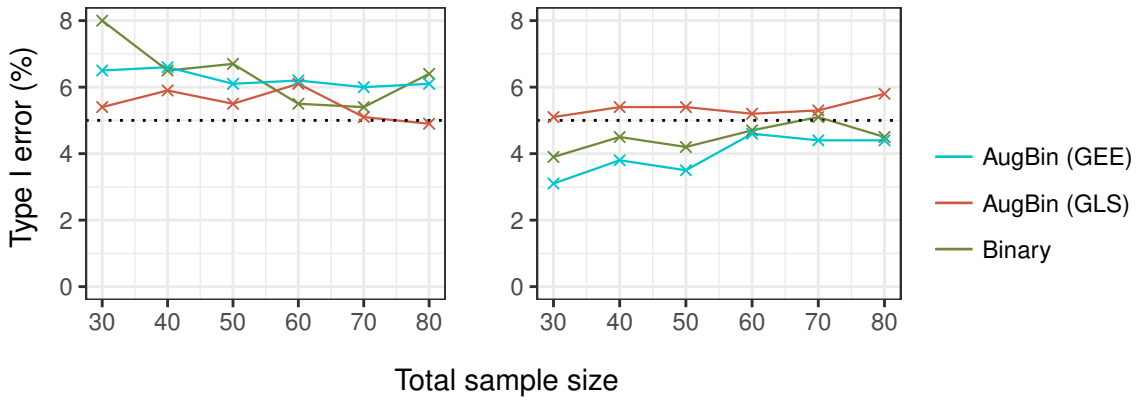


Figure 2.8: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR50 risk difference treatment effect estimate

Risk Difference

The type I error rate of the methods is shown in Figure 2.8. For sample size N<60 the standard binary method has inflated type I error rate. The augmented binary method has nominal type I error rate. This indicates that the augmented binary method still performs well under lower response rates, even when using the risk difference treatment effect. The right panel in the figure shows the type I error rate of the methods with the small sample corrections. The Firth adjustment corrects the type I error rate for the binary method. The augmented binary (GEE) has a reduced type I error rate from having both the Firth and the GEE variance correction implemented. The augmented binary (GLS) type I error rate remains close to nominal.

The power of the methods for the ACR50 risk difference treatment response is shown in Figure 2.9. The power for the standard binary method is 24-47% for the sample sizes investigated. The power of the augmented binary methods, for both GLS and GEE, is 20-50%. This means that the models have similar power for the ACR50 risk difference, however while the type I error rate for the standard binary method is inflated, the type I error rate for the augmented binary method is not. The power for the small sample adjusted methods is shown in the right panel of Figure 2.9. When the type I error rate is corrected, the power of the standard binary method is 16-43%. The power of the augmented binary method (GLS) is 21-50% with nominal type I error rate. Correcting the augmented binary (GEE) method means it has power of 13-45%. Therefore the augmented binary method (GLS) is still the most favourable method in this case.

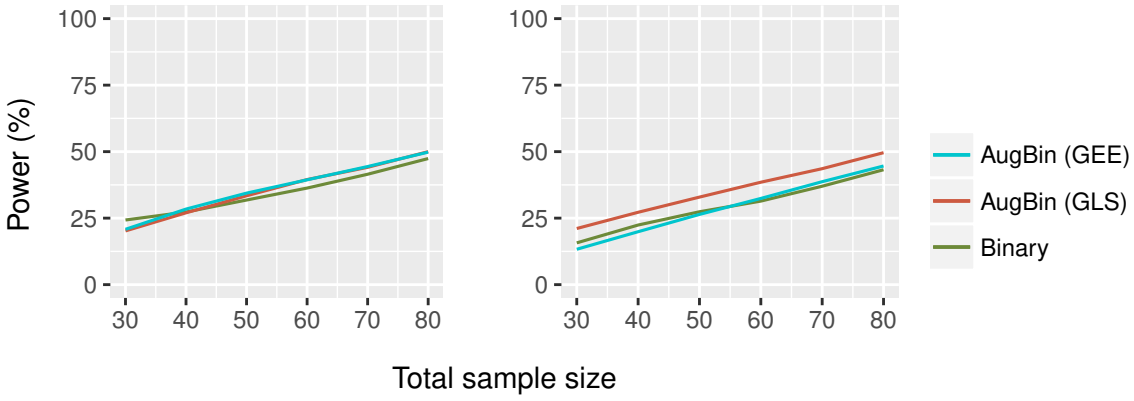


Figure 2.9: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR50 risk difference treatment effect estimate

2.3.5.3 ACR70

In order to understand how the methods work with even lower response rates we investigate the type I error rate and power for the ACR70 endpoint. In this context the patients must improve their ACR by 70%. This is never used as a primary endpoint as response rates at this level are so low, however it will often be considered as one of the secondary endpoints in RA trials and so it is important to understand if the augmented binary method is applicable in this setting.

Odds Ratio

The type I error rate for the log-odds ACR70 treatment response is shown in the left panel of Figure 2.10. The standard binary method has a type I error rate equal to zero for all sample sizes investigated. The type I error rate of the augmented binary method (GLS) is inflated for N<50 and for N<60 for GEE. The small sample adjusted type I error rates are shown in the right panel. The adjusted type I error rate for the standard binary and augmented binary (GLS) method remains unchanged. The adjusted type I error rate for the augmented binary (GEE) is reduced to below nominal.
Figure 2.11 contains the power of the methods for the log-odds ACR70 treatment effect. The standard binary method has power equal to zero. This makes it inappropriate for use when using the log-odds treatment effect estimate when response rates are low. The power of the augmented binary method using GLS and GEE is similar, namely 19-48% for the sample sizes investigated. The right panel of the figure shows the small sample

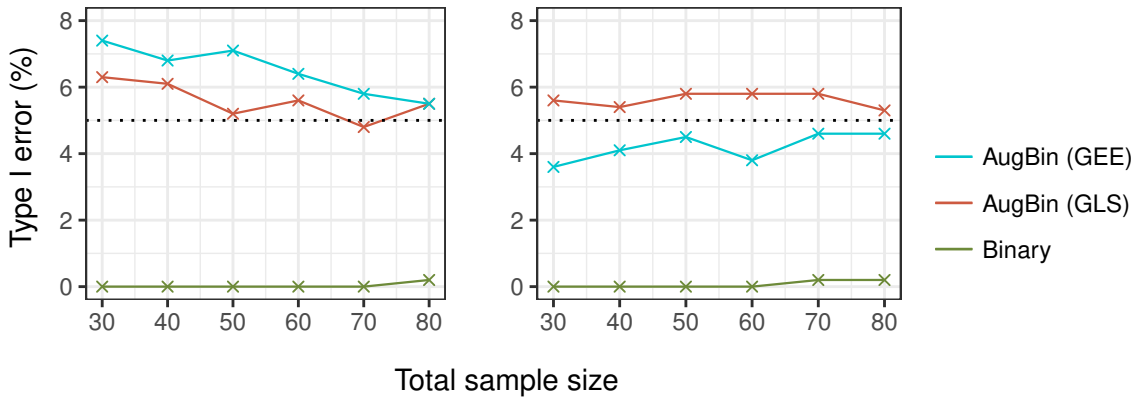


Figure 2.10: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR70 log-odds treatment effect estimate

adjusted power for the methods. The power for the standard binary and augmented binary (GEE) methods remains unchanged. The small sample adjusted power for the augmented binary (GEE) is 14-42%. This indicates that the augmented binary method (GLS) continues to work well with lower response rates when working on the log-odds scale and is the preferred method.

Risk Difference

The type I error rate of the methods for the ACR70 risk difference response is shown in Figure 2.12. The standard binary has below nominal type I error rate for N<70 in this instance, however there is inflation present for N= 30. This is likely due to the combination of a small sample size, very low response rates and a violation of the normality assumption made when calculating the standard errors on the probability scale. Both the augmented binary (GLS) and (GEE) have below nominal type I error rates. The results for the small sample adjusted methods are shown in the right panel of the figure. The adjustment overcorrects the standard binary and augmented binary (GEE) methods in this case with type I error rate 0-3%. The type I error rate for the augmented binary (GLS) remains unchanged.

Figure 2.13 shows the power of the methods for the ACR70 risk difference treatment effect. The standard binary method has power 11-29%. The augmented binary method with GLS and GEE has power 11-42% whilst also having a smaller type I error rate than the standard binary method. The power of the small sample adjusted standard

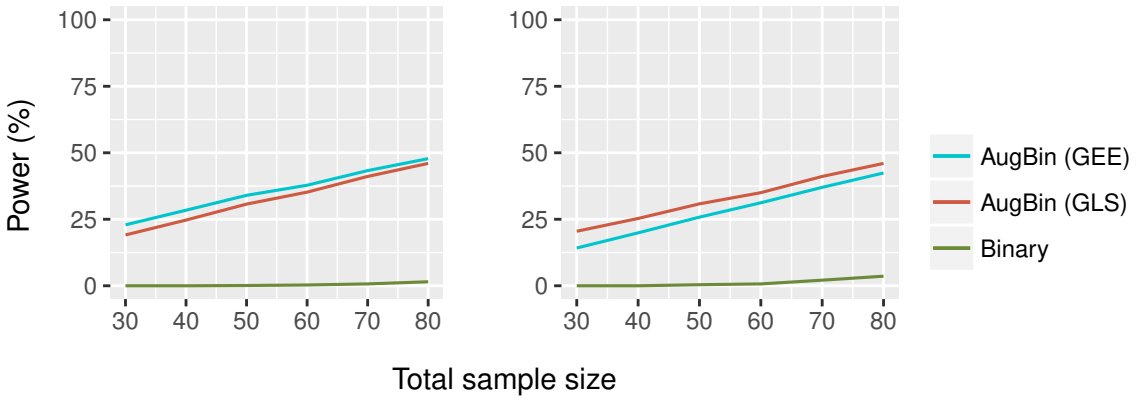


Figure 2.11: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR70 log-odds treatment effect estimate

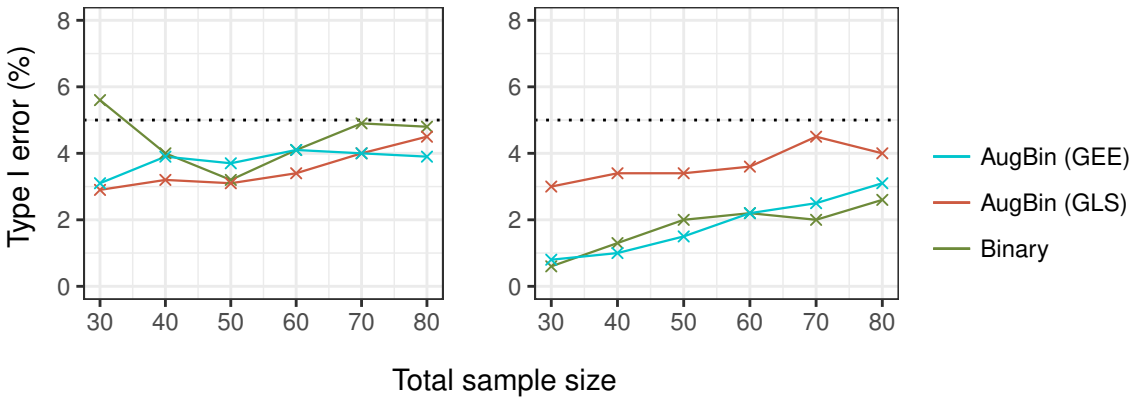


Figure 2.12: Type I error rate of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR70 risk difference treatment effect estimate

binary method drops to 3-18% and the augmented binary (GEE) method to 3-33%. The power of the small sample adjusted augmented binary (GLS) method remains unchanged at 12-42%. Again, the augmented binary method using GLS performs well even with low response rates in each arm.

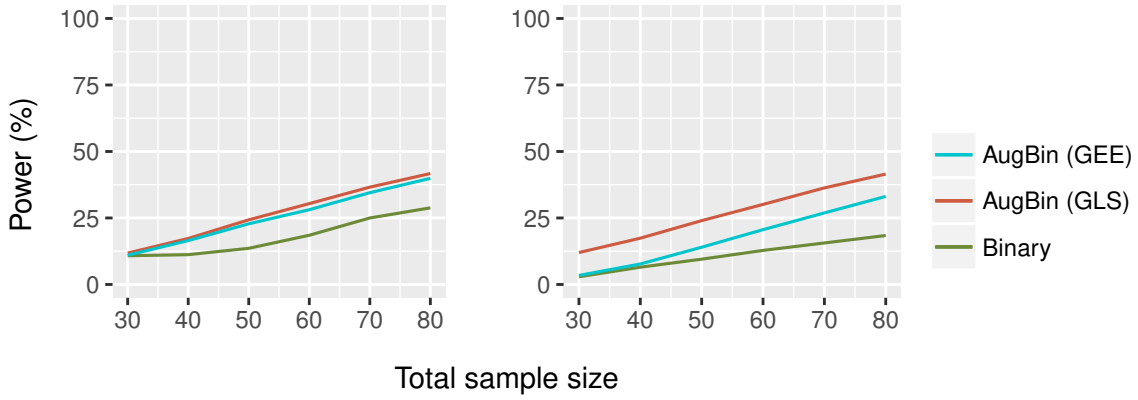


Figure 2.13: Power of the unadjusted standard binary, augmented binary (GEE) and augmented binary (GLS) methods (left) and the corresponding small sample adjusted methods (right) for total sample sizes between 30 and 80 when reporting the ACR70 risk difference treatment effect estimate

2.4 Assessing Properties: Simulated Example

Despite the advantages in using re-sampling as a means of assessing model performance, it is limited in the sense that we do not know the 'correct' answer in each scenario. To verify the findings from the re-sampling, we consider a simulated example from a known distribution. The simulated scenarios are for the ACR20 response, given that this is usually the primary outcome. We begin by setting the probability of response equal to 0.470 in the treatment arm and 0.336 in the placebo arm, similar to the OSKIRA-1 study. Secondly, we simulate under the null where the probability of response equals 0.336 in both arms. We investigate power, type I error rate, average treatment effect estimates and average confidence interval width for the small sample adjusted binary and augmented binary methods.

2.4.1 Data Generating Model

The data generating model used is shown below, which is based on the augmented binary model. As re-sampling investigated the performance of the method under a

realistic data structure where assumptions may be violated, simulating data from the augmented binary model will provide an indication of how the method works when assumptions are satisfied and we know the true parameter values. In this case we consider only the GLS for modelling the continuous component due to its superior performance over GEE in the resampling analyses.

$$\begin{gathered} Y_{ij} = \alpha_{F0}+\alpha_{F1}T_i I\{j=1\} + \alpha_{F2}T_i I\{j=2\} + \alpha_{F3}y_{i0} + \alpha_j + \varepsilon_{ij} \\ (\varepsilon_{i1}, \varepsilon_{i2}) \sim N\left((0,0), \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \end{gathered} \tag{2.3}$$

$$logit\left(P(F_{i1}=1|T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \beta_{F0} + \beta_{F1}T_i + \beta_{F2}y_{i0} \tag{2.4}$$

$$logit\left(P(F_{i2}=1|F_{i1}=0, T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \gamma_{F0} + \gamma_{F1}T_i + \gamma_{F2}Y_{i1} \tag{2.5}$$

We investigate the small sample adjusted measures for the risk difference estimator δ_1.

2.4.2 Results

Table 2.9 shows the average ACR20 risk difference treatment effect for the small sample adjusted methods. The true treatment effect estimate is 0.134. The methods perform similarly with both slightly underestimating the treatment effect in smaller samples and slightly overestimating in larger samples (n> 40). The variability in estimated treatment effects is larger for the binary method in all sample sizes. The power of the small sample adjusted binary method is 15-29% for the sample sizes investigated. The corresponding small sample corrected power for the augmented binary method is 17-43%. The absolute power estimates for both methods differ from those in the re-sampling results, however the comparative conclusions are the same. Namely that the power of the augmented binary method is always larger than that of the standard binary method. Table 2.10 shows the average confidence interval width and the reduction in required sample size from the augmented binary method. The efficiency gains from the augmented binary method amount to reducing the required sample size by 38% to show the same treatment effect. This is true for all sample sizes investigated.

Table 2.11 shows the average treatment effects in the null case from the standard binary and augmented binary methods. The methods are shown to be unbiased for all sample sizes, with smaller variability in treatment effect estimates from the augmented

Table 2.9: Average risk difference ACR20 response with standard deviation in parenthesis (S.D.) and power for the small sample adjusted standard binary and augmented binary methods in 5000 simulations for total sample size between 30 and 80

Total sample	δ_1 (S.D.)		Power	
size	Binary	Augmented binary	Binary	Augmented binary
30	0.128 (0.167)	0.130 (0.121)	0.145	0.172
40	0.132 (0.145)	0.133 (0.106)	0.179	0.226
50	0.138 (0.129)	0.135 (0.097)	0.213	0.278
60	0.137 (0.120)	0.136 (0.088)	0.240	0.329
70	0.135 (0.113)	0.136 (0.083)	0.269	0.367
80	0.138 (0.103)	0.138 (0.077)	0.293	0.425

$\alpha_{F0} = -15, \alpha_{F1} = 2.5, \alpha_{F2} = 2, \alpha_{F3} = 4.1, \alpha_1 = 6, \alpha_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6, _{F0} = -3.8, \beta_{F1} = -0.1, \beta_{F2} = 0.4, \gamma_{F0} = -0.8, \gamma_{F1} = -0.08, \gamma_{F2} = -0.008, \delta_1 \approx 0.134$

binary method. The augmented binary method has nominal type I error rate, which is consistent with the re-sampling results. However, the type I error for the adjusted standard binary method is 6.8-8.1%, which is higher than the type I error rates found from re-sampling. The average confidence interval width in the null case is shown in Table 2.12. The results are consistent with the previous findings from re-sampling.

2.5 Discussion

In this chapter we have explored the small sample properties of the standard binary and augmented binary methods and proposed adjustments to improve them, when necessary. Our findings suggest that the increased efficiency of the augmented binary method does indeed translate to a small sample setting. The method performs better on the log-odds scale, where normality assumptions made when employing the delta method are best satisfied. These assumptions are more questionable when working with samples of this size on the probability scale, which is partly reflected in the differences in inflation present. These findings have been published [58] and the paper is included in Appendix B along with the supplementary material in Appendix C and D.

Taking a societal view of power, as discussed in [59], we can say that rare disease trials are restricted in their capacity to detect treatment effects both because of small studies and few studies running in any given disease. Therefore it follows that maximising power within a single study is perhaps even more crucial than in more common diseases,

Table 2.10: Average confidence interval width and reduction in required sample size (%) for the risk difference ACR20 response for the small sample adjusted standard binary and augmented binary methods in 5000 simulations for total sample size between 30 and 80

Total sample size	Average confidence interval width		Sample size
	Binary	Augmented binary	reduction (%)
30	0.630	0.496	38.0
40	0.550	0.431	38.6
50	0.493	0.386	38.7
60	0.452	0.353	39.0
70	0.419	0.328	38.7
80	0.392	0.306	39.1

$\alpha_{F0} = -15, \alpha_{F1} = 2.5, \alpha_{F2} = 2, \alpha_{F3} = 4.1, \alpha_1 = 6, \alpha_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6,_{F0} = -3.8, \beta_{F1} = -0.1, \beta_{F2} = 0.4, \gamma_{F0} = -0.8, \gamma_{F1} = -0.08, \gamma_{F2} = -0.008, \delta_1 \approx 0.134$

Table 2.11: Average risk difference ACR20 response with standard deviation in parenthesis (S.D.) and type I error in the null case for the small sample adjusted standard binary and augmented binary methods in 5000 simulations for total sample size between 30 and 80

Total sample	δ_1 (S.D.)		Type I error	
size	Binary	Augmented binary	Binary	Augmented binary
30	0.002 (0.157)	0.001 (0.102)	0.068	0.047
40	-0.001 (0.143)	0.001 (0.092)	0.080	0.047
50	0.000 (0.128)	-0.002 (0.081)	0.081	0.044
60	-0.001 (0.118)	0.000 (0.075)	0.079	0.043
70	-0.001 (0.107)	0.000 (0.070)	0.073	0.043
80	0.000 (0.104)	0.000 (0.065)	0.081	0.049

$\alpha_{F0} = -15, \alpha_{F1} = 2.5, \alpha_{F2} = 2, \alpha_{F3} = 4.1, \alpha_1 = 6, \alpha_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6,_{F0} = -3.8, \beta_{F1} = -0.1, \beta_{F2} = 0.4, \gamma_{F0} = -0.8, \gamma_{F1} = -0.08, \gamma_{F2} = -0.008, \delta_1 \approx 0.134$

Table 2.12: Average confidence interval width for the ACR20 risk difference response in the null case for the small sample adjusted standard binary and augmented binary methods in 5000 simulations for total sample size between 30 and 80

Total sample size	Average confidence interval width		Reduction (%)
	Binary	Augmented binary	
30	0.596	0.426	28.5
40	0.517	0.370	28.4
50	0.465	0.332	28.6
60	0.425	0.303	28.7
70	0.394	0.282	28.4
80	0.369	0.263	28.7

$\alpha_{F0} = -15, \alpha_{F1} = 2.5, \alpha_{F2} = 2, \alpha_{F3} = 4.1, \alpha_1 = 6, \alpha_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6, _{F0} = -3.8, \beta_{F1} = -0.1, \beta_{F2} = 0.4, \gamma_{F0} = -0.8, \gamma_{F1} = -0.08, \gamma_{F2} = -0.008, \delta_1 \approx 0.134$

which can accumulate power over many studies as well as large individual studies. This additional power may be realised in practice when conducting meta-analysis. Consequently, the increased power offered from the augmented binary method is an important development for analysing small sample data and should be considered for the primary analysis method in trials of rare diseases using these endpoints.

Our findings show that the treatment effect scale and estimation method used is important when conducting analysis in small samples. When implementing the augmented binary method in rare disease trials we recommend the use of the Firth adjustment for the logit models as it reduces the bias and variance of the estimates. This is especially valuable in this setting due to the restrictive nature of sample size. For the continuous component, we recommend the GLS estimator. As well as offering the best power and precision, GLS methods make more realistic assumptions about the mechanism for missing responses, namely that they are missing at random rather than missing completely at random. Moreover they experience fewer convergence issues in very small samples.

Another important consideration is the role of response rate in each arm on the operating characteristics of interest. The ACR50 and ACR70 results indicate that power and type I error are highly dependant on responder rates. For the standard binary method, the results show deflations in the type I error rate on the log-odds scale and inflations on the probability scale, with type I error rates ranging from 0 to 8%. This

is likely due to logistic regression methods having poorly estimated standard errors when there are few events per parameter, as is the case for the ACR50 and ACR70 endpoints [60]. Overall, the augmented binary method shows fewer deviations from nominal type I error rates whilst exhibiting increased power over the standard binary method in every scenario investigated, indicating that it is still an appropriate analysis method in small samples when response rates in each arm are low.

Using re-sampling from real data has the advantage that we test the performance of the methods under realistic data structures and can understand how the methods behave when the assumptions are not necessarily satisfied. However, as the power and type I error rates are proxies for the true quantities it is useful to supplement the re-sampling with simulations from a known distribution. The comparative findings from the simulated example are in agreement with re-sampling and further reiterate the problems with type I error rate control in the standard binary method. As the type I error rate is more stable for the augmented binary method both in the re-sampling and the simulated example, the overall findings show that it is more robust in the rare disease setting than logistic regression methods.

Although it is recognised that novel methods developed for use in rare diseases may be of more immediate utility than in common diseases, some resistance to implementing the augmented binary method in real rare disease trials may be experienced due to its increased complexity. To assist with this we have made the R code fully available when publishing the work [58]. It is of paramount importance that the efficiency gains provided by this method are not used as a substitute for other important efforts and considerations undertaken when running rare disease trials. That is, the method should be used to complement efforts in establishing international, multi-centre trials with maximum feasible enrolment periods, alongside other achievable strategies to increase sample size; not to replace them.

There are some limitations in what we have presented. We have only investigated the performance of the method in small samples by re-sampling from rheumatoid arthritis data. Similar procedures may be carried out in other data sets and the methods applied directly to rare disease data, to ensure these gains are always experienced across a range of responder indices and response rates. Moreover, due to the increased number of parameters, the augmented binary method starts to experience some problems when the total sample size is reduced to N=20. This is unlikely to be a problem in practice, as a randomised trial as small as this would be unusual. If required, it may be possible to make further assumptions in order to reduce the number of parameters

to be estimated, such as assuming that the effect of the continuous measure on failure probability is the same across all time points.

A further extension which will be considered in Chapter 3, is the development of joint modelling methods for the instance when the composite is a more complicated combination of outcomes, namely multiple continuous, ordinal and binary components. We expect these methods to exhibit even larger efficiency gains due to using information in multiple continuous and ordinal components. This will provide the potential to further improve the frequency and quality of evidence generated in many rare disease areas.

Chapter 3

Complex Composite Structures

3.1 Motivation

The augmented binary method discussed in Chapter 2 appears to perform well when the composite is formed from one continuous measurement and one binary indicator. Furthermore, it can be employed in any responder endpoint with multiple components, provided that at least one is continuous. However, given that modelling one continuous component through the augmented binary method is shown to reduce the required sample size by at least 32%, we hypothesise that modelling additional continuous components would result in an even greater improvement in efficiency.

The aim of the work in this chapter is to extend and employ methodology to appropriately model composite endpoints with a structure more complex and information rich than what was previously considered. Table 3.1 shows examples of endpoints with multiple continuous and discrete components. Response definitions vary, for instance responders in fibromyalgia must respond in two continuous and one ordinal category however responders in trials for frailty or soft tissue infections must respond in a total of five continuous and discrete components. It is therefore desirable that the methods developed in this chapter allow for these variations in response definitions, either by modelling all of the outcomes or collapsing additional outcomes into a single binary indicator. The primary motivating example for this work is a composite endpoint used in SLE, which is made up of two continuous, one ordinal and one binary component. We will explore this endpoint in detail and propose methodology for analysing it however the developments will be applicable to many other diseases that use similar endpoints.

Table 3.1: Examples of diseases that use complex composite endpoints combining multiple discrete and continuous measures to determine effectiveness of a treatment including criteria for response and how each component is typically measured

Disease	Responder endpoint	Measured by
Fibromyalgia	• 30% improvement in pain	Electronic diary
	• 30% improvement in functional status	Subscale of Fibromyalgia Impact Questionnaire (FIQ)
	• improved, much improved, or very much improved	7-point Patient Global Impression of Change (PGIC)
Frailty	• BMI<18.5 kg/m2 OR >10% weight loss since last wave	weight and height
	• One positive answer to exhaustion questions	CES-D questionnaire
	• Low grip strength (M < 31.12 kg, F < 17.60 kg)	Jamar hand dynamometer
	• Gait speed (M < 0.691 m/s, F < 0.619 m/s)	Distance/time
	• Low activity (M < 16.5 activity units F < 13.5 activity units)	Activity units derived using intensity vs. frequency
Necrotizing Soft Tissue Infections	• Alive until day 28	yes/no
	• Day 14 debridements ≤ 3	surface area
	• No amputation if debridement	yes/no
	• Day 14 mSOFA score ≤ 1	mSOFA score
	• Reduction of at least 3 score	mSOFA score

3.1.1 Application: Systemic Lupus Erythematosus

SLE is a complex multisystem autoimmune disease resulting from a diversity of clinical features and factors (genetic, hormonal and environmental) [61]. In the US population, the prevalence was 52.2 per 100,000, with a comparative figure of 26.2 in the UK [62]. These figures are substantially higher in some ethnic populations. Treatment of SLE is typically challenging because of the limited efficacy and poor tolerability of standard therapy. Furthermore, due to its intricate nature, it is challenging to effectively measure disease status and indeed improvement. In an attempt to capture the complexity of the disease, SLE makes use of a Systemic Lupus Responder Index (SRI) which is comprised of a continuous SLE Disease Activity Index (SLEDAI), a continuous Physicians Global Assessment (PGA) and an ordinal British Isles Lupus Assessment Group measure (BILAG) [63, 64]. To determine efficacy in many SLE trials, the SRI endpoint is combined with a binary indicator containing information about additional medication, such as the tapering of oral corticosteroids, to form the responder index. The structure of the composite is shown in Figure 3.1. Note that the continuous SLEDAI measure is a scoring system composed of 102 items and the ordinal BILAG measure arises from 24 items.

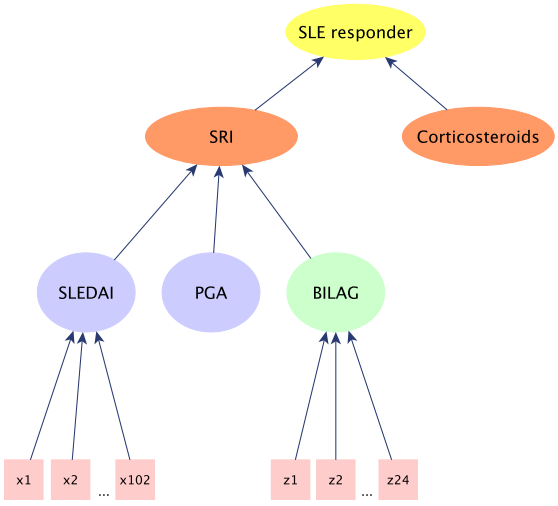


Figure 3.1: Structure of the composite endpoint used in trials of systemic lupus erythematosus (SLE), where patients must respond in all components to be responders overall. The continuous Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), continuous Physicians Global Assessment (PGA) and ordinal British Isles Lupus Assessment Group (BILAG) measures are dichotomised and combined to form the binary Systemic Lupus Erythematosus Response Indicator (SRI) which is then combined with the binary corticosteroids taper variable to form the overall binary SLE responder index

Table 3.2: Response definition in the four components of the Systemic Lupus Erythematosus Responder Index (SRI): Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), British Isles Lupus Assessment Group (BILAG), Physician's Global Assessment (PGA) and corticosteroid taper measure

Component	Response definition
SLEDAI-2K	Reduction from baseline of at least 4 points in the Systemic Lupus Erythematosus Disease Activity Index 2000 according to the SRI-4 response definition
BILAG	No new organ systems affected as defined by 1 or more British Isles Lupus Assessment Group A or greater than one Group B item compared to baseline
PGA	No worsening from baseline in subjects lupus disease activity defined by an increase 0.30 points on a 3-point visual analogue scale (VAS)
Corticosteroids	No discontinuation of investigational drug or use of restricted medications beyond the protocol-allowed threshold with a sustained reduction in the dose of corticosteroids

As is standard in responder analysis, this is also analysed by dichotomising each component and using logistic regression on the overarching responder endpoint. Employing the standard binary method in this instance is subject to even greater losses in efficiency due to losing information in multiple continuous and ordinal outcomes. In order to appropriately reflect the structure of the endpoint and retain information from each component's original scale the method must allow for joint modelling two continuous, an ordinal and a binary outcome. Table 3.2 shows the response definition in each of the four components. SRI-4 is the endpoint commonly employed in trials which means that the improvement threshold from baseline is set at 4 points on the SLEDAI scale. Other endpoints of interest are SRI-5 and SRI-6, which result in lower response rates.

SLEDAI-2K

The SLEDAI-2K index is an assessment which consists of 24 lupus-related items that a physician will complete to decide whether each of the 24 items is 'present' or 'absent' in the last 4 weeks. It is a weighted instrument, in which the presence of a descriptor is multiplied by the particular organ's 'weight', for example, renal

Table 3.3: Grading system in the British Isles Lupus Assessment Group index (BILAG) used to measure disease activity across nine organ systems in systemic lupus erythematosus

BILAG grade	Definition
Grade A	Active disease requiring immunosuppressive drugs and/or a prednisone dose of >20 mg/day or equivalent
Grade B	Moderate disease activity requiring a lower dose of corticosteroids, topical steroids, immunosuppressives, antimalarials, or NSAIDs
Grade C	Mild, stable disease
Grade D	No disease activity but the system has previously been affected
Grade E	No current or previous disease activity

descriptors are multiplied by 4 and central nervous descriptors by 8. These weighted organ manifestations are subsequently totalled to obtain the final score. The assessment also includes the collection of blood and urine to evaluate the laboratory categories. The overall SLEDAI-2K score range is 0 to 105 points with 0 indicating inactive disease [65].

BILAG

The BILAG-2004 is a translational index with nine organ systems, namely General, Mucocutaneous, Neuropsychiatric, Musculoskeletal, Cardiorespiratory, Gastrointestinal, Ophthalmic, Renal and Haematology. It has ordinal scales by design and records disease activity across the different organ systems by comparing manifestations occurring in the last 4 weeks with the previous 4 weeks. The nine organ systems incorporate a total of 97 items, each of which will receive a grade based on the grading system in Table 3.3.

PGA

The PGA is a global assessment, factoring in all aspects of the subjects lupus disease activity and is completed by a certified investigator. It represents the physician's overall assessment of average SLE disease severity on a visual analogue scale (VAS)

with 0 representing no disease to 3 indicating severe disease activity over the last 4 weeks. A score of 3 refers to the most severe possible disease in all SLE subjects and therefore the rating should virtually never reach 3. Any disease rated greater than 2.5 is very severe, moderate disease covers approximately 1.5 to 2.4 and mild disease falls below 1.5. The instrument is similar to a logarithmic scale, with greater distances or demarcations possible among more mild-moderate symptoms. When scoring the PGA, the mark on the VAS should be moved relative to the score from the previous visit and wherever possible should be completed by the same physician for a given patient.

Corticosteroid Tapering

In this instance the binary indicator contains information on oral corticosteroid use. Specifically, in order to be a responder in this outcome a patient must have no discontinuation of the investigational drug or use of restricted medications beyond the protocol-allowed threshold prior to assessment with a sustained reduction in the dose of corticosteroids. However, this binary indicator may contain any combination of additional criteria that clinicians or patients may feel is important to meet in order to demonstrate improvement. It is also the case that the outcome of interest may be the SRI-4 endpoint alone and so the methodology should be flexible to the inclusion or exclusion of this component in the overall composite.

3.2 Background

The main obstacle when jointly modelling variables of a different nature is the non-existence of an obvious multivariate distribution. Over the past 20 years there have been substantial developments in statistical methodology for the analysis of mixed data. Many of these ideas have roots in much earlier work but advances in computing have made them practical for use more recently.

3.2.1 Copulas

One family of models used to model mixed outcome types which feature frequently in economics and finance are copulas. These are functions that join or couple multivariate distribution functions to their uniform one-dimensional marginal distribution functions [66]. Copulas offer a flexible framework in this setting, as the marginal distribution functions need not come from the same parametric family. While the construction of

copulas is considered to be mathematically elegant and the flexibility with which we can model appealing, they are not without their shortcomings. Extensions beyond the bivariate setting are difficult and have failed to perform well in many applications [67]. Other practical implications include poor out-of-sample predictions due to the wide variety of copulas available. These restrictions, along with difficulties in longitudinal settings with unbalanced data structures, have seen few applications of copulas for mixed outcome types in the medical statistics literature [68]. Applications of copulas in mixed outcome settings include [69, 70].

3.2.2 Factorisation

One likelihood based method for handling mixed data is the factorisation model. The objective is to factorise the joint distribution and fit a univariate model to each component of the factorisation [67]. This accounts for correlations between the outcomes by including one response as a covariate in the model for the other response. In the graphical modelling literature this has been termed the 'Conditional Gaussian Distribution'(CGD) [71, 72] and is the basis for the augmented binary method [35, 36, 58]. An advantage of these methods in relation to the composite endpoint problem is that we may account for correlations between measurements whilst making inference directly on the outcomes that we have measured, hence they fall within a broader class of 'direct methods'. Examples of other applications of these ideas, which build on the work of Olkin and Tate [73], include developmental toxicity studies [74]. One difficulty with these methods beyond the bivariate scenario is the range of possibilities for the factorisations, with no consensus on how this should be determined. In the case of the SLE responder endpoint containing four components, this amounts to 24 possible factorisations, each of which may result in different conclusions [67, 68]. Furthermore, modelling the endpoint using these methods would account for correlation between outcomes, but only in a restricted way as this is accounted for by including the outcome measured on one component as a covariate in the model for the other component.

3.2.3 Latent Variable Models

Another likelihood based method that allows for more flexibility when modelling the correlations between outcomes falls within the framework of latent variable models [75]. The multiple outcomes are assumed to be physical manifestations of some underlying

latent process, by including the same latent variable in each of the models for the observed responses. The outcomes are then assumed to be independent conditional on this latent variable. This solves the problem of deciding the order of factorisations in previously discussed methods however this formulation results in the inclusion of some covariance parameters in the mean structure, leaving the model sensitive to misspecification of the correlation structure [76]. One example of these models is seen in [77], where effects of covariates of interest are modelled through this shared latent variable. Although these models have the intuitive interpretation that each outcome is attempting to capture underlying disease activity, the correlation matrix is restricted to allow for the same correlation between each pair of outcomes, which is unlikely in practice. This structure is relaxed in [78], where the effects of covariates are included in the model separately from the latent variable. The correlation structure can be further relaxed to allow for a different latent variable for each outcome, meaning that pairs of outcomes are not assumed to have the same correlation. However these models would require integrating out each of the latent variables in order to obtain the joint distribution of interest [79]. Furthermore, they are relevant in applications with multiple time points however less so for a single time point, as is the case for the composite endpoint problem.

3.2.4 Extensions to Multivariate Probit Models

Latent variables have also been used in the setting of mixed continuous and discrete variables to a different end. Namely, the outcomes adopt a correlated Gaussian distribution by assuming that the discrete outcomes are coarsely measured manifestations of underlying continuous variables subject to some threshold specifications [80, 81]. Specifying discrete variables in terms of a partitioning of the latent variable space dates back to Pearson in 1904 [82] in relation to his generalised theory of alternative inheritance, and has received much consideration in the literature since. Terminology surrounding these models is inconsistent but they are often referred to as multivariate probit models [80]. In the graphical modelling literature they have been termed 'conditional grouped continuous models' (CGCMs) [83] and elsewhere have been referred to as 'multivariate ordered probit models' [84], 'correlated probit models' [85] and 'generalized multivariate probit models' [86]. The general mixed-data model introduced by [87] for mixed nominal, ordinal, and continuous data also reduces to a CGCM in the absence of nominal outcomes. By formulating the distribution in this way, we can correlate the error terms for the components and work within the familiar paradigm of

Gaussian distributions and maximum likelihood theory. The theory and application of these ideas for a mixture of continuous and binary outcomes has featured in the statistics literature, see for example [88–90]. Generalisations of these ideas, which appear less frequently in the literature, lead to methods for modelling continuous and ordinal variables, with applications in developmental toxicology and the joint modelling of hybrid traits in genetics [91–96]. A CGCM has also been proposed in clinical trials to deal with the problem of multiple continuous and binary co-primary endpoints, where a treatment effect must be achieved in all outcomes to conclude it is successful overall [97, 98]. Despite the advantages, the multivariate probit model has not realised its full potential in the applied biostatistics literature. This was noted by [99] and we believe it still to be the case today, where any applications that do appear tend to demonstrate bivariate scenarios. Other work has combined thresholding the response variables and introducing latent variables in the model however this is simply a reparameterisation of the case with latent outcomes only. Examples of this in the binary and continuous cases, including generalisations for the longitudinal setting, can be found in [67, 85], and the continuous and ordinal case in [100, 101].

The latent outcome framework employed in the CGCM is sufficiently complex and we propose its use for the composite endpoint problem. However, the purpose of this work is to employ the framework to a different end. Rather than using the latent Gaussian distribution to make inference on multivariate outcomes we will use it to model the multiple components within a composite, while still making inference on the one-dimensional composite endpoint based on proportions of patients crossing responder thresholds on each outcome. By employing the latent structure to collapse the multiple outcomes after the model is fitted rather than before, we aim to greatly improve efficiency whilst still providing the same overall treatment effect measure on the composite.

3.3 Latent Variable Model

3.3.1 Notation

Let $\mathbf{Y_i} = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})^T$ represent the vector of observed outcomes for patient $i \in N$ with mean values $(\mu_1, \mu_2, \mu_3, \mu_4)^T$ and $\mathbf{Y} = (\mathbf{Y_1}, \ldots \mathbf{Y_N})^T$ represent the observed outcomes for all patients. Y_{i1} and Y_{i2} are the observed continuous SLEDAI and PGA measures. Let Y_{i3} denote BILAG, the observed ordinal manifestation of Y^*_{i3} with

mean μ_3^* and Y_{i4} the observed binary taper variable for latent Y_{i4}^* with mean μ_4^*. We therefore let $\mathbf{Y_i^*} = (Y_{i1}, Y_{i2}, Y_{i3}^*, Y_{i4}^*)^T$ denote the vector of observed and latent continuous measures for patient i and $\mathbf{Y}^* = (\mathbf{Y_1^*}, \dots \mathbf{Y_N^*})^T$. T_i represents the treatment indicator for patient i, y_{i10} and y_{i20} are the baseline measures for Y_{i1} and Y_{i2} respectively.

3.3.2 Model

The mean structure for the outcomes is shown in (3.1). The baseline measures y_{10} and y_{20} are included in the model for Y_1 and Y_2 respectively.

$$\begin{aligned} Y_{i1} &= \alpha_0 + \alpha_1 T_i + \alpha_2 y_{i10} + \varepsilon_{i1} \\ Y_{i2} &= \beta_0 + \beta_1 T_i + \beta_2 y_{i20} + \varepsilon_{i2} \\ Y_{i3}^* &= \gamma_1 T_i + \varepsilon_{i3}^* \\ Y_{i4}^* &= \psi_0 + \psi_1 T_i + \varepsilon_{i4}^* \end{aligned} \tag{3.1}$$

The observed discrete variables are related to the latent continuous variables by partitioning the latent variable space, as shown in (3.2). The lower and upper thresholds for both discrete variables are set at $\tau_{03} = \tau_{04} = -\infty, \tau_{53} = \tau_{24} = \infty$. The intercept term for the ordinal variable in (3.1) is set at $\gamma_0 = 0$ so that the cut-points $\tau_{13}, \tau_{23}, \tau_{33}, \tau_{43}$ may be estimated. The intercept for the binary outcome ψ_0 may be estimated, as $\tau_{14} = 0$.

$$Y_{i3} = \begin{cases} \text{Grade E} & \text{if } \tau_{03} \leq Y_{i3}^* < \tau_{13}, \\ \text{Grade D} & \text{if } \tau_{13} \leq Y_{i3}^* < \tau_{23}, \\ \text{Grade C} & \text{if } \tau_{23} \leq Y_{i3}^* < \tau_{33}, \\ \text{Grade B} & \text{if } \tau_{33} \leq Y_{i3}^* < \tau_{43}, \\ \text{Non-responder} & \text{if } \tau_{43} \leq Y_{i3}^* < \tau_{53} \end{cases} \quad Y_{i4} = \begin{cases} 0, & \text{if } \tau_{04} \leq Y_{i4}^* < \tau_{14}, \\ 1, & \text{if } \tau_{14} \leq Y_{i4}^* < \tau_{24} \end{cases} \tag{3.2}$$

Following these assumptions, we can model the error terms in (3.1) as multivariate normal with zero mean and variance-covariance matrix Σ, as shown in (3.3). Note that the error variances for $\varepsilon_3^*, \varepsilon_4^*$ are $\sigma_3 = 1$ and $\sigma_4 = 1$ however this does not represent a constraint on the model but rather a rescaling required for identifiability.

$$(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon^*_{i3}, \varepsilon^*_{i4}) \sim N(\mathbf{0}, \Sigma) \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 & \rho_{14}\sigma_1 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 & \rho_{24}\sigma_2 \\ \rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 & \rho_{34} \\ \rho_{14}\sigma_1 & \rho_{24}\sigma_2 & \rho_{34} & 1 \end{pmatrix} \tag{3.3}$$

The joint likelihood contribution for patient i with for instance, Y_{i3} = Grade C and $Y_{i4} = 0$, can be factorised as shown below.

$$l\left(\boldsymbol{\theta}; \mathbf{Y}^*_{\mathbf{i}}\right) = f(Y_{i1}, Y_{i2}; \boldsymbol{\theta}) \int_{\tau_{23}}^{\tau_{33}} \int_{-\infty}^{0} f(Y^*_{i3}, Y^*_{i4} | Y_{i1}, Y_{i2}; \boldsymbol{\theta}) dy^*_4 dy^*_3 \tag{3.4}$$

where,
$\boldsymbol{\theta} = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \gamma_1, \psi_0, \psi_1, \sigma_1, \sigma_2, \rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}, \tau_{13}, \tau_{23}, \tau_{33}, \tau_{43})$

Note that it is possible to evaluate the joint likelihood contribution for patient i using $f(Y_{i1}, Y_{i2}, Y^*_{i3}, Y^*_{i4}; \boldsymbol{\theta})$ however factorising as in (3.4) may reduce computational times, particularly in high-dimensional models. This formulation also allows us to express the observed likelihood as shown in (3.5).

$$l(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{N} \prod_{w=1}^{5} \prod_{k=1}^{2} f(Y_{i1}, Y_{i2}; \boldsymbol{\theta}) \left[pr\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = y_{i1}, Y_{i2} = y_{i2}; \boldsymbol{\theta}\right)\right]^{I\{Y_{i3}=w, Y_{i4}=k\}} \tag{3.5}$$

The joint probability of patients having discrete measurements $Y_{i3} = w$ and $Y_{i4} = k$ must be multiplied over the five ordinal levels and two binary levels resulting in ten combinations of the probabilities in (3.6) to be calculated.

$$\begin{aligned} &P\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = y_{i1}, Y_{i2} = y_{i2}; \boldsymbol{\theta}\right) = \\ &\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \\ &\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) \end{aligned} \tag{3.6}$$

where Φ_2 is the bivariate standard normal distribution function, $\mu_{3|1,2}, \mu_{4|1,2}$ are the conditional means of $Y_{3,4|1,2}$ and $\Sigma_{3,4|1,2}$ is the corresponding covariance matrix. These

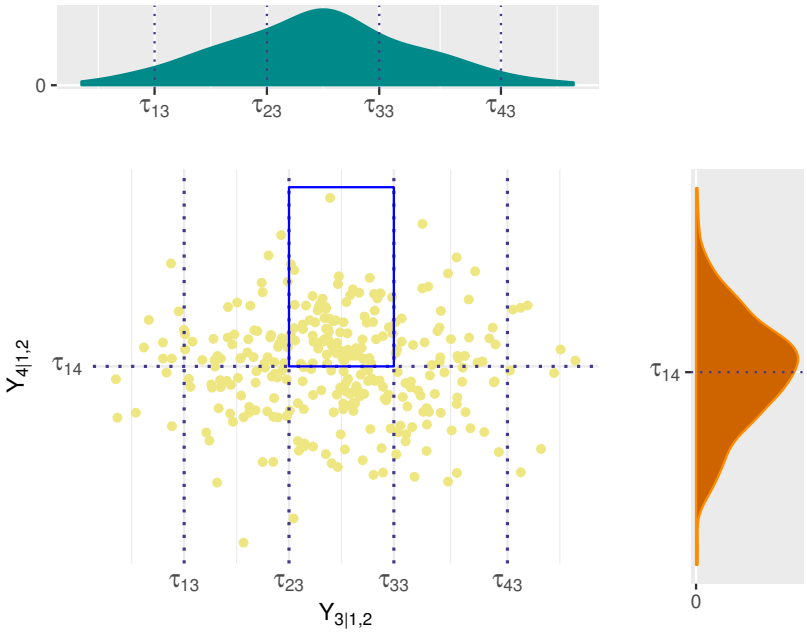


Figure 3.2: The figure shows the continuous space of $Y_{3|1,2}$ and $Y_{4|1,2}$ including the thresholds which are assumed to partition the continuous space to form discrete variables. The joint probabilities in each of the 10 combinations can then be determined using the bivariate distribution function and the thresholds

are derived using conditional multivariate normality rules, resulting in (3.7).

$$\begin{aligned}\mu_{3|1,2} =& \mu_3 + \frac{(\rho_{13}-\rho_{12}\rho_{23})}{\sigma_1\left(1-\rho_{12}^2\right)}\left(Y_{i1}-\mu_1\right) + \frac{(\rho_{23}-\rho_{12}\rho_{13})}{\sigma_2\left(1-\rho_{12}^2\right)}\left(Y_{i2}-\mu_2\right) \\ \mu_{4|1,2} =& \mu_4 + \frac{(\rho_{14}-\rho_{12}\rho_{24})}{\sigma_1\left(1-\rho_{12}^2\right)}\left(Y_{i1}-\mu_1\right) + \frac{(\rho_{24}-\rho_{12}\rho_{14})}{\sigma_2\left(1-\rho_{12}^2\right)}\left(Y_{i2}-\mu_2\right)\end{aligned} \tag{3.7}$$

$$\Sigma_{3,4|1,2} = \begin{pmatrix} 1-\frac{\rho_{13}^2-2\rho_{12}\rho_{13}\rho_{23}+\rho_{23}^2}{1-\rho_{12}^2} & \rho_{34}-\frac{\rho_{13}\rho_{14}-\rho_{12}\rho_{13}\rho_{24}-\rho_{12}\rho_{14}\rho_{23}+\rho_{23}\rho_{24}}{1-\rho_{12}^2} \\ \rho_{34}-\frac{\rho_{13}\rho_{14}-\rho_{12}\rho_{13}\rho_{24}-\rho_{12}\rho_{14}\rho_{23}+\rho_{23}\rho_{24}}{1-\rho_{12}^2} & 1-\frac{\rho_{14}^2-2\rho_{12}\rho_{14}\rho_{24}+\rho_{24}^2}{1-\rho_{12}^2} \end{pmatrix}$$

For the SLE case, where $w = 5$ and $k = 2$, the intuition for the bivariate conditional probability in (3.6) is shown in Figure 3.2. The probability of a given patient falling within the highlighted section, which indicates a Grade C BILAG reading and non-

response in the oral corticosteroids measure, is determined by (3.8).

$$\begin{aligned} &P\left(Y_{i3}=3, Y_{i4}=2 | Y_{i1}=y_{i1}, Y_{i2}=y_{i2}; \boldsymbol{\theta}\right)= \\ &\quad \Phi_2\left(\tau_{33}-\mu_{3|1,2}, \tau_{24}-\mu_{4|1,2}; \Sigma_{3,4|1,2}\right)-\Phi_2\left(\tau_{(23}-\mu_{3|1,2}, \tau_{24}-\mu_{4|1,2}; \Sigma_{3,4|1,2}\right)- \\ &\quad \Phi_2\left(\tau_{33}-\mu_{3|1,2}, \tau_{(14}-\mu_{4|1,2}; \Sigma_{3,4|1,2}\right)+\Phi_2\left(\tau_{23}-\mu_{3|1,2}, \tau_{14}-\mu_{4|1,2}; \Sigma_{3,4|1,2}\right) \end{aligned} \quad (3.8)$$

3.3.3 Estimation

As the variance parameters (σ_1, σ_2) are required to be greater than 0, we introduce parameters (υ_1, υ_2) such that

$$\begin{aligned} \sigma_1 &= exp(\upsilon_1) \\ \sigma_2 &= exp(\upsilon_2) \end{aligned}$$

This transformation ensures that the variance is above 0 whilst allowing the estimated parameter to take any real value. We must also ensure that the correlation parameters $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$ are estimated within (-1,1) by introducing $(\upsilon_{12}, \upsilon_{13}, \upsilon_{14}, \upsilon_{23}, \upsilon_{24}, \upsilon_{34})$, where

$$\begin{aligned} \rho_{12} &= 2logit^{-1}(\upsilon_{12}) - 1 \\ \rho_{13} &= 2logit^{-1}(\upsilon_{13}) - 1 \\ \rho_{14} &= 2logit^{-1}(\upsilon_{14}) - 1 \\ \rho_{23} &= 2logit^{-1}(\upsilon_{23}) - 1 \\ \rho_{24} &= 2logit^{-1}(\upsilon_{24}) - 1 \\ \rho_{34} &= 2logit^{-1}(\upsilon_{34}) - 1 \end{aligned}$$

The model is fit in R by coding the likelihood function, where the bivariate distribution functions in (3.6) are estimated using 'pmvnorm', adopting the method of Genz [102]. The optimisation is conducted using the 'optimx' package, which is an extension of 'optim' that enables the comparison of optimisation methods. A commonly used optimisation method that can be implemented is that of Nelder and Mead [103], which is robust but relatively slow. Another is that of Fletcher [104] which is a quasi-Newton technique that updates an approximation to the inverse Hessian function and is referred

to as the 'BFGS' method. A limited memory alternative to this can also be implemented which allows for box constraints [105]. Another quasi-Newton method which uses port routines is available under the 'nlminb' option and is the best performing method in this setting in terms of accuracy and convergence rate, however also is the slowest. The nlminb function does not facilitate altering the tolerance of the computation of the Hessian separately to that of the mean parameters and so we use the 'Hessian' function in the 'numDeriv' package to obtain the Hessian matrix using Richardson extrapolation [106]. The covariance matrix of the model parameters is obtained by inverting the Hessian. In a small number of cases the Hessian may not be positive definite because of computational error, meaning that it cannot be inverted. This is rectified in these cases by using the 'near PD' function, which implements the algorithm of Higham [107] to compute the nearest positive definite matrix. Another consideration for estimation is that the optimisation performs better when the components are positively correlated, which is in agreement with suggestions in [108]. Consequently, it may be necessary to transform some of the outcomes, depending on the data structure.

3.3.4 Inference

We wish to make inference on the probability of response at time point one. Let S_i be an indicator for patient i denoting whether or not they achieved response and let S_i=1 if $Y_{i1} \leq \eta_1, Y_{i2} \leq \eta_2, Y^*_{i3} \leq \eta_3, Y^*_{i4} \leq \eta_4$, where η_k represents the dichotomisation threshold for outcome k. Therefore,

$$P(S = 1 \mid T, y_{10}, y_{20}) = \int_{-\infty}^{\eta_1} \int_{-\infty}^{\eta_2} \int_{-\infty}^{\eta_3} \int_{-\infty}^{\eta_4} f_{\mathbf{Y}^*}(\mathbf{Y}^*; T, y_{10}, y_{20}) dy_4^* dy_3^* dy_2 dy_1 \qquad (3.9)$$

where $f_{\mathbf{Y}^*}(\mathbf{Y}^*; .)$ is the multivariate normal density function for the observed and latent continuous measures. We obtain the integrand in (3.9) by using the fitted values of the parameters in the conditional mean and conditional covariance matrix in (3.7), assuming that each patient was treated and not treated. The integral in (3.9) is evaluated using the 'R2Cuba' package. Parameter estimates from these methods are maximum likelihood estimates and so we avail of asymptotic maximum likelihood theory. Note that this reliance on asymptotic theory in addition to the large number of parameters may be problematic for application in rare disease trials. The standard error estimates are obtained using the delta method.

3.3.5 Pragmatic Considerations

The potential gains from retaining more of the information are offset somewhat by the increased complexity in implementation. Some practical considerations related to the implementation of the method are discussed below.

Starting Values

One important consideration in applying the latent variable method is how to choose the starting values for the likelihood optimisation algorithm. Initially, we try setting all the starting values at zero, to determine if this is a practical solution. In this setting, the algorithm is extremely slow to converge. Furthermore, setting the thresholds at the same starting values can be problematic due to the necessary ordering. That is, if at any point the lower limits of the integration in (3.6) exceed the upper limits, the probability cannot be computed and the optimisation fails.
We are interested in choosing more appropriate starting values to reduce the chances of these computational problems as well as to reduce the computational time and increase our chances of converging at the global maximum. One suggestion for this is to use random restarts, starting the algorithm at randomly chosen start points repeatedly [109]. This is computationally intensive and often runs in to difficulty due to restrictions in place, for example that the thresholds must be ordered. Another suggestion is to use the data to inform the choice of start value. This is more challenging in this context due to the fact that we are treating the discrete outcomes as latent.
Proceeding by using the data, the parameters related to the observed continuous variables can be set by fitting separate linear models. The variance and correlation parameters related to these can be set to the values estimated using the data. Treating the ordinal variable as continuous and fitting a linear model with no intercept provides a good starting value for the BILAG treatment parameter. Fitting a linear model to the binary outcome provides poor starting values for the the binary intercept and treatment parameter, however the algorithm still performs well if only these values are poorly specified. Treating these discrete variables as continuous in order to determine the correlations performs well and provides good starting values for these parameters. Having identified proposed starting values for the mean, variance and correlation parameters the latent outcome corresponding to the ordinal BILAG measure can be simulated. Ordinal threshold starting values can be obtained by comparing the distribution of the latent measure with the corresponding observed frequencies in the

Table 3.4: Average execution time in seconds across 1000 runs for the latent variable, augmented binary and standard binary methods to produce log-odds treatment effect estimates and standard errors

Method	Elapsed	User	System
Binary	0.464	0.439	0.020
Augmented Binary	9.591	9.480	0.092
Latent Variable	4925.2	4862.3	50.42

ordinal measure. However, this technique for selecting the starting values is clearly simplistic. A more elaborate search strategy could be conducted that may substantially speed up the optimisation. This will not be investigated within this thesis and is identified as an area for further research.

Computational Time

Another factor in the application of the latent variable model is increased computational time. The average execution time to provide the outcome of interest with standard errors is shown in Table 3.4. The time of interest is the elapsed time, which expresses the wall clock time taken to fit the model, get maximum likelihood parameter estimates, obtain the probability of interest and its standard error. The standard binary estimate is obtained in less than a second and the augmented binary estimate requires approximately 10 seconds however the latent variable requires much longer, taking 4920 seconds or approximately 82 minutes.

As mentioned previously, the application of these models has been limited in the past by availability of sufficient computational power, which is now more readily available. However, applications of similar latent variable methods in the literature are still largely limited to modelling two outcomes, due to the non-linearity of execution times with increasing outcomes. This is in agreement with our findings, as modelling one outcome using the binary method is nine times faster than modelling two outcomes using the augmented binary method, whilst modelling these two outcomes is over 500 times faster than modelling four outcomes using the latent variable method. Of course these timings depend on many factors, in particular the type of outcome and the number of levels in the ordinal variable. In our case, we find the number of ordinal levels to be the most influential factor in computational time. This is due to the fact that 5 levels in the ordinal variable leads to 10 probability calculations in (3.6), however 3 levels would require the computation of 6 of these joint probabilities. Consequently,

Table 3.5: Benchmarked time in seconds of each of the processes required to fit the latent variable method to the systemic lupus erythematosus composite endpoint

Function	Elapsed time
Likelihood maximisation	3683.1
Hessian	845.6
Probability of response	3.064
Partial derivatives	266.3

the run time will be substantially increased if there are multiple ordinal levels and decreased if the discrete variables are binary.

The increased complexity means that many factors may be responsible for the much slower progression of this model fitting. As we program the likelihood, rather than using a package to do this, it may be possible to code the method more efficiently. Another factor is that due to the increased number of outcomes, the unstructured covariance matrix and the thresholds, the number of parameters to model is greatly increased from nine in the augmented binary method to 21 in the latent variable method. Searching over this 21 parameter space to find a global maximum is complex and computationally intensive and therefore relatively slow. Furthermore, for reasons discussed previously, the Hessian is calculated separately with increased tolerance and the partial derivatives are computed to obtain the standard errors. Table 3.5 shows the benchmarked times of the processes involved in fitting the latent variable method. This highlights that obtaining the maximum likelihood estimates of the parameters accounts for 77% of the required computational time. Benchmarking the optimisation process provides a clearer picture of the bottleneck. Within each iteration, the most time consuming task is the calculation of the bivariate probabilities in (3.6). It is possible to parallelise this calculation using the 'parapply' function in R however for our problem we find that this slows down the overall computation. This is due to the fact that there are many of these calculated repeatedly but that each individual calculation does not require much time. In other words, in the time it takes to redistribute the calculations to separate cores, the result is already available on one core. The true bottleneck comes from the fact that the algorithm iterates many times in order to converge. For a problem of this nature it is common to consider coding it in a low level language such as C++. Due to the fact that the process requiring the most time is the optimisation itself, we conclude that it is not worthwhile given that although the model would be written in C++ it would still have to be optimised in R using a similar algorithm.

We conclude that when fitting to one dataset in an applied problem, a computation time of 82 minutes is not infeasible. However for exploring the performance of the methods through simulation, we require an alternative. The solution we propose for this, and apply in our case, is to parallelise at a simulation level across many cores on a High Performance Computer (HPC). For 1000 simulated data sets, using 200 cores, the simulation would complete in under 7 hours.

Model Fit

Goodness-of-fit statistics are well established when fitting univariate models however the assessment of multivariate methods is more challenging. Graphical techniques that involve inspecting plots of the residuals to determine the validity of assumptions such as homoscedasticity and normality are limited in their capacity to capture the structure in more than two dimensions. Furthermore, solutions providing comparative values must add an appropriate penalty for the additional outcomes, for example a modified Akaike Information Criterion (AIC). This is exacerbated by the fact that a subset of the outcomes are latent and therefore difficult to visualise or test. One suggestion in the literature for assessing goodness-of-fit is introduced in [92] for the case when there is one continuous and one ordinal variable. This may be extended to allow for two continuous, one ordinal and one binary outcome for application in SLE, as shown below.

As before, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ be the vector of observed responses for patient i. Partitioning the observed and latent continuous measures, we let $\mathbf{Y_{cts}} = (Y_1, Y_2)$ and $\mathbf{Y_{dis}} = (Y_3, Y_4)$. Then, $\widehat{\Sigma}_{11} = \widehat{Var}(\mathbf{Y_{cts}}), \widehat{\Sigma}_{22} = \widehat{Var}(\mathbf{Y_{dis}}), \widehat{\Sigma}_{12} = \widehat{\Sigma}_{21} = \widehat{Cov}(\mathbf{Y_{cts}}, \mathbf{Y_{dis}})$. The modified Pearson residuals taking in to account the correlation between responses are shown in (3.10).

$$r_i^p = \widehat{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i) \tag{3.10}$$

where,

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = (\widehat{E}(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}))^T \tag{3.11}$$

and

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{pmatrix} \tag{3.12}$$

A Cholesky decomposition may be used to obtain $\widehat{\Sigma}^{-\frac{1}{2}}$ in (3.10). The covariance between the vector of observed continuous and observed discrete responses is shown below.

$$\begin{aligned}\Sigma_{12} =& E(\mathbf{Y_{cts}}\mathbf{Y_{dis}}) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\ =& E(\mathbf{Y_{cts}}E(\mathbf{Y_{dis}} \mid \mathbf{Y_{cts}})) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\ =& E(Y_1Y_2E(Y_3, Y_4 \mid Y_1, Y_2)) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\ =& \int_{y_1}\int_{y_2} y_1y_2 \sum_{y_3}\sum_{y_4} y_3y_4 P\left(Y_3 = w, Y_4 = k | Y_1 = y_1, Y_2 = y_2\right) f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 \\ & - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}})\end{aligned}$$

Where,

$$\begin{aligned}&P\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = y_{i1}, Y_{i2} = y_{i2}; \boldsymbol{\theta}\right) = \\ &\quad \Phi\left(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \\ &\Phi\left(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right)\end{aligned}$$

$\mu_{3|1,2}$, $\mu_{4|1,2}$ and $\Sigma_{3,4|1,2}$ are defined in (3.7). Furthermore,

$$E(\mathbf{Y_{cts}}) = \int_{y_1}\int_{y_2} y_1y_2 f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2$$

$$E(\mathbf{Y_{dis}}) = \sum_{y_3}\sum_{y_4} y_3y_4 P\left(Y_3 = w, Y_4 = k\right)$$

and

$$\begin{aligned}P(Y_{i3} = w, Y_{i4} = k) = \Phi(\tau_{w3} - \mu_3, \tau_{k4} - \mu_4; \rho_{3,4}) - \Phi(\tau_{(w-1)3} - \mu_3, \tau_{k4} - \mu_4; \rho_{3,4}) - \\ \Phi(\tau_{w3} - \mu_3, \tau_{(k-1)4} - \mu_4; \rho_{3,4}) + \Phi(\tau_{(w-1)3} - \mu_3, \tau_{(k-1)4} - \mu_4; \rho_{3,4})\end{aligned}$$

The Pearson residual is based on the Pearson goodness-of-fit statistics,

$$\chi^2_p = \sum_{i=1}^{N} \chi^2_p(\boldsymbol{Y}_i, \hat{\boldsymbol{\mu}}_i) \tag{3.13}$$

where $p = (w \times k) - 1$ with ith component,

$$\chi^2_p(\boldsymbol{Y}_i, \hat{\boldsymbol{\mu}}_i) = (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)' \widehat{\Sigma}^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i) \tag{3.14}$$

Comparing the residuals to the chi-squared value allows us to identify observations which the model does not fit well, as the residuals should follow a chi-squared distribution with p degrees of freedom if the model fits well. If there are many observations unexplained by the model then it may indicate a poor choice, which may be due to the covariance structure $\widehat{\Sigma}$ and its assumed distribution. Otherwise the joint normality of the errors may be an unreasonable assumption indicating that the latent variable model may not be appropriate. It is possible to fit latent variable models which assume a different multivariate distribution for the error terms, however this will not be investigated within this thesis.

3.4 Models for Comparison

3.4.1 Augmented Binary Method

We modify the augmented binary model presented in Chapter 1 to allow for one time point, as shown below. The baseline measures for Y_{i1} and Y_{i2} are included for comparison, as they are included in the mean structure of the latent variable method. As only one time point is modelled we can use a linear model for Y_{i1}. Note that Y_{i1} or Y_{i2} may be chosen as the continuous measure to be retained and will be determined by which is the most informative.

$$Y_{i1} = \alpha_0 + \alpha_{F1} Ti + \alpha_{F2} y_{i10} + \alpha_{F3} y_{i20} + \varepsilon_{i1} \tag{3.15}$$

In this case, the failure time binary indicator will contain information from the remaining three components. F_i is set to equal 0 if $Y_{i2} \leq \eta_2, Y_{i3}$ is Grade B-E and $Y_{i4} = 0$, otherwise the patient is a non-responder in these components and $F_i = 1$.

$$logit(Pr(F_i = 1 | T_i, y_{i10}, y_{i20}) = \beta_{F0} + \beta_{F1} T_i + \beta_{F2} y_{i10} + \beta_{F3} y_{i20} \tag{3.16}$$

F_i is modelled using one logistic regression model in this case.

3.4.2 Standard Binary Method

The standard binary method models treatment and both baseline measures, as shown below.

$$logit(Pr(S_i = 1|T_i, y_{i10}) = \psi_{F0} + \psi_{F1}T_i + \psi_{F2}y_{i10} + \psi_{F3}y_{i20} \tag{3.17}$$

The probability of response and standard errors are obtained from the logistic regression as detailed in Chapter 1.

3.5 Simulation Study

3.5.1 Data Generating Models

Initially we investigate the properties of the methods when the assumptions of the latent variable model are satisfied. The parameter values in the 'baseline' case are chosen to simulate a scenario where composite endpoints are typically recommended for use. Namely, that all four components drive response and items are correlated but not so highly that the composite becomes redundant. The parameter values have been informed by the MUSE trial dataset, in particular the correlation structure. The response probability in the control arm is 0.275 and in the treatment arm is 0.381, resulting in an odds ratio equal to 1.6. The parameter values selected for the model in (3.1) are shown below.

Baseline: $N = 300, \alpha_0 = -4.9, \alpha_1 = -0.28, \alpha_2 = -0.5, \beta_0 = -1.2, \beta_1 = -0.35, \beta_2 = -0.5, \gamma_1 = -0.24, \psi_0 = -0.2, \psi_1 = -0.18, \sigma_1 = \sigma_2 = 1, \rho_{12} = 0.5, \rho_{13} = \rho_{24} = 0.35, \rho_{14} = 0.25,\ \rho_{23} = 0.4, \rho_{34} = 0.3, \tau_{13} = -1, \tau_{23} = -0.1, \tau_{33} = 0.45, \tau_{43} = 1.3, \eta_1 = -4, \eta_2 = -0.6, \eta_3 = 0.45, \eta_4 = 0$

From this baseline case, we vary parameters to determine how the methods behave under various scenarios, the values of which are detailed in Table 3.6.

3.5.2 Performance Measures

The performance measures and Monte Carlo standard errors (MCSE) are shown in Table 3.7. More details can be found in [110].

Table 3.6: Parameter values for the simulated scenarios which investigate the effect of varying the responder threshold η_1, changing the components driving response and differing treatment effects on the performance of the latent variable, augmented binary and standard binary methods for the systemic lupus erythematosus composite endpoint

Scenario	Parameters	Investigates
$\eta_1 = -2$	$\eta_1 = -2$	100% of patients respond in Y_1
$\eta_1 = -3$	$\eta_1 = -3$	96% of patients respond in Y_1
$\eta_1 = -4$	$\eta_1 = -4$	82% of patients respond in Y_1
$\eta_1 = -5$	$\eta_1 = -5$	52% of patients respond in Y_1
$\eta_1 = -6$	$\eta_1 = -6$	20% patients respond in Y_1
Y_1, Y_4	$\eta_1 = -5, \eta_2 = 2, \eta_3 = 2$	Continuous and binary variable driving response
Y_4	$\eta_1 = -2, \eta_2 = 2, \eta_3 = 2$	Binary variable driving response
Y_1, Y_2, Y_3	$\eta_4 = 2$	Two continuous and ordinal drive response
Treat case 1	$\alpha_0 = -4.9, \alpha_1 = -0.09, \beta_0 = -1.2, \beta_1 = -0.11, \gamma_1 = -0.145, \psi_0 = -0.2, \psi_1 = -0.07$	Odds ratio = 1.217
Treat case 2	$\alpha_0 = -4.9, \alpha_1 = -0.20, \beta_0 = -1.2, \beta_1 = -0.25, \gamma_1 = -0.2, \psi_0 = -0.2, \psi_1 = -0.12$	Odds ratio = 1.426
Treat case 3	$\alpha_0 = -4.9, \alpha_1 = -0.30, \beta_0 = -1.2, \beta_1 = -0.50, \gamma_1 = -0.3, \psi_0 = -0.2, \psi_1 = -0.22$	Odds ratio = 1.794
Treat case 4	$\alpha_0 = -4.9, \alpha_1 = -0.32, \beta_0 = -1.2, \beta_1 = -0.65, \gamma_1 = -0.39, \psi_0 = -0.2, \psi_1 = -0.27$	Odds ratio = 2.007
Treat case 5	$\alpha_0 = -4.9, \alpha_1 = -0.33, \beta_0 = -1.2, \beta_1 = -0.72, \gamma_1 = -0.45, \psi_0 = -0.2, \psi_1 = -0.33$	Odds ratio = 2.198
Null	$\alpha_1 = \beta_1 = \gamma_1 = \psi_1 = 0$	Type I error rate

Table 3.7: Performance measures and Monte Carlo standard errors used to assess the behaviour of the latent variable, augmented binary and binary methods in a simulation study for the systemic lupus erythematosus composite endpoint

Performance measure	Estimate	MCSE
Bias	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \hat{\delta}_j - \delta$	$\sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{j=1}^{n_{sim}} (\hat{\delta}_j - \bar{\delta})^2}$
Coverage	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(\hat{\delta}_{low,j} \leq \delta \leq \hat{\delta}_{upp,j})$	$\sqrt{\frac{\widehat{cov.}(1-\widehat{cov.})}{n_{sim}}}$
Bias-corrected coverage	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(\hat{\delta}_{low,j} \leq \bar{\delta} \leq \hat{\delta}_{upp,j})$	$\sqrt{\frac{\widehat{BEcov.}(1-\widehat{BEcov.})}{n_{sim}}}$
Power	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(p_j < cv)$	$\sqrt{\frac{\widehat{Power}(1-\widehat{Power})}{n_{sim}}}$
MSE	$\sum_{j=1}^{n_{sim}} (\hat{\delta}_j - \delta)^2$	$\sqrt{\frac{\sum_{j=1}^{n_{sim}} [(\hat{\delta}_j-\delta)^2-\widehat{MSE}]^2}{n_{sim}(n_{sim}-1)}}$
Empirical SE	$\sqrt{\frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} (\hat{\delta}_i - \bar{\delta})^2}$	$\frac{\widehat{EmpSE}}{\sqrt{2(n_{sim}-1)}}$
Model SE	$\sqrt{\frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} \widehat{Var}(\hat{\delta}_j)}$	$\sqrt{\frac{\widehat{Var}[\widehat{Var}(\hat{\delta})]}{4n_{sim}\widehat{ModSE}^2}}$†
Relative precision A vs. B	$\frac{\widehat{Var}(\hat{\delta}_j)_B}{\widehat{Var}(\hat{\delta}_j)_A}$	-

$\hat{\delta}_j$: estimated log-odds treatment effect in simulated data j
$\bar{\delta}$: mean log-odds treatment effect over n_{sim} datasets
$\hat{\delta}_{low,j}, \hat{\delta}_{upp,j}$ lower and upper limit of confidence interval for iteration j
† $\widehat{Var}[\widehat{Var}(\hat{\delta})] = \frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} \{\widehat{Var}(\hat{\delta}_i) - \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \widehat{Var}(\hat{\delta}_j)\}^2$

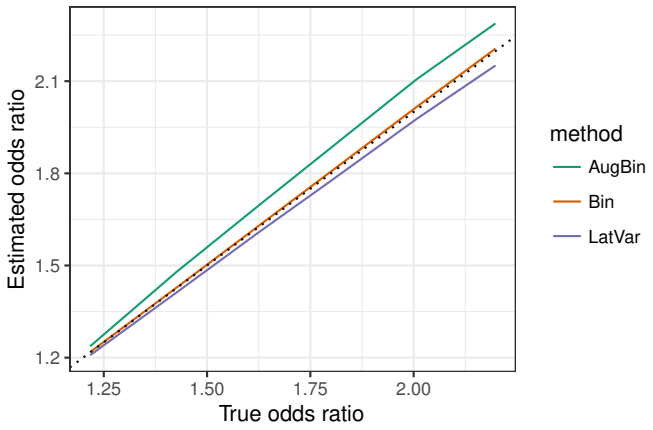


Figure 3.3: Bias reported from the latent variable method, augmented binary method and standard binary method when n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

3.5.3 Findings

The simulation results for the different scenarios are presented in Tables 3.8 - 3.15. In what follows, we discuss the most interesting and relevant findings in more detail.

3.5.3.1 Varying Treatment Effect

An important property for an estimator in clinical trials is that it is unbiased. Figure 3.3 shows the bias of the methods as the treatment effect varies. The standard binary method is unbiased, as we would expect for a logistic regression in a large sample. The latent variable method is unbiased for smaller treatment effects but a small bias towards the null is introduced as the treatment effect increases. The augmented binary method is biased away from the null in this setting and the bias increases as the treatment effect increases. Given that this performance is worse than is suggested from previous applications of the augmented binary method in [35, 36], this would suggest that the augmented binary method may be biased when the true data generating mechanism is more similar to the latent variable model.

Figure 3.4 shows the coverage of the methods. The binary method has approximately nominal coverage. For smaller treatment effects the latent variable method has nominal coverage, however the coverage probability decreases as the treatment effect increases. The augmented binary method has coverage of approximately 0.91, which also decreases when the treatment effect increases. In order to diagnose this under-coverage in the

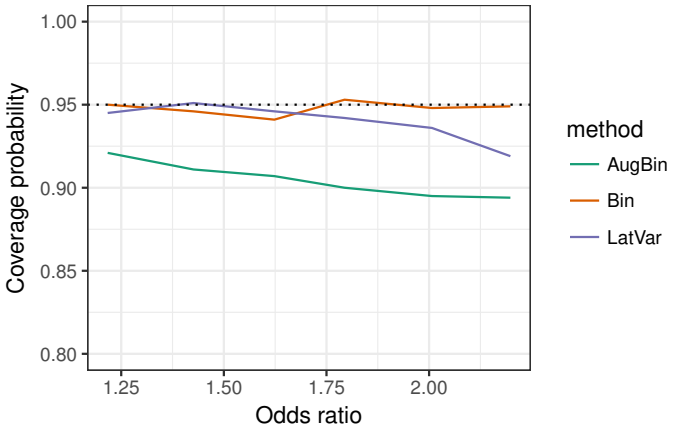


Figure 3.4: Coverage probability reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

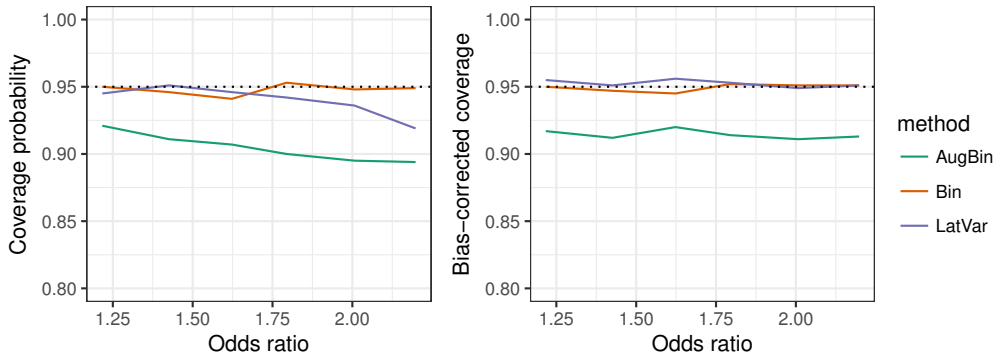


Figure 3.5: Coverage probability (left) and bias-corrected coverage probability (right) reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

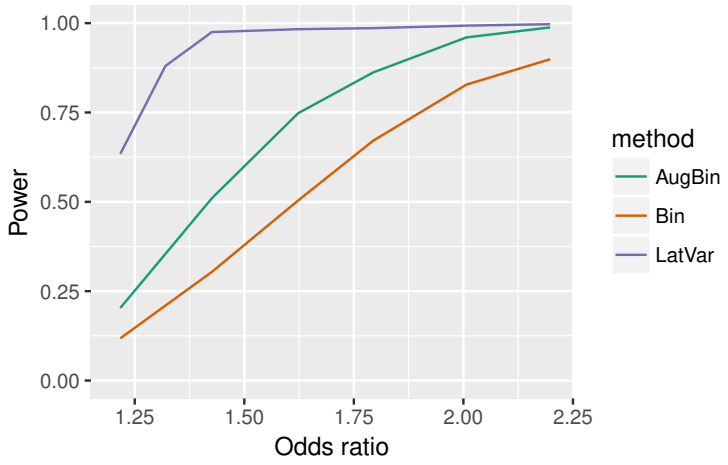


Figure 3.6: Statistical power reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

joint modelling methods we can look at bias-corrected coverage, as recommended in [110]. Figure 3.5 shows both the coverage and bias-corrected coverage. The properties of the standard binary method remain unchanged. The bias-corrected coverage of the latent variable method is 0.95, which indicates that any under-coverage is due to the bias present. This is not true for the augmented binary method which shows small improvements in bias-corrected coverage so that it is approximately 0.92. This indicates that under-coverage is present in this method due to bias as well as other factors, which is likely to be model misspecification. The power of the three methods is shown in Figure 3.6. The performance of the binary and augmented binary method is as we would expect based on previous findings in [110] and the latent variable method offers much higher power. In this setting it has close to 100% power for odds ratios larger than 1.6, an effect that is plausible to observe in a trial.

To investigate improvements in efficiency we consider the relative precision of each of the methods versus another. Obtaining the relative precision in each of the simulated data sets and plotting the median, 10th centile and 90th centile facilitates an intuitive interpretation, as illustrated in Figure 3.7. The augmented binary method is 1.5 times as precise as the binary method and consistently so across the different odds ratios considered. The latent variable method offers much larger gains in precision over both the augmented and standard binary methods however the variability in precision gains is much larger than those demonstrated with the augmented binary method.

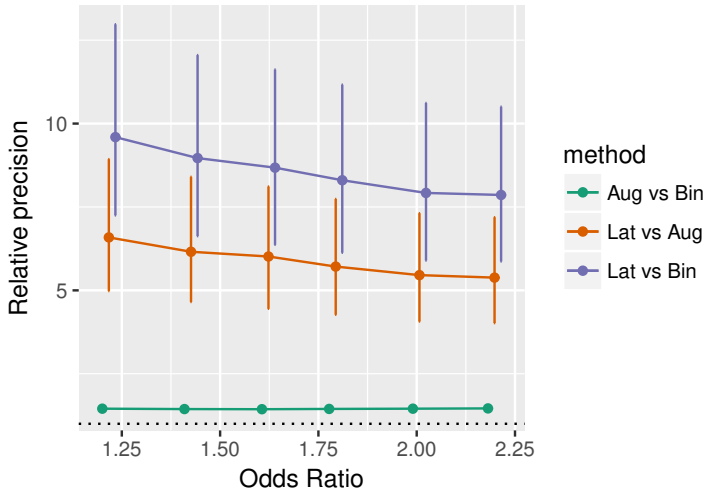


Figure 3.7: Median, 10th centile and 90th centile estimated relative precision reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

In this setting, the latent variable method is approximately 8 times as precise as the binary method. These findings have indicated that the standard binary method has the smallest bias and that the latent variable method has the smallest variance. The mean squared error (MSE) provides a combined measure of bias and variance. Figure 3.8 shows the MSE of the three methods as the treatment effect varies. The MSE for the standard and augmented binary methods is approximately 6.5 times that of the latent variable method. However, this measure should be interpreted with care due to the fact that the MSE is more sensitive to the sample size than comparisons of bias or empirical SE alone [110].

3.5.3.2 Varying η_1

To understand more about the precision performance of the augmented binary method in particular, we vary the responder threshold η_1 to change the proportion of responders in that outcome. Figure 3.9 shows the density of the Y_1 variable and the relative precision of the methods as the responder threshold varies. The precision gains from the augmented binary method diminish as the threshold increases, which is intuitive as improvements in efficiency fall as the continuous component becomes less responsible for driving response. It is interesting to note that all precision gains are lost for any thresholds above -4. Therefore, even when 20% of patients are non-responders, all

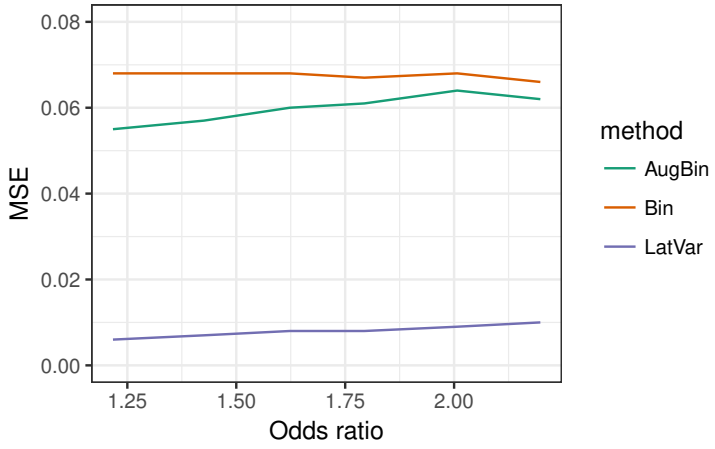


Figure 3.8: Mean Squared Error (MSE) reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

efficiency gains are lost. The percentage of responders needed to improve efficiency using the augmented binary method will of course depend on the correlation structure employed. Due to the additional information in the other components, the latent variable method is five times as precise as the other methods. Figure 3.10 shows the power of the methods as the Y_1 dichotomisation threshold changes. The power of the latent variable reduces slightly as the proportion of responders increases. Although this loss in power appears negligible, it is worth noting that a power of 0.999 vs 0.998 may be substantial in terms of sample size required. There are power gains available from employing the augmented binary method, even when only a very small proportion of patients are non-responders.

3.5.3.3 Components Contributing to Response

An important consideration when investigating performance is how the precision changes when different combinations of outcomes are responsible for driving response. Figure 3.11 shows boxplots of the relative precision for the methods when response is driven by (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_4) and (Y_4). When all four components contribute to response, the latent variable method offers large precision gains over the other two methods. The latent variable method always outperforms the other methods in this setting however the variability in the magnitude of gains offered is large. The median result is that the treatment effect reported by the latent variable

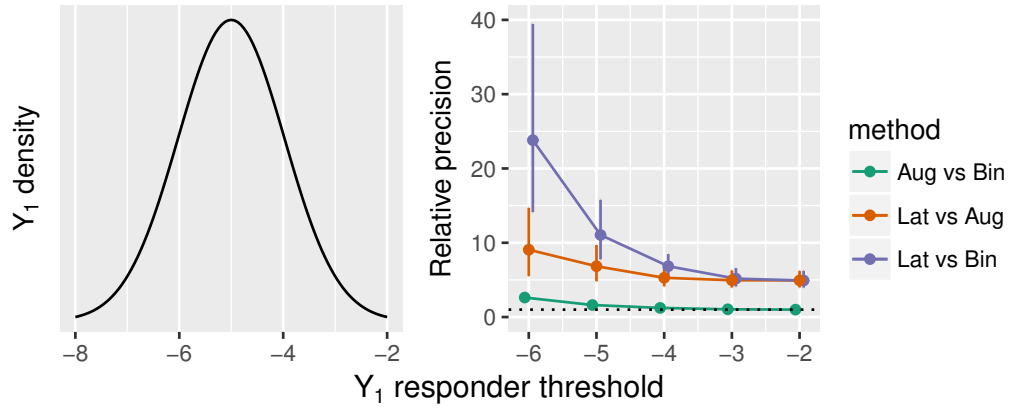


Figure 3.9: Density of continuous Y_1 variable (left) and estimated relative precision of augmented binary versus standard binary method, latent variable versus augmented binary method and latent variable versus standard binary method as the Y_1 responder threshold η_1 varies between $\eta_1 = -6$ and $\eta_1 = -2$ (right) for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

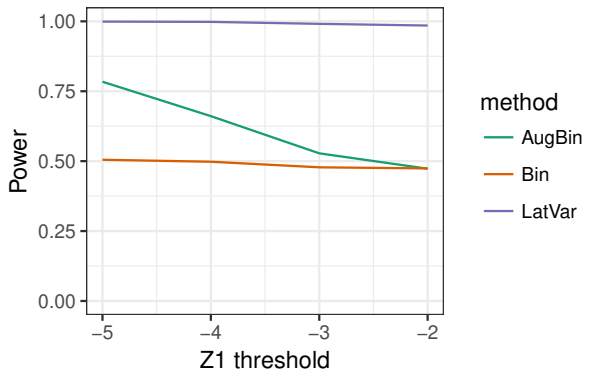


Figure 3.10: Statistical power of latent variable method, augmented binary method and standard binary method as the Y_1 responder threshold η_1 varies between $\eta_1 = -5$ and $\eta_1 = -2$ (right) for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

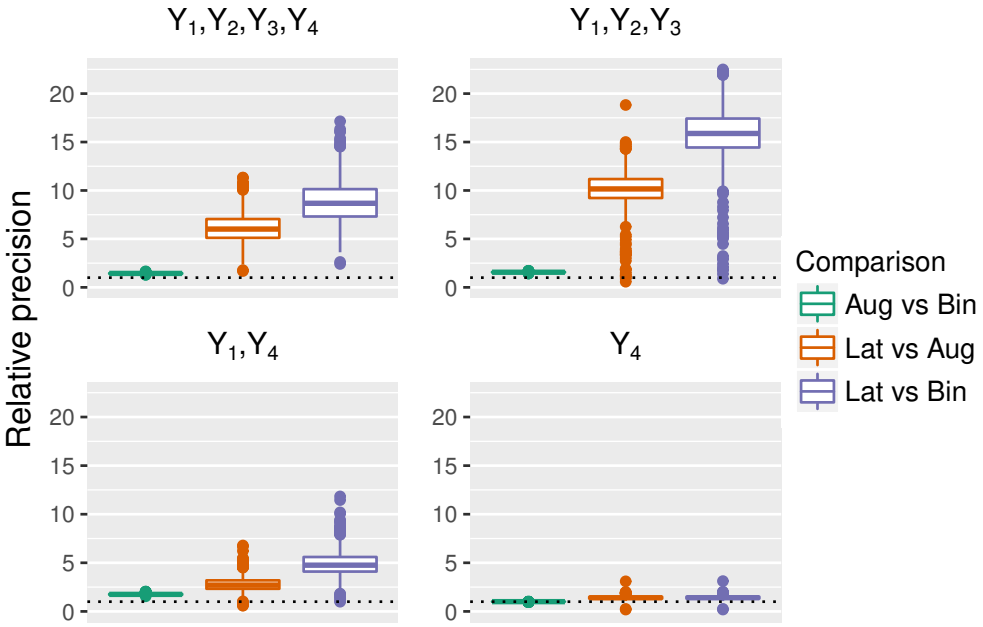


Figure 3.11: Estimated relative precision gains from augmented binary versus standard binary method, latent variable versus augmented binary method and latent variable versus standard binary method when different combinations of components drive response. Response driven by (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_4) and (Y_4) where Y_1 and Y_2 are continuous, Y_3 is ordinal, Y_4 is binary for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

method is eight times as precise as that reported by the binary method and six times as precise as the augmented binary method. If response is driven by (Y_1, Y_2, Y_3) then the relative gains for the latent variable method are larger, however note that in less than 2% of cases the treatment effect is reported equally or less precisely than both of the other methods. The findings are similar for when response is driven by (Y_1, Y_4), however the median gains are smaller. The treatment effect is reported five times more precisely from the latent variable method than the binary method here. Note that as the augmented binary method models the relevant components, it still performs well and again better than the latent variable method in a very small number of cases. When binary Y_4 determines response, the augmented binary method offers no improvement in precision whereas the latent variable method is approximately 1.5 times more precise. It is clear from the results that the magnitude of the precision gain from the latent variable method is highly dependent on the structure of the data.

3.5.3.4 Probability of Response in Each Arm

We are also interested in how well the probability of response in each arm is estimated. Figure 3.12 shows that the binary method estimates the probability of response in each arm well. The latent variable method slightly underestimates the probability of response as the treatment effect increases. The augmented binary method largely underestimates the probability of response in both arms. Figure 3.13 shows the estimation of the probability of response in each arm as the treatment effect varies. The standard binary method estimates the probability of response in both arms perfectly for all treatment effects considered. The latent variable method is underestimating the probability of response by less than 0.01 in the control arm however underestimates the probability in the treatment arm by 0.005-0.02, with the magnitude of the underestimation increasing with increasing treatment effect. This explains the increasing bias as the true treatment effect increases. The augmented binary method underestimates the probability of response in both arms by approximately 0.05 for all treatment effects considered.

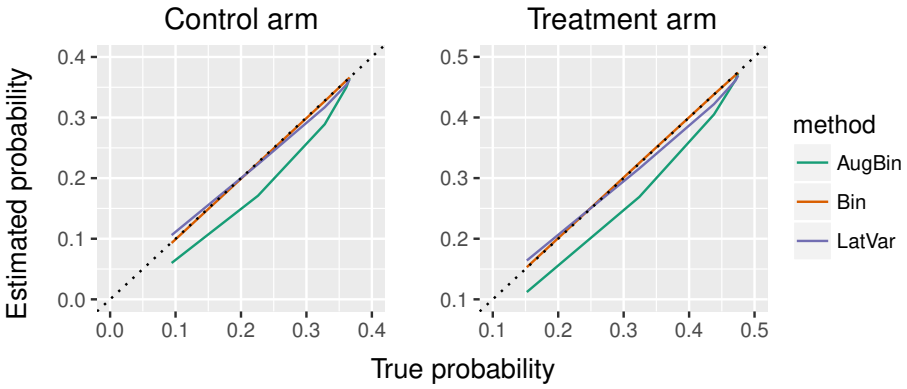


Figure 3.12: Estimation of the probability of response in each arm by the latent variable method, augmented binary method and standard binary method for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

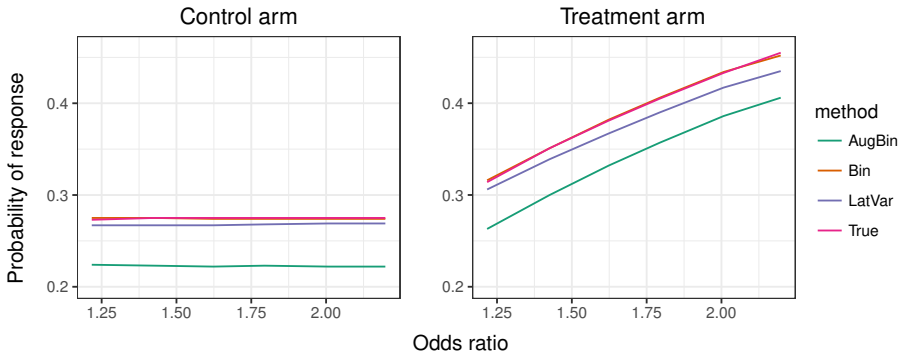


Figure 3.13: Probability of response in each arm reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

Table 3.8: Median estimates of the operating characteristics (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
Bias	Baseline	-0.017 (0.085)	0.070 (0.239)	0.004 (0.251)
	$\eta_1 = -2$	-0.008 (0.080)	0.004 (0.245)	0.000 (0.245)
	$\eta_1 = -3$	-0.010 (0.086)	0.018 (0.243)	0.000 (0.244)
	$\eta_1 = -4$	-0.014 (0.079)	0.049 (0.245)	0.003 (0.247)
	$\eta_1 = -5$	-0.022 (0.090)	0.091 (0.256)	0.004 (0.279)
	$\eta_1 = -6$	-0.040 (0.092)	0.137 (0.292)	0.006 (0.366)
	Y_1, Y_4	-0.026 (0.090)	0.031 (0.200)	0.005 (0.251)
	Y_4	-0.003 (0.081)	-0.010 (0.238)	-0.006 (0.237)
	Y_1, Y_2, Y_3	-0.020 (0.079)	0.064 (0.222)	0.004 (0.236)
	Treat case1	-0.009 (0.071)	0.020 (0.234)	0.001 (0.260)
	Treat case2	-0.014 (0.080)	0.053 (0.238)	0.002 (0.260)
	Treat case3	-0.024 (0.079)	0.083 (0.247)	0.006 (0.258)
	Treat case4	-0.030 (0.088)	0.099 (0.253)	0.009 (0.261)
	Treat case5	-0.047 (0.087)	0.090 (0.250)	0.008 (0.257)
	Null	-0.003 (0.036)	-0.004 (0.104)	0.000 (0.116)
Coverage	Baseline	0.949 (0.004)	0.907 (0.005)	0.944 (0.004)
	$\eta_1 = -2$	0.951 (0.008)	0.945 (0.007)	0.949 (0.007)
	$\eta_1 = -3$	0.948 (0.008)	0.938 (0.008)	0.948 (0.007)
	$\eta_1 = -4$	0.949 (0.008)	0.913 (0.009)	0.946 (0.008)
	$\theta_1 = -5$	0.946 (0.009)	0.883 (0.011)	0.948 (0.009)
	$\eta_1 = -6$	0.931 (0.012)	0.858 (0.013)	0.954 (0.008)
	Y_1, Y_4	0.952 (0.008)	0.918 (0.009)	0.945 (0.008)
	Y_4	0.948 (0.008)	0.952 (0.007)	0.952 (0.007)
	Y_1, Y_2, Y_3	0.953 (0.008)	0.893 (0.010)	0.949 (0.008)
	Treat case1	0.951 (0.007)	0.921 (0.009)	0.950 (0.008)
	Treat case2	0.947 (0.007)	0.911 (0.009)	0.946 (0.008)
	Treat case3	0.940 (0.008)	0.900 (0.010)	0.951 (0.008)
	Treat case4	0.931 (0.009)	0.895 (0.010)	0.948 (0.008)
	Treat case5	0.919 (0.009)	0.894 (0.010)	0.949 (0.008)
	Null	0.951 (0.003)	0.921 (0.004)	0.948 (0.003)

Table 3.9: Median estimates of the operating characteristics (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
Bias-corrected coverage	Baseline	0.956 (0.003)	0.920 (0.004)	0.945 (0.004)
	$\eta_1 = -2$	0.951 (0.008)	0.943 (0.007)	0.949 (0.007)
	$\eta_1 = -3$	0.949 (0.008)	0.942 (0.008)	0.948 (0.007)
	$\eta_1 = -4$	0.951 (0.008)	0.920 (0.009)	0.946 (0.008)
	$\eta_1 = -5$	0.952 (0.009)	0.897 (0.011)	0.949 (0.009)
	$\eta_1 = -6$	0.958 (0.012)	0.903 (0.013)	0.955 (0.008)
	Y_1, Y_4	0.954 (0.008)	0.923 (0.009)	0.948 (0.008)
	Y_4	0.949 (0.008)	0.950 (0.007)	0.954 (0.007)
	Y_1, Y_2, Y_3	0.961 (0.008)	0.904 (0.010)	0.950 (0.008)
	Treat case1	0.955 (0.007)	0.917 (0.009)	0.950 (0.008)
	Treat case2	0.951 (0.007)	0.912 (0.009)	0.947 (0.008)
	Treat case3	0.953 (0.008)	0.914 (0.010)	0.952 (0.008)
	Treat case4	0.949 (0.009)	0.911 (0.010)	0.951 (0.008)
	Treat case5	0.951 (0.009)	0.913 (0.010)	0.951 (0.008)
	Null	0.952 (0.003)	0.925 (0.004)	0.948 (0.003)
Power	Baseline	0.983 (0.002)	0.748 (0.007)	0.504 (0.008)
	$\eta_1 = -2$	0.976 (0.004)	0.473 (0.016)	0.474 (0.016)
	$\eta_1 = -3$	0.979 (0.003)	0.528 (0.016)	0.478 (0.016)
	$\eta_1 = -4$	0.981 (0.002)	0.661 (0.015)	0.498 (0.016)
	$\eta_1 = -5$	0.987(0.001)	0.784 (0.015)	0.505 (0.018)
	$\eta_1 = -6$	0.990 (0.001)	0.835 (0.014)	0.359 (0.018)
	Y_1, Y_4	0.920 (0.008)	0.699 (0.015)	0.434 (0.016)
	Y_4	0.316 (0.015)	0.228 (0.013)	0.224 (0.013)
	Y_1, Y_2, Y_3	0.993 (0.001)	0.845 (0.012)	0.595 (0.016)
	Treat case1	0.634 (0.016)	0.203 (0.013)	0.118 (0.011)
	Treat case2	0.975 (0.004)	0.510 (0.016)	0.304 (0.015)
	Treat case3	0.986 (0.001)	0.862 (0.011)	0.671 (0.015)
	Treat case4	0.993 (0.001)	0.960 (0.006)	0.828 (0.012)
	Treat case5	0.997 (0.001)	0.988 (0.004)	0.899 (0.010)

Table 3.10: Median estimates of the mean squared error (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

Scenario	Method		
	Latent Variable	Augmented Binary	Binary
Baseline	0.007 (<0.001)	0.057 (<0.001)	0.063 (<0.001)
$\eta_1 = -2$	0.012 (<0.001)	0.060 (<0.001)	0.060 (0.003)
$\eta_1 = -3$	0.011 (<0.001)	0.059 (0.003)	0.059 (0.003)
$\eta_1 = -4$	0.009 (<0.001)	0.060 (0.003)	0.061 (0.003)
$\eta_1 = -5$	0.007 (<0.001)	0.066 (0.003)	0.078 (0.004)
$\eta_1 = -6$	0.007 (<0.001)	0.085 (0.005)	0.134 (0.008)
Y_1, Y_4	0.013 (<0.001)	0.040 (0.002)	0.063 (0.003)
Y_4	0.041 (0.002)	0.057 (0.002)	0.056 (0.002)
Y_1, Y_2, Y_3	0.004 (<0.001)	0.049 (0.002)	0.056 (0.003)
Treat case1	0.006 (<0.001)	0.055 (0.002)	0.068 (0.003)
Treat case2	0.007 (<0.001)	0.057 (0.003)	0.068 (0.003)
Treat case3	0.008 (<0.001)	0.061 (0.003)	0.067 (0.003)
Treat case4	0.009 (<0.001)	0.064 (0.003)	0.068 (0.003)
Treat case5	0.011 (<0.001)	0.062 (0.003)	0.066 (0.003)
Null	0.006 (<0.001)	0.055 (0.001)	0.068 (0.001)

Table 3.11: Median estimates of the empirical standard error and model standard error (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
EmpSE	Baseline	0.084 (0.001)	0.229 (0.003)	0.251 (0.003)
	$\eta_1 = -2$	0.108 (0.002)	0.245 (0.006)	0.245 (0.006)
	$\eta_1 = -3$	0.106 (0.002)	0.243 (0.006)	0.244 (0.006)
	$\eta_1 = -4$	0.096 (0.002)	0.240 (0.006)	0.247 (0.006)
	$\eta_1 = -5$	0.079 (0.002)	0.240 (0.006)	0.279 (0.007)
	$\eta_1 = -6$	0.074 (0.002)	0.257 (0.007)	0.363 (0.010)
	Y_1, Y_4	0.111 (0.003)	0.198 (0.004)	0.251 (0.006)
	Y_4	0.203 (0.005)	0.238 (0.005)	0.237 (0.005)
	Y_1, Y_2, Y_3	0.055 (0.001)	0.213 (0.005)	0.236 (0.005)
	Treat case1	0.080 (0.002)	0.233 (0.005)	0.260 (0.006)
	Treat case2	0.084 (0.002)	0.232 (0.005)	0.260 (0.006)
	Treat case3	0.088 (0.002)	0.233 (0.005)	0.258 (0.006)
	Treat case4	0.089 (0.002)	0.233 (0.005)	0.260 (0.006)
	Treat case5	0.088 (0.002)	0.233 (0.005)	0.257 (0.006)
	Null	0.080 (0.001)	0.234 (0.002)	0.260 (0.003)
ModSE	Baseline	0.008 (0.004)	0.042 (0.001)	0.061 (0.001)
	$\eta_1 = -2$	0.011 (0.003)	0.055 (0.001)	0.055 (0.001)
	$\eta_1 = -3$	0.011 (0.003)	0.053 (0.001)	0.056 (0.001)
	$\eta_1 = -4$	0.009 (0.004)	0.046 (0.001)	0.057 (0.001)
	$\eta_1 = -5$	0.006 (0.006)	0.041 (0.001)	0.067 (0.001)
	$\eta_1 = -6$	0.006 (0.012)	0.048 (0.002)	0.130 (0.003)
	Y_1, Y_4	0.012 (0.005)	0.032 (0.001)	0.056 (0.001)
	Y_4	0.041 (0.004)	0.056 (0.001)	0.057 (0.001)
	Y_1, Y_2, Y_3	0.004 (0.012)	0.034 (0.001)	0.053 (0.001)
	Treat case1	0.007 (0.007)	0.044 (0.001)	0.063 (0.001)
	Treat case2	0.007 (0.004)	0.043 (0.001)	0.062 (0.001)
	Treat case3	0.008 (0.004)	0.042 (0.001)	0.060 (0.001)
	Treat case4	0.008 (0.004)	0.041 (0.001)	0.060 (0.001)
	Treat case5	0.008 (0.005)	0.041 (0.001)	0.059 (0.001)
	Null	0.007 (0.002)	0.046 (0.001)	0.066 (0.001)

Table 3.12: Median estimate of the probability of response from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

	Pr(resp$\mid T = 0$)				Pr(resp$\mid T = 1$)			
Scenario	True	Lat Var	Aug Bin	Bin	True	Lat Var	Aug Bin	Bin
Baseline	0.275	0.267	0.222	0.274	0.381	0.367	0.332	0.382
$\eta_1 = -2$	0.366	0.363	0.365	0.366	0.475	0.469	0.473	0.474
$\eta_1 = -3$	0.361	0.354	0.350	0.361	0.471	0.461	0.462	0.470
$\eta_1 = -4$	0.328	0.317	0.289	0.327	0.438	0.422	0.405	0.438
$\eta_1 = -5$	0.226	0.223	0.171	0.225	0.324	0.316	0.269	0.325
$\eta_1 = -6$	0.094	0.106	0.060	0.093	0.152	0.164	0.112	0.153
Y_1, Y_4	0.302	0.302	0.269	0.303	0.397	0.390	0.366	0.399
Y_4	0.571	0.567	0.577	0.578	0.643	0.638	0.645	0.646
Y_1, Y_2, Y_3	0.388	0.375	0.325	0.389	0.512	0.491	0.459	0.513
Treat case1	0.273	0.267	0.224	0.275	0.314	0.306	0.263	0.316
Treat case2	0.275	0.267	0.223	0.275	0.351	0.339	0.300	0.351
Treat case3	0.275	0.268	0.223	0.274	0.405	0.390	0.357	0.406
Treat case4	0.275	0.269	0.222	0.274	0.433	0.417	0.386	0.434
Treat case5	0.275	0.269	0.222	0.274	0.455	0.435	0.406	0.452
Null	0.275	0.269	0.222	0.274	0.275	0.269	0.221	0.274

Table 3.13: Median estimates of the odds ratio treatment effect estimate (95% confidence intervals in parentheses) from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

	Treatment effect			
Scenario	True	Lat Var	Aug Bin	Bin
Baseline	1.623	1.592 (1.348, 1.880)	1.744 (1.165, 2.611)	1.643 (1.013, 2.666)
$\eta_1 = -2$	1.566	1.555 (1.263, 1.914)	1.573 (0.991, 2.496)	1.567 (0.988, 2.486)
$\eta_1 = -3$	1.573	1.560 (1.273, 1.912)	1.602 (1.019, 2.518)	1.575 (0.992, 2.500)
$\eta_1 = -4$	1.600	1.578 (1.312, 1.897)	1.680 (1.101, 2.563)	1.612 (1.009, 2.574)
$\eta_1 = -5$	1.642	1.606 (1.375, 1.874)	1.797 (1.206, 2.677)	1.665 (1.002, 2.768)
$\eta_1 = -6$	1.721	1.653 (1.425, 1.918)	1.973 (1.282, 3.037)	1.800 (0.890, 3.638)
Y_1, Y_4	1.522	1.475 (1.189, 1.829)	1.570 (1.105, 2.230)	1.530 (0.960, 2.436)
Y_4	1.353	1.348 (0.910, 1.998)	1.339 (0.841, 2.133)	1.335 (0.838, 2.128)
Y_1, Y_2, Y_3	1.655	1.613 (1.436, 1.811)	1.582 (1.230, 2.529)	1.670 (1.059, 2.609)
Treat case1	1.217	1.207 (1.028, 1.417)	1.242 (0.824, 1.870)	1.218 (0.744, 1.996)
Treat case2	1.426	1.407 (1.194, 1.657)	1.503 (1.001, 2.256)	1.437 (0.882, 2.340)
Treat case3	1.794	1.751 (1.480, 2.073)	1.947 (1.305, 2.907)	1.816 (1.122, 2.938)
Treat case4	2.007	1.948 (1.642, 2.312)	2.215 (1.489, 3.296)	2.041 (1.264, 3.295)
Treat case5	2.198	2.097 (1.766, 2.490)	2.406 (1.620, 3.572)	2.203 (1.366, 3.552)
Null	1.000	0.996 (0.828, 1.197)	0.995 (0.655, 1.513)	1.000 (0.604, 1.656)

Table 3.14: Median estimates of the relative precision [with 10th centile and 90th centile values in parentheses] from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

	Treatment effect		
Scenario	Lat Var vs Bin	Lat Var vs Aug Bin	Aug Bin vs Bin
Baseline	8.678 [6.365, 11.62]	6.015 [4.447, 8.112]	1.437 [1.381, 1.499]
$\eta_1 = -2$	4.919 [3.962, 6.225]	4.935 [3.973, 6.220]	1.000 [0.989, 1.018]
$\eta_1 = -3$	5.172 [4.105, 6.584]	4.952 [3.967, 6.285]	1.043 [1.028, 1.063]
$\eta_1 = -4$	6.483 [5.058, 8.468]	5.283 [4.125, 6.859]	1.231 [1.192, 1.279]
$\eta_1 = -5$	11.05 [7.763, 15.78]	6.838 [4.806, 9.691]	1.618 [1.532, 1.716]
$\eta_1 = -6$	23.81 [14.12, 39.46]	9.059 [5.507, 14.71]	2.627 [2.268, 2.991]
Y_1, Y_4	4.756 [3.595, 6.563]	2.709 [2.028, 3.763]	1.753 [1.672, 1.855]
Y_4	1.407 [1.257, 1.617]	1.402 [1.251, 1.605]	1.005 [0.993, 1.024]
Y_1, Y_2, Y_3	15.89 [13.05, 19.09]	10.16 [8.321, 12.17]	1.562 [1.508, 1.623]
Treat case1	9.594 [7.247, 12.98]	6.587 [4.980, 8.933]	1.453 [1.391, 1.522]
Treat case2	8.965 [6.625, 12.05]	6.157 [4.652, 8.402]	1.441 [1.384, 1.499]
Treat case3	8.301 [6.127, 11.17]	5.713 [4.269, 7.739]	1.445 [1.387, 1.510]
Treat case4	7.920 [5.894, 10.62]	5.457 [4.066, 7.312]	1.453 [1.397, 1.513]
Treat case5	7.860 [5.865, 10.51]	5.382 [4.019, 7.191]	1.461 [1.404, 1.522]
Null	9.939 [7.437, 13.35]	6.833 [5.099, 9.127]	1.455 [1.386, 1.521]

Table 3.15: Median confidence interval width (standard deviation in parentheses) for log-odds treatment effects reported from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) applied to the systemic lupus erythematosus endpoint when the assumptions of the latent variable method are satisfied, N=300 and $n_{sim} = 5000$

	Average CI width			% reduction CI width	
Scenario	Lat Var	Aug Bin	Bin	Lat Var	Aug Bin
Baseline	0.328 (0.04)	0.806 (0.02)	0.966 (0.03)	66.00	16.50
$\eta_1 = -2$	0.416 (0.04)	0.922 (0.01)	0.921 (0.01)	54.85	-0.11
$\eta_1 = -3$	0.407 (0.04)	0.903 (0.01)	0.923 (0.01)	55.93	2.11
$\eta_1 = -4$	0.368 (0.04)	0.844 (0.02)	0.936 (0.02)	60.66	9.82
$\eta_1 = -5$	0.305 (0.05)	0.797 (0.03)	1.013 (0.04)	69.89	21.39
$\eta_1 = -6$	0.287 (0.07)	0.862 (0.04)	1.394 (0.12)	79.44	38.17
Y_1, Y_4	0.427 (0.06)	0.703 (0.02)	0.930 (0.02)	54.13	24.45
Y_4	0.784 (0.07)	0.929 (0.02)	0.930 (0.02)	15.66	0.16
Y_1, Y_2, Y_3	0.226 (0.05)	0.721 (0.02)	0.901 (0.01)	74.90	19.92
Treat case1	0.319 (0.05)	0.818 (0.02)	0.984 (0.03)	67.61	16.88
Treat case2	0.326 (0.04)	0.811 (0.02)	0.973 (0.03)	66.50	16.62
Treat case3	0.333 (0.04)	0.799 (0.02)	0.960 (0.03)	65.29	16.76
Treat case4	0.339 (0.04)	0.793 (0.02)	0.956 (0.03)	64.53	16.96
Treat case5	0.339 (0.04)	0.790 (0.02)	0.954 (0.03)	64.50	17.22
Null	0.320 (0.06)	0.836 (0.03)	1.006 (0.03)	68.25	16.90

3.6 Sensitivity Analysis

The latent variable method has been shown to perform well when the assumptions of joint normality are satisfied. It is important to understand the robustness of the method when these assumptions are not satisfied, especially as the joint normality assumption cannot be tested in this instance. A distribution which should sufficiently represent these deviations from normality is the skew-normal distribution.

3.6.1 Multivariate Skew-Normal Distribution

The multivariate skew-normal is an extension of the univariate skew normal distribution introduced by Azzalini and Dalla Valle [111], which they define as follows. A random vector $\mathbf{Y}=(Y_1, ..., Y_k)^T$ has k-variate skew-normal distribution, if its density function is

$$f_k(\mathbf{y}) = 2\phi_k(\mathbf{y}; \Omega)\Phi(\boldsymbol{\alpha}^T\mathbf{y}), \mathbf{y} \in \mathbf{R}^k \tag{3.18}$$

where $\phi_k(\mathbf{y}; \Omega)$ is the probability density function of the k-variate normal distribution with standardised marginals and correlation matrix Ω. The shape parameter vector $\boldsymbol{\alpha}$ determines the skew, where a large positive α value results in a large right skew and conversely a large negative α value results in a large left skew. When $\boldsymbol{\alpha} = \mathbf{0}$ it reduces the density in (3.18) to the N$(\mathbf{0}, \Omega)$ density and when $\boldsymbol{\alpha} \to \pm\infty$ it is reduced to the half-normal density. Scenarios of interest are shown in Table 3.16. The first scenario considers when all four components are mild-moderately skewed. We have not considered large values for skew given that the continuous outcomes can be transformed in this scenario. Scenarios 2-3 consider different magnitudes of skew in the latent continuous components only. This tests the robustness of the method to the assumption that the discrete variables manifest from a true normal continuous variable. Scenario 4 investigates when a small amount of skew is present and there is no effect of the intervention.

It is often useful to understand the deviations from normality being investigated through visualisation. Figure 3.14 shows histograms for a random sample of univariate error terms when $\alpha = 0.1$ and $\alpha = 0.05$ for N=300.

3.6.2 Results

The bias, coverage, bias-corrected coverage and power are shown in Table 3.17 for all four scenarios. In scenarios 1-3, the non-normality introduces bias which results

Table 3.16: Simulation scenarios considered to investigate deviations from joint normality for the components of the systemic lupus erythematosus composite endpoint based on the multivariate skew-normal distribution where $\boldsymbol{\alpha}$ determines the magnitude of the skew in each component

Scenario	$\boldsymbol{\alpha}$	Purpose
skew1	(0.1, 0.1, 0.1, 0.1)	Skew in all four components
skew2	(0, 0, 0.1, 0.1)	Skew in discrete components only
skew3	(0, 0, 0.05, 0.05)	Smaller skew in discrete components only
skew4	(0, 0, 0.05, 0.05)	Skew in discrete components only: null case

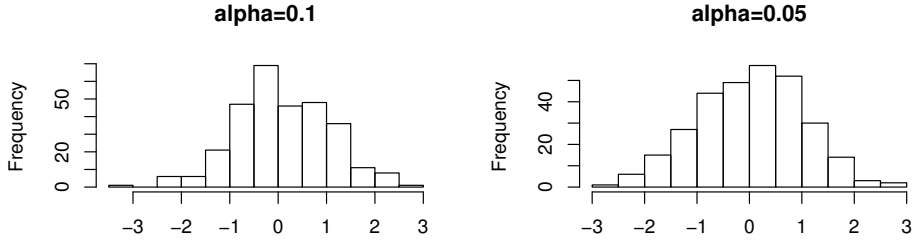


Figure 3.14: Histogram of univariate skew-normal distributed error terms with expectation equal to zero for $\alpha = 0.1$ (left) and $\alpha = 0.05$ (right)

in under-coverage. The bias-corrected coverage is close to nominal for all scenarios. However, the coverage of the latent variable method is nominal in the null case. This is consistent with our findings when the joint normality assumption is satisfied, in that bias is introduced in the estimation of the treatment arm. However the magnitude of this bias is much smaller when the assumptions are satisfied. It is worth noting that this case investigates mild-moderate skew and that for larger values of skew we would expect the bias and under-coverage to be substantial. The augmented binary and standard binary methods behave similarly to when the joint normality assumptions are satisfied, which is expected given that the model assumptions are violated in both contexts. The latent variable method still offers large power gains over the other methods. Table 3.18 shows the MSE, Empirical SE and Model SE of the three methods. The latent variable method consistently performs best across these performance measures. As the MSE is a combined bias and variance estimator, it is a useful summary for the overall performance of the methods. The augmented binary and standard binary methods have an MSE across all scenarios of approximately 0.06 whilst the MSE of the latent variable method is between 0.01 and 0.04. This indicates that the large reduction in

variance is useful despite the introduction of bias. We acknowledge however that this may not hold across all sample sizes [110].

Table 3.19 shows the probability of response in each arm for each of the methods. The findings are consistent with when the assumptions are satisfied. Namely, the latent variable method estimates the probability of response in the control arm well however underestimates the probability of response in the treatment arm. The magnitude of this underestimation is unaffected by the degree of skew or whether the skew is present in the observed continuous components. The odds ratio treatment effect estimates are shown in Table 3.20. The latent variable method is biased towards the null and the augmented binary method is biased away from the null. The binary method slightly underestimates the treatment effect in this setting however all are close to true for the null case. The median relative precision of the methods is shown in Table 3.21, with their 10th centile and 90th centile values. These are again consistent with our previous findings indicating that the violation of joint normality only affects the bias and not the variance. This is further reiterated by the reduction in confidence interval width shown in Table 3.22.

Table 3.17: Operating characteristics (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
Bias	skew1	-0.173 (0.012)	0.041 (0.252)	-0.015 (0.258)
	skew2	-0.103 (0.008)	0.036 (0.251)	-0.020 (0.255)
	skew3	-0.068 (0.008)	0.038 (0.244)	-0.016 (0.245)
	skew4	-0.033 (0.008)	0.007 (0.254)	0.001 (0.255)
Coverage	skew1	0.556 (0.018)	0.933 (0.009)	0.939 (0.009)
	skew2	0.811 (0.013)	0.928 (0.008)	0.941 (0.008)
	skew3	0.884 (0.010)	0.934 (0.008)	0.950 (0.007)
	skew4	0.933 (0.009)	0.923 (0.009)	0.950 (0.008)
Bias-corrected coverage	skew1	0.962 (0.007)	0.929 (0.009)	0.943 (0.008)
	skew2	0.936 (0.008)	0.930 (0.008)	0.943 (0.007)
	skew3	0.940 (0.008)	0.929 (0.008)	0.954 (0.007)
	skew4	0.948 (0.008)	0.926 (0.009)	0.950 (0.008)
Power	skew1	0.897 (0.011)	0.646 (0.017)	0.487 (0.018)
	skew2	0.959 (0.006)	0.637 (0.015)	0.471 (0.016)
	skew3	0.982 (0.004)	0.641 (0.015)	0.495 (0.016)
	skew4	-	-	-

Table 3.18: Operating characteristics (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
MSE	skew1	0.039 (0.001)	0.063 (0.003)	0.066 (0.003)
	skew2	0.021 (0.001)	0.063 (0.003)	0.065 (0.003)
	skew3	0.014 (0.001)	0.060 (0.003)	0.060 (0.003)
	skew4	0.010 (0.001)	0.064 (0.004)	0.065 (0.003)
EmpSE	skew1	0.097 (0.003)	0.248 (0.006)	0.257 (0.007)
	skew2	0.102 (0.002)	0.249 (0.006)	0.254 (0.006)
	skew3	0.099 (0.002)	0.241 (0.005)	0.245 (0.006)
	skew4	0.094 (0.002)	0.254 (0.006)	0.255 (0.006)
ModSE	skew1	0.010 (0.006)	0.052 (0.001)	0.064 (0.001)
	skew2	0.010 (0.003)	0.050 (0.001)	0.060 (0.001)
	skew3	0.010 (0.015)	0.048 (0.001)	0.059 (0.001)
	skew4	0.009 (0.004)	0.051 (0.001)	0.063 (0.001)

Table 3.19: Median estimated probability of response in the treatment and placebo arms (with standard deviation in parentheses) from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Pr(resp$\mid T = 0$)				Pr(resp$\mid T = 1$)			
Scenario	True	Lat Var	Aug Bin	Bin	True	Lat Var	Aug Bin	Bin
skew1	0.259	0.263	0.221	0.258	0.365	0.330	0.326	0.359
		(0.024)	(0.031)	(0.035)		(0.031)	(0.037)	(0.040)
skew2	0.290	0.287	0.253	0.290	0.398	0.370	0.361	0.392
		(0.025)	(0.033)	(0.037)		(0.033)	(0.039)	(0.041)
skew3	0.309	0.302	0.271	0.308	0.418	0.394	0.382	0.413
		(0.025)	(0.34)	(0.037)		(0.033)	(0.040)	(0.041)
skew4	0.309	0.299	0.269	0.307	0.309	0.292	0.270	0.307
		(0.025)	(0.034)	(0.037)		(0.029)	(0.034)	(0.038)

Table 3.20: Estimated odds ratio treatment effect from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Method			
Scenario	True	Lat Var	Aug Bin	Bin
skew1	1.640	1.379 (1.140, 1.668)	1.708 (1.093, 2.668)	1.616 (0.985, 2.651)
skew2	1.617	1.459 (1.203, 1.770)	1.676 (1.083, 2.594)	1.586 (0.980, 2.565)
skew3	1.611	1.505 (1.243, 1.822)	1.674 (1.089, 2.572)	1.585 (0.987, 2.548)
skew4	1.000	0.967 (0.807, 1.160)	1.007 (0.647, 1.566)	1.001 (0.613, 1.634)

Table 3.21: Relative precision of the latent variable, augmented binary and binary methods when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Method comparison		
Scenario	Lat Var vs Bin	Lat Var vs Aug Bin	Aug Bin vs Bin
skew1	6.903 [5.336, 8.972]	5.579 [4.376, 7.313]	1.231 [1.189, 1.275]
skew2	6.263 [5.013, 7.917]	5.177 [4.096, 6.518]	1.213 [1.178, 1.252]
skew3	6.326 [5.016, 7.995]	5.192 [4.098, 6.548]	1.219 [1.184, 1.257]
skew4	7.384 [5.729, 9.343]	5.985 [4.655, 7.629]	1.231 [1.192, 1.273]

Table 3.22: Median confidence interval width (standard deviation in parentheses) of the treatment effect reported from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Average CI width			% reduction CI width	
Scenario	Lat Var	Aug Bin	Bin	Lat Var	Aug Bin
skew1	0.376 (0.05)	0.890 (0.03)	0.986 (0.03)	61.86	9.75
skew2	0.384 (0.04)	0.872 (0.02)	0.960 (0.02)	60.00	9.18
skew3	0.377 (0.07)	0.858 (0.02)	0.946 (0.02)	60.17	9.37
skew4	0.362 (0.04)	0.882 (0.03)	0.978 (0.03)	63.02	9.86

3.7 Case Study

3.7.1 Trial Data

To understand more about how the methods perform in real data, we apply them to the MUSE trial [112]. The trial was a phase IIb, randomised, double-blind, placebo-controlled study investigating the efficacy and safety of anifrolumab in adults with moderate to severe SLE. Patients (n=305) were randomised to receive anifrolumab (300mg or 1000mg) or placebo, in addition to standard therapy every 4 weeks for 48 weeks. The primary end point was the percentage of patients achieving an SRI response at week 24 with sustained reduction of oral corticosteroids (<10mg/day and less than or equal to the dose at week 1 from week 12 through 24).

3.7.2 MUSE Primary Analysis

Due to data sharing policy, we conduct the analysis for a subset of the patients, namely n=278 rather than n=305 reported in the paper, thus the results will differ from the original paper. Table 3.23 shows the demographics and baseline clinical characteristics of the patients enrolled. Table 3.24 shows the efficacy results at week 24. We present the results for both the tapered SRI(4) endpoint including the oral corticosteroid information and the results for the SRI(4) excluding the tapering information. Within each case we also present the results for the high and low interferon (IFN) gene signature subgroups, as was the case for the primary analysis. Using this subset of the original trial data, we find the point estimates differ from the primary analysis, however the conclusions are the same. Therefore, the effect sizes from our re-analysis should not be taken as a contradiction of the original findings and our interest here will instead be focused on the comparable model performance.

3.7.3 Exploratory Data Analysis

The simulation results indicate that the structure of the data is an important factor in the performance of the methods and in particular that the latent variable method is sensitive to the assumptions made. A complication for exploratory data analysis in this context is that the key assumption of joint normality of the four components cannot be assessed, due the two latent outcomes. This is an obstacle for the application of these methods in practice, given that we know that violation of these assumptions can be problematic. Although univariate normality of the observed continuous components

Table 3.23: Demographic and baseline clinical characteristics of the patients enrolled in each of the three arms of the phase IIb MUSE trial in adults with moderate to severe systemic lupus erythematosus (modified ITT population)*

	Placebo (n=87)	Anifrolumab 300mg (n=95)	Anifrolumab 1000mg (n=96)
Age, years	39.2 ± 12.5	38.9 ± 11.9	41.0 ± 11.6
Sex, no. (%) female	79 (90.3)	89 (93.7)	91 (94.8)
Weight, kg	69.7 ± 19.4	69.3 ± 17.1	71.2 ± 17.0
Height, cm	161.4 ± 8.4	161.5 ± 8.6	162.0 ± 6.7
Race, no. (%)			
White	38 (43.7)	33 (34.7)	48 (50.0)
African American	8 (9.2)	18 (18.9)	10 (10.4)
Asian	7 (8.0)	3 (3.2)	5 (5.2)
American Indian/Alaska Native	0 (0.0)	4 (4.2)	0 (0.0)
Other	34 (39.1)	37 (38.9)	33 (34.4)
Ethnicity, no. (%) non-Hispanic	48 (55.2)	50 (52.6)	59 (61.5)
High IFN gene signature, no. (%)	64 (73.6)	72 (75.8)	72 (75.0)
SLEDAI-2K global score	11.0 ± 4.3	10.8 ± 3.8	10.8 ± 4.1
BILAG 2004 global score	19.9 ± 6.0	19.6 ± 5.8	18.3 ± 5.5
Physician's global assessment	1.75 ± 0.41	1.86 ± 0.39	1.85 ± 0.39

Table 3.24: Summary of efficacy results for the anifrolumab 200mg arm versus placebo and the anifrolumab 1000mg arm versus placebo in the phase IIb MUSE trial in adults with moderate to severe systemic lupus erythematosus

	Placebo (n=87)	Anifrolumab 300mg (n=95)	OR (95% CI)	P	Anifrolumab 1000mg (n=96)	OR (95% CI)	P
Week 24							
SRI(4) (including taper)	18/87 (20.7)	34/95 (35.8)	2.14 (1.14, 4.00)	0.017	29/96 (30.2)	1.66 (0.87, 3.15)	0.122
High IFN gene signature	10/64 (15.6)	27/72 (37.5)	3.24 (1.48, 7.11)	0.003	21/72 (29.2)	2.22 (1.00, 4.92)	0.048
Low IFN gene signature	8/23 (0.35)	7/23 (30.4)	0.82 (0.26, 2.56)	0.746	8/24 (33.3)	0.94 (0.29, 3.06)	0.922
SRI(4) (excluding taper)	38/87 (43.7)	52/95 (54.7)	1.56 (0.87, 2.80)	0.137	57/96 (59.4)	1.88 (1.05, 3.37)	0.033
High IFN gene signature	27/64 (42.2)	40/72 (55.6)	1.71 (0.87, 3.38)	0.120	45/72 (62.5)	2.28 (1.15, 4.52)	0.017
Low IFN gene signature	11/23 (47.8)	12/23 (52.2)	1.19 (0.39, 3.63)	0.773	12/24 (0.50)	1.09 (0.35, 3.38)	0.889

* adjusted for three stratification factors

does not imply joint normality of these, we will look at the outcomes individually and jointly in order to learn more about the structure of the data. Although the covariate adjustment will be informed by clinical relevance in practice, we will look at the distribution of the residuals under various covariate adjustments to determine under what covariate structures the assumptions are most justified.
Table 3.25 shows the decomposition of responders and non-responders in each of the components by treatment arm. In both the treatment and the control arm, almost all patients are responders in both the PGA and BILAG measures. This indicates that these components do not enrich the composite endpoint and so it is the SLEDAI and taper measures that are responsible for driving response rates. From the simulated scenarios, we may expect smaller precision gains than if three or four components determined response. We also expect that the augmented binary method may perform more similarly to the latent variable model in terms of efficiency, due to modelling the two informative components.

Physician's Global Assessment (PGA)

The PGA measure used in the primary analysis is the change score at week 24. To derive this change score, the baseline PGA measure is subtracted from the observed measure at week 24. If this change is <0.3, then the patient is considered to be a responder in this endpoint. Figure 3.15 shows the histograms for the PGA change measures and PGA raw measures. These are also shown by treatment arm. Figure 3.16 shows the distribution of the residuals for two different covariate adjustments. The model on the left includes the treatment arm. The mean structure of the PGA measure of interest in our analysis will include both the treatment and PGA baseline measure, as in the right hand figure.

SLEDAI-2K

The SLEDAI-2K measure included in the primary analysis is the change score at week 24. This change score is derived by obtaining the difference between an imputed SLEDAI score at week 24 and the observed baseline SLEDAI score. The imputation involves using the raw SLEDAI score at week 24, if it has been observed. If not, it is imputed using the last observed measure i.e. last observation carried forward. This imputation procedure was planned and conducted by AstraZeneca for the primary analysis. It is possible to investigate the influence of their choice of imputation method

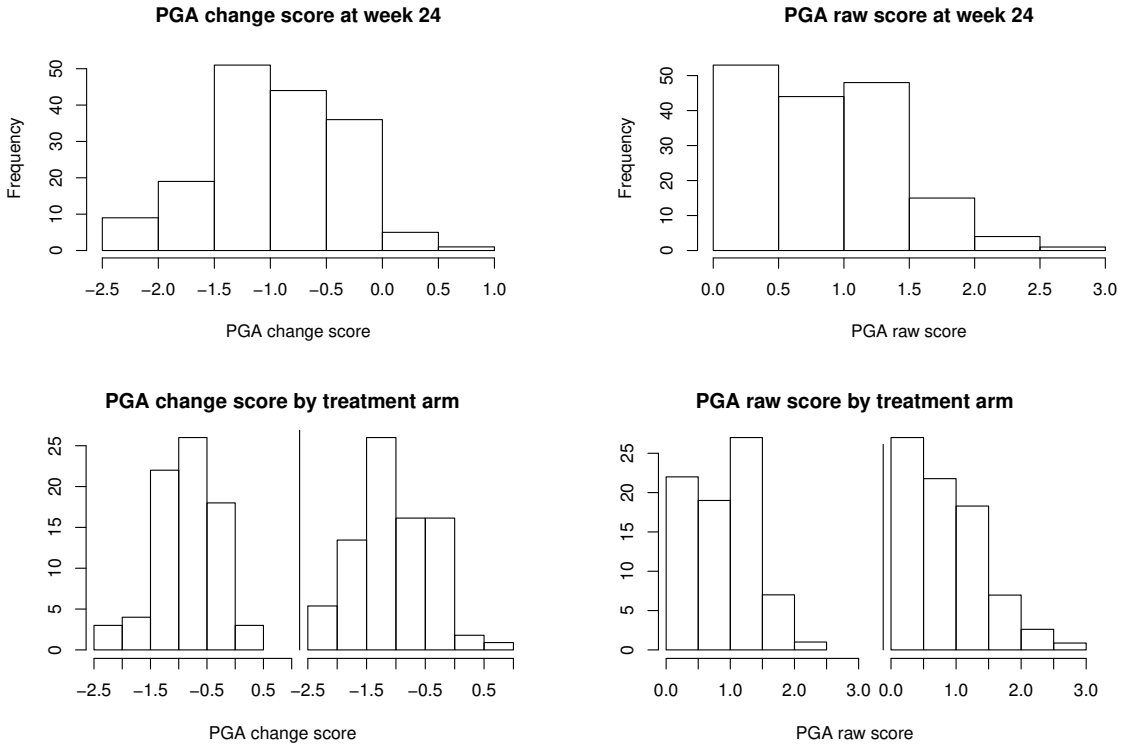


Figure 3.15: Histogram plots for Physicians Global Assessment (PGA) measures: PGA change score (top left), PGA raw score (top right), PGA change score in anifrolumab 300mg arm and placebo arm (bottom left) and PGA raw score in anifolumab 300mg arm and placebo arm (bottom right) in the phase IIb MUSE trial in patients with systemic lupus erythematosus

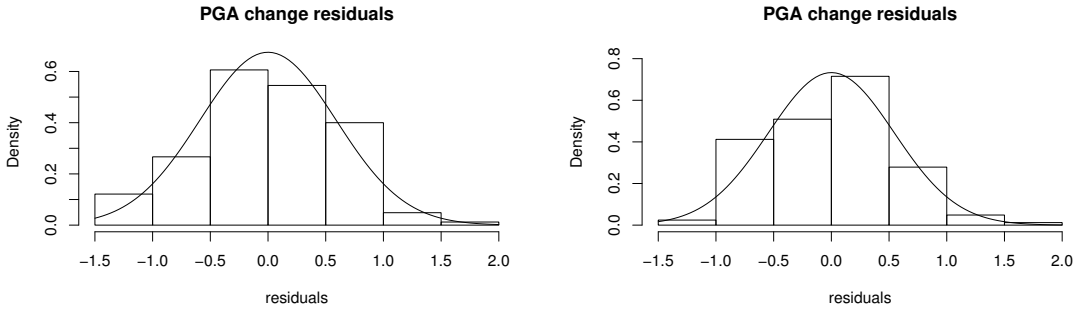


Figure 3.16: Histogram plots for residuals in the Physicians Global Assessment (PGA) measure adjusted for treatment arm (left) and adjusted for treatment arm and baseline PGA measure (right) in the phase IIb MUSE trial in patients with systemic lupus erythematosus

Table 3.25: Observed response rates in each of the SLE responder index components in the anifrolmab 300mg arm and placebo arm of the phase IIb MUSE trial. SLE index is comprised of a continuous SLEDAI outcome, continuous PGA outcome, ordinal BILAG outcome and binary taper outcome where response in each component is achieved when the patient meets the criteria shown

Components	Response criteria	Treatment arm	
		Anifrolumab 300mg	Placebo
SLEDAI	Change in SLEDAI $\leq$ -4	58/89	41/76
PGA	Change in PGA $<$ 0.3	87/89	75/76
BILAG	No Grade A or more than one Grade B	86/89	72/76
Taper	Sustained reduction in oral corticosteroids	53/95	37/87
SLE responder endpoint	Responder in all four components above	34/95	18/87

on the analysis and conclusions by employing different imputation procedures and determining how this affects the conclusions. However, for the purpose of this work, we will proceed using the imputation in the primary analysis.
Figure 3.17 shows the histograms for the change score and imputed week 24 score. These are both shown by treatment arm. Figure 3.18 shows the distribution of the residuals for two different covariate adjustments. The analysis of interest includes treatment and baseline SLEDAI measure, as in the right hand figure.

BILAG-2004

The BILAG component grades patients A-E on an ordinal scale across nine organ systems at each visit. They are then given a global BILAG score based on these grades. Each grade is assigned points and this is used to calculate the continuous global score. Grade A is scored 12, Grade B is scored 8, Grade C scored 1 and Grade D and E are both scored 0. If the global score is not observed, the last observation is imputed. A patient is considered to be a non-responder in BILAG if one or more body systems has been scored 12 at week 24 which had been scored 8 or less at baseline or two

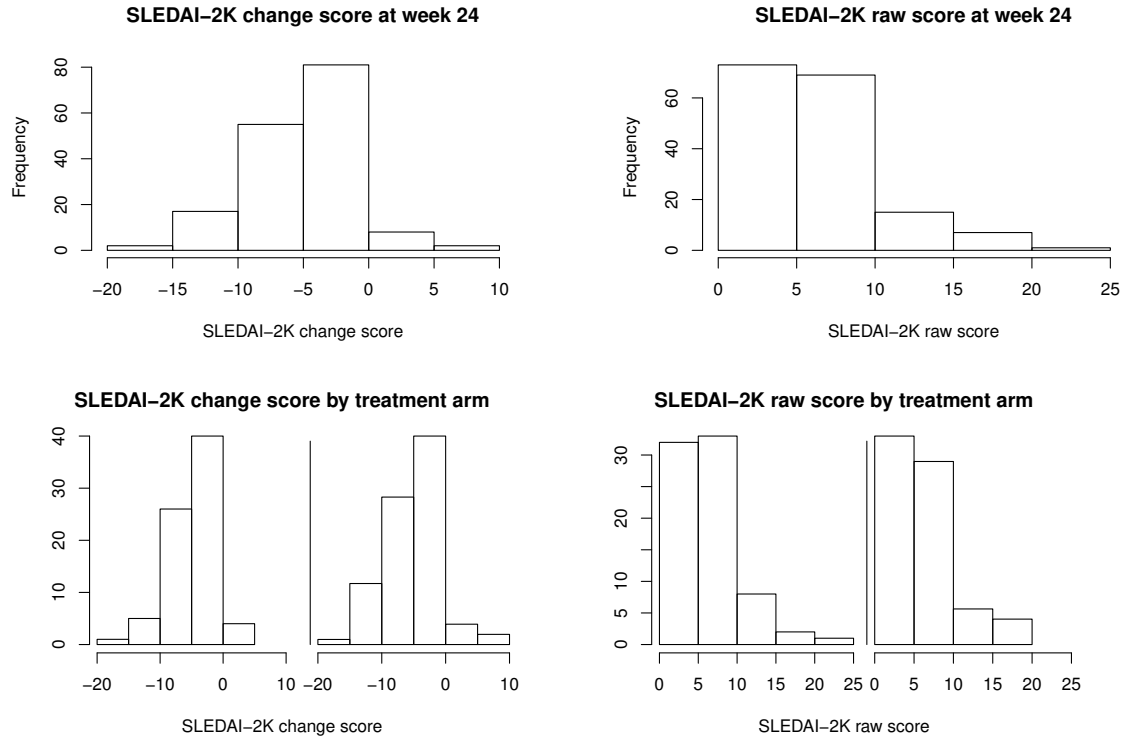


Figure 3.17: Histogram plots for Systemic Lupus Erythematous Disease Activity Index (SLEDAI) measures: SLEDAI change score (top left), SLEDAI raw score (top right), SLEDAI change score in anifrolumab 300mg arm and placebo arm (bottom left) and SLEDAI raw score in anifolumab 300mg arm and placebo arm (bottom right) in the phase IIb MUSE trial in patients with systemic lupus erythematosus

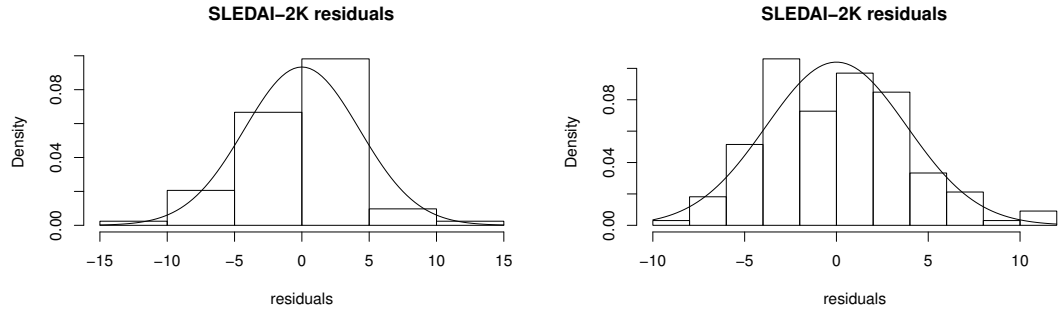


Figure 3.18: Histogram plots for residuals in the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) measure adjusted for treatment arm (left) and adjusted for treatment arm and baseline SLEDAI measure (right) in the phase IIb MUSE trial in patients with systemic lupus erythematosus

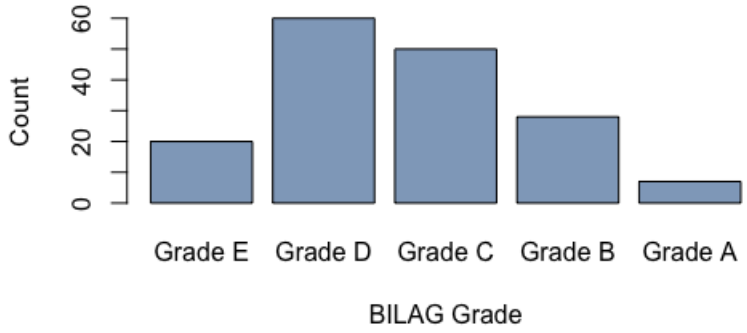


Figure 3.19: Barplot showing the British Isles Lupus Assessment Group (BILAG) measure in the Phase IIb MUSE trial in patients with systemic lupus erythematosus

or more body systems scored 8 at week 24 which scored 1 or less at baseline. As with the other components, BILAG is treated as a binary variable in the primary analysis. A responder will have no new grade A's and no more than one new grade B from baseline at week 24. Table 3.25 shows that 158/165 patients are responders in BILAG. Due to the implicit ordinal scaling, we model BILAG on the ordinal scale for the overall measure. The plot in Figure 3.19 shows the frequency of patients in each category. Note that very small numbers of non-responders may make the estimation of the thresholds more difficult.

Tapering

In order to be a responder in the tapering component, patients had to achieve a sustained reduction of oral corticosteroids. This is defined as <10mg/day and less than or equal to the dose at week 1 from week 12 through 24. This outcome contributed substantially in discriminating between responders and non-responders, as can be seen in Table 3.25. The observed binary variable will be used in the analysis, setting the threshold to zero, allowing for estimation of an intercept term.

Multivariate Plots

We can visualise the 4-D mixed outcome data as shown in Figure 3.20. The continuous PGA and SLEDAI measures are shown on the x and y axis respectively. The ordinal BILAG measure is shown as different coloured data points, where non-responders

are shown in purple. The taper responders are shown in the left panel and the non-responders in the right. Overall SLE responders are shown in the bottom left hand quadrant of the left hand panel. It is worth noting we can also determine visually that responder status is completely specified by SLEDAI and taper outcomes. This is demonstrated as there are no BILAG or PGA non-responders in the bottom half of the left panel.

A heatmap for the correlations is shown in Figure 3.21. We can see that the components of the SRI-4 variable are strongly positively correlated and that none of the variables are strongly correlated with the tapering outcome. SLEDAI and PGA are strongly negatively correlated with their respective baseline measures, as we would expect due to their inclusion in the change outcome. Violin plots are useful to visualise the density of the observed continuous measures by treatment arm across the categories of the observed discrete variables. Figure 3.22 shows the violin plots for the SLE index components. These plots are similar to boxplots however also show the probability density of the data values which has been smoothed using a Gaussian kernal. PGA and SLEDAI measures are not substantially different across the tapering levels, which we would expect due to their low correlation. The correlation between BILAG and both measures is indicated by rising values of PGA and SLEDAI change when approaching the response threshold.

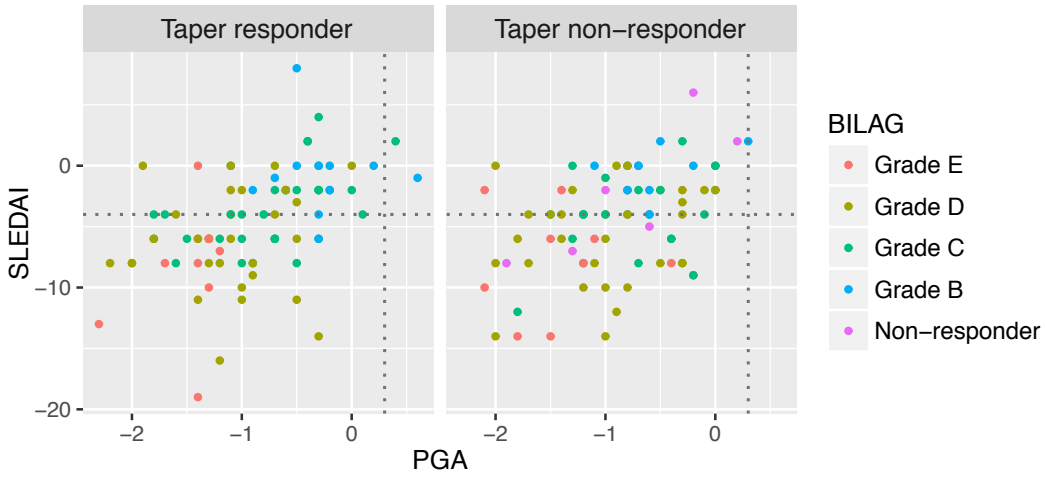


Figure 3.20: Observed response rates in each of the SLE responder index components in the phase IIb MUSE trial. To be classed a responder a patient must be: below -4 to respond in the continuous SLEDAI outcome (left), below 0.3 in the continuous PGA measure (bottom), any colour but purple for BILAG and be placed in the left hand panel for taper response. Overall responders are shown in the bottom left quadrant of the left hand panel

Figure 3.21: Heatmap showing correlations between the four systemic lupus erythematosus composite endpoint components and their baseline variables: British Isles Lupus Assessment Group (BILAG), oral corticosteroid tapering (tapering), Physicians Global Assessment (PGA), Physicians Global Assessment baseline measure (PGAbase), Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) and Systemic Lupus Erythematosus Disease Activity Index baseline measure (SLEDAIbase) in the phase IIb MUSE trial

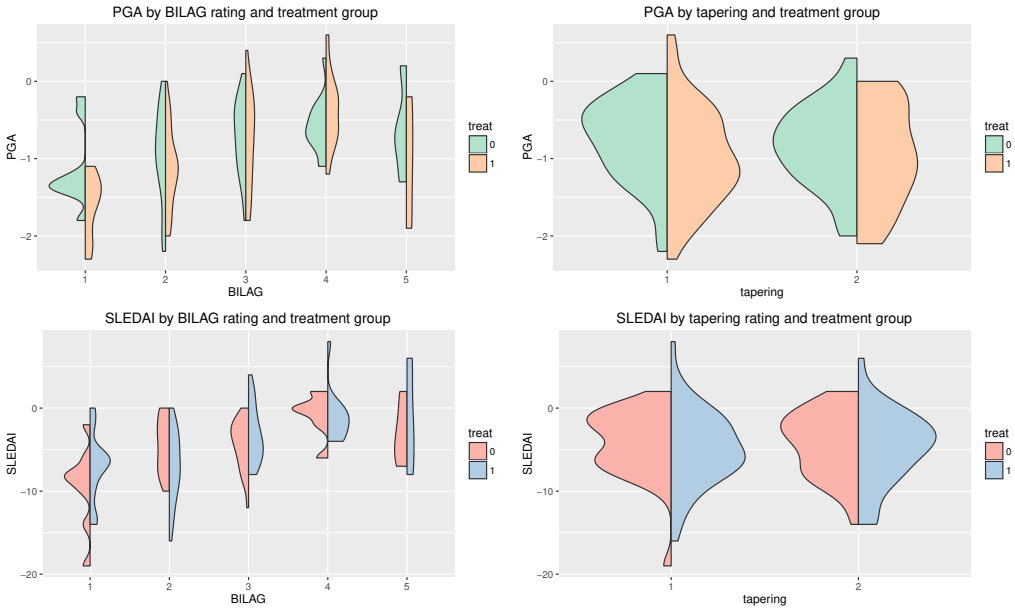


Figure 3.22: Violin plots of Physicians Global Assessment (PGA) by British Isles Lupus Assessment Group (BILAG) rating and treatment group (top left), PGA by oral corticosteroid tapering and treatment group (top right), Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) by BILAG rating and treatment group (bottom left), SLEDAI by tapering rating and treatment group (bottom right)

3.7.4 MUSE Trial Re-Analysis

Table 3.26 shows the estimated probability of response in each arm of the trial for each of the three methods. The probability of response in the placebo arm is similar for all methods. A much larger discrepancy is shown in the treatment arm, which agrees with the findings from the simulation study.

Table 3.26: Estimated probability of response from the latent variable, augmented binary and standard binary methods in the anifrolumab 300mg arm and placebo arm of the phase IIb MUSE trial in adults with systemic lupus erythematosus

Method	Anifrolumab 300mg	Placebo
Latent Variable	0.311	0.199
Augmented Binary	0.324	0.211
Binary	0.382	0.224

The log-odds treatment effect point estimates and confidence intervals are shown in Figure 3.23. Both joint modelling methods estimate the treatment effect more precisely. Although there may be bias towards the null present in the point estimates for the joint modelling methods, the confidence intervals entirely overlap with that of the binary method. All three methods indicate that anifrolumab 300mg performs better than placebo, as in the original findings.

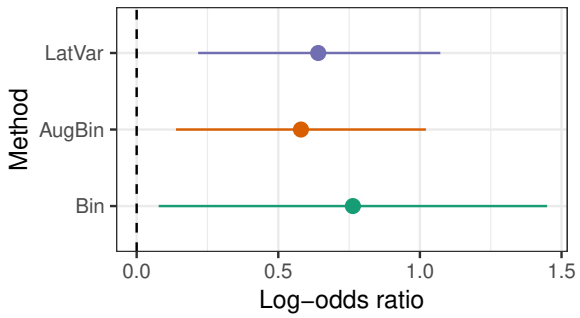


Figure 3.23: Estimated log-odds treatment effect estimates from the latent variable, augmented binary and standard binary methods in the phase IIb MUSE trial in adults with systemic lupus erythematosus

In terms of estimated precision, it is interesting to determine where the trial data set lies in the distribution of datasets generated in the simulation study. Table 3.27 shows the estimated relative precision gains in the MUSE data. The latent variable method is

Table 3.27: Relative precision comparison of the treatment effect estimates reported from the latent variable, augmented binary and standard binary methods in the Phase IIb MUSE trial in adults with systemic lupus erythematosus

Comparison	Relative precision	Reduction required sample size
Lat Var vs Bin	2.549	60.4%
Lat Var vs Aug Bin	1.056	4.40%
Aug Bin vs Bin	2.415	58.6%

2.5 times as precise as the binary method in this setting, whilst the augmented binary method is 2.4 times as precise. This similar performance is expected as the augmented binary method models the SLEDAI and taper variables - the only components driving response. This increase in precision from the latent variable method compared with the binary method amounts to a 60% reduction in required sample size.

3.7.5 Model Fit

We assess the goodness-of-fit of the latent variable model in the MUSE trial dataset using the modified Pearson residuals introduced earlier. Figure 3.24 shows the residuals for each patient in the trial. The model appears to fit well with only two observations poorly explained. Figure 3.25 shows a histogram of the residuals. If the model fits well, the residuals should follow the χ^2_9 distribution shown. The residuals seem to follow the χ^2_9 distribution except for the observations with residuals equal to 40, which confirms that the model fits the data well apart from these two measurements.

3.8 Bias Correction Using the Bootstrap

The simulations have shown that the variance of the treatment effect reported by the latent variable method is always largely reduced from that of the standard binary method, however bias is introduced in settings with large treatment effects and when joint normality assumptions are not satisfied. Furthermore, the bias-corrected coverage of the confidence interval is nominal even when the coverage is poor. In theory, if the bias could be estimated from the observed data then we could subsequently correct for this to obtain the desirable property of an unbiased estimator with nominal coverage.

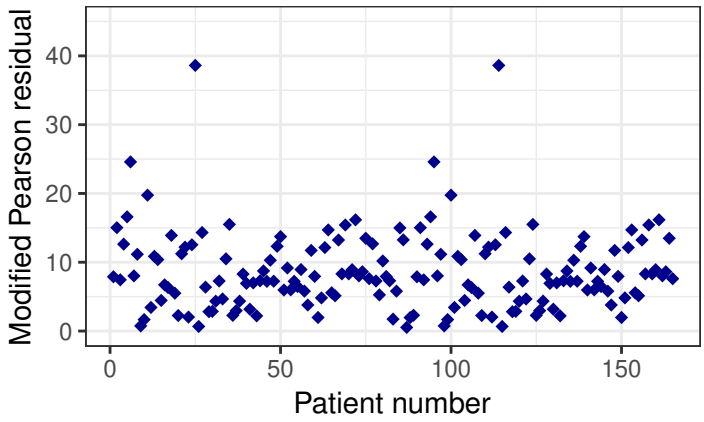


Figure 3.24: Plot of the Modified Pearson residuals from the latent variable model for each patient in the MUSE trial. The residuals highlight that two patients observations are poorly explained by the model but that the model is a good fit for the remaining patients

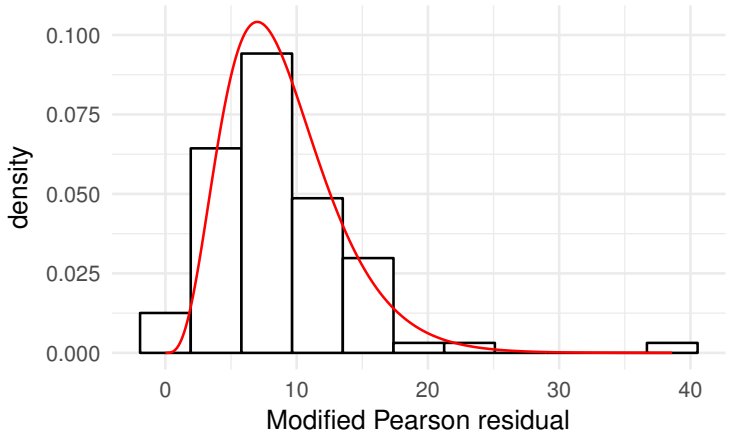


Figure 3.25: Histogram of the modified Pearson residuals from the latent variable model in the MUSE trial dataset with the corresponding χ^2_9 density. The modified Pearson residuals should follow the distribution of the χ^2_9 density shown if the model fits well

3.8.1 Bootstrap Method

Efron's bootstrap [113] is a general method which is used for estimating various properties of a given statistic, in particular its bias and variance. The concept is based on treating the observed sample as the population and then sampling with replacement from this n_{boot} times. Note that this is not a permutation distribution due to the replacement when sampling. Efron constructs the problem by assuming a random sample of size N is observed from an unspecified distribution F,

$$X_i = x_i, \quad X_i \sim_{ind} F \qquad i = 1, 2, ..., N \tag{3.19}$$

Given a specified random variable R($\mathbf{X}$, F), we wish to estimate the sampling distribution of R on the basis of the observed data $\mathbf{x}$. The bootstrap method to solve this problem is as follows:

1. Construct the sample probability distribution $\hat{F}$, putting mass $\frac{1}{N}$ at each point $x_1, x_2, ..., x_N$.

2. With $\hat{F}$ fixed, draw a random sample of size N from $\hat{F}$ to obtain the bootstrap sample
$$X_i^* = x_i^*, \quad X_i^* \sim_{ind} \hat{F} \qquad i = 1, 2, ..., N \tag{3.20}$$

3. Approximate the sampling distribution of R($\mathbf{X}$,F) by the bootstrap distribution of
$$R^* = R(\mathbf{X}^*, \hat{F}) \tag{3.21}$$

In theory, the distribution of R^* can be calculated exactly once the data $\mathbf{x}$ is observed and will equal the desired distribution of R if $F = \hat{F}$.

3.8.2 Application in the One-Sample Multivariate Case

To investigate how the bootstrap would work in practice, we demonstrate it using the MUSE trial data. In this scenario N=182 and due to the computational complexity we choose n_{boot}=1000. Therefore the procedure is as follows:

1. Sample with replacement N=182 patients from the MUSE trial

2. Compute the treatment effect using the latent variable, augmented binary and standard binary methods

Table 3.28: Log-odds treatment effect estimates and 95% confidence intervals from the latent variable method, augmented binary method and standard binary method in the phase IIb MUSE trial and the bootstrap sample when N=182 and $n_{boot} = 1000$

Method	Log-odds treatment effect	
	MUSE trial estimate	Bootstrap estimate
Latent Variable	0.641 (0.217, 1.072)	0.682 (0.275, 1.137)
Augmented binary	0.580 (0.139, 1.021)	0.608 (0.096, 1.111)
Binary	0.763 (0.078, 1.449)	0.809 (0.112, 1.561)

3. Repeat step 1 and 2 n_{boot}=1000 times

4. Obtain an estimate of the bias using the difference between the treatment effect in the MUSE trial and the mean of the bootstrap treatment effects

Importantly, a 95% bootstrap confidence interval for the treatment effect estimate can be obtained by ordering the 1000 bootstrap estimates of the treatment effect and taking the 25th and 975th estimate. The point estimates and 95% confidence intervals from the MUSE trial and from the re-sampling are shown in Table 3.28.
The log-odds point estimate from the latent variable method has shifted away from the null by approximately 0.04. This is the magnitude of bias that the simulation results suggested for this treatment effect. The width of the confidence interval has remained the same in the bootstrap sample, indicating that the variance is well estimated in the trial dataset. Ideally, we would investigate this further across a larger number of datasets however this is too computationally intensive. To perform this on one replicate, where $n_{boot} = 1000$ using 200 cores on the HPC currently takes 7 hours. Exploring this further through bootstrapping or employing alternative multivariate distributions is an area for future research.

3.9 Discussion

The work in this chapter aimed to address the large loss of information in modelling complex composite endpoints. One challenge in this work was determining an appropriate joint model for the components when these are measured on different scales. By partitioning latent variable outcome spaces we were able to model the observed

structure of the composite endpoint which resulted in large gains in efficiency. These gains in efficiency were offset by the introduction of a small bias when the treatment effect is large. Sensitivity analyses showed that this bias is exacerbated when the assumptions of joint normality were not satisfied, however similar reductions in variance were observed. Application to the MUSE trial data reinforced the simulation findings, in that the treatment effect reported from the latent variable method was 2.5 times as precise as that reported from the logistic regression and appeared to be biased towards the null.

Bias correction seems to perform well in the real data, where the crucial assumptions cannot be tested. The point estimate is shifted by a magnitude that would have been expected from the simulation results and the estimate of the variance is similar to that obtained in the single trial dataset. Furthermore the latent variable bootstrap confidence interval for the treatment effect is contained within that for the binary method, which offers further reassurance for application. However, these results are not definitive and more work could be done on investigating different structures and scenarios to ensure that the bias correction is always what we would expect from simulation results.

The potential precision gains offered by the latent variable method offer justification for the additional complexity however the magnitude of these gains are highly dependent on the components that drive response. The baseline case in the simulations was chosen to reflect when a composite endpoint is recommended for use, i.e. when all four components were responsible for driving response. In this scenario the precision gains achieved resulted in the latent variable method reporting the effect 2.5 to 17.5 times more precisely than the standard binary method. However, in practice in SLE trials this has not been found to be the case. A review of two phase III trials (N= 2262) using the SRI-5 index found the SRI-5 response rate at week 52 for all patients was 32.8% [114]. Non-response due to a lack of SLEDAI improvement, concomitant medication non-compliance or dropout was 31, 16.5 and 19.1%, respectively. Non-response due to deterioration in BILAG or PGA after SLEDAI improvement, concomitant medication compliance and trial completion was 0.5%. This is in agreement with our findings from the MUSE trial data, which suggests that the precision gains in the baseline case are optimistic. The simulation results show that when one continuous and one binary component drive response, the latent variable method may be anywhere between 1 and 12 times as precise. This means that in a very small number of cases (<2%) there are no precision gains from the increased complexity of the latent variable method. The

potential gains available in 98% of cases ensure that implementing the latent variable method is very much a worthwhile endeavour, for all stakeholders in a clinical trial.
Another useful metric in considering whether the method should be employed in practice is the MSE, as this is a combined measure of bias and variance. The simulation results show that the MSE of the reported treatment effect from the latent variable method (0.01-0.04) is always smaller than that of the standard binary method (0.06).
Another important consideration comes from an ethical context. Having the skills to interpret these performance measures means that statisticians also have an ethical obligation when recommending methods for use. If the confidence interval has close to nominal coverage, it is important to consider whether an unbiased point estimate is crucial, especially when the required sample size may be reduced by 60%+. This sample size reduction would mean that fewer patients are subjected to placebo, effective drugs may make it to market sooner and could allow randomisation ratio to be moved away from 1:1 without affecting power. We therefore recommend the latent variable method for use in practice in SLE trials. Should the method not be employed as the primary analysis method, it should at least be fitted as a secondary analysis measure to enhance understanding of the trial data.
In addition to SLE, we have identified other disease areas that have a similar complex composite structure, meaning the potential to improve efficiency extends well beyond the SLE paradigm. However, it must be acknowledged that the exact structure of the endpoint may offer different magnitudes of bias, precision and computational time. In addition, as we have coded the likelihood ourselves with no generic package available to do this, the likelihood and probability of response code will have to be tailored specifically to each endpoint. In order to promote implementation in the general case of multiple continuous and discrete outcomes, we will need to develop a software package. This is beyond the remit of this thesis but is an important consideration for future work.
Obtaining maximum likelihood estimates from latent variable models has been achieved in different ways throughout the literature. In this work we have used a quasi-Newton algorithm however these and Newton type algorithms are not without their limitations, such as tending to be slow or intractable in higher dimensions [85]. The EM algorithm has been proposed in this setting as it lends itself well to situations with unknown parameters such as the τ-thresholds, however conditioning on these parameters as in (3.4) violates regularity conditions. Hence a Parameter-Expanded EM algorithm which transforms the latent variables and expands the parameter space may be more

appropriate [115]. For an implementation of this estimation method when identifying genetic factors for comorbid conditions, see the work conducted by Zhang [96]. Implementing the method as we have done in this paper is computationally demanding however we would not expect the Parameter Expanded EM algorithm to rectify this and may actually lead to increased computational time. More work is required to compare estimation methods for latent variable models in general.

The work in this chapter advocates the use of novel methodology to extract more available information from a complex composite endpoint. An obstacle for the uptake and implementation of this method is the lack of an existing method to perform a sample size calculation for a given trial. We explore this in the following chapter.

Chapter 4

Sample Size Estimation using the Latent Variable Model

4.1 Motivation

Sample size estimation plays an integral role in the design of a clinical trial. The objective is to determine the minimum sample size that is large enough to detect, with a specified power, a predetermined clinically meaningful treatment effect. Although it is crucial that investigators have enough patients enrolled to detect this effect, overestimating the sample size also has ethical and practical implications. Namely, in a placebo-controlled trial, more patients are subjected to a placebo arm than is necessary therefore withholding access to potentially beneficial drugs from them and delaying access to future patients. Furthermore it results in longer, more expensive trials, using resources that could be allocated elsewhere.

Mixed outcome components may be collapsed into a binary composite endpoint based on response thresholds, as we have seen previously. If a composite is selected as the primary endpoint in a trial then a sample size calculation is needed and this is typically based on the overall binary responder endpoint analysed using logistic regression. Sample size calculations performed in this way are valid but when applying a novel analysis approach that increases power, such as the latent variable model in Chapter 3, it is desirable to have the option to take this into account in the sample size calculation. If we can develop an approach to calculate the sample size using the latent variable method then the potential efficiency gains are much more likely to be realised in practice.

4.2 Literature Review

A recent and comprehensive overview of the existing literature for sample size determination in clinical trials with multiple endpoints is provided by Sozu et al. [116]. The review found many proposals for power and sample size calculations for multiple continuous outcomes. Some of these suggestions were based on assuming that the endpoints were bivariate normally distributed [97, 117]. Extensions of these methods discussed the case where there are more than two endpoints and provided practical formulae for implementation [98, 118]. Other work focused on testing procedures and found two-sample t-tests, which reject only if each t-statistic is significant, to be conservative and biased, sometimes resulting in large sample sizes [119, 120]. Other efforts were focused on investigating and controlling the type I error rate [33, 121–123]. All of these methods focus on the requirement of effects on all endpoints. Methods for effects on at least one endpoint also exist [33, 123–125].

Substantially less consideration has been given to the case of multiple binary endpoints. Five methods of power and sample size calculation based on three association measures are introduced for co-primary binary endpoints by Sozu et al. [126]. Sample size calculation for trials using multiple risk ratios and odds ratios for treatment effect estimation is discussed by Hamasaki et al. [127]. Song [128] explores co-primary endpoints in non-inferiority clinical trials. Consideration has also been given to the case where two co-primary endpoints are both time-to-event measures where effects are required in both endpoints [129–131] and at least one of the endpoints [132].

Despite these advances for multiple outcomes measured on the same scale, very little consideration has been given to the mixed outcome setting. One paper considers overall power functions and sample size determinations for multiple co-primary endpoints that consist of mixed continuous and binary variables [133]. They assume that response variables follow a multivariate normal distribution, where binary variables are observed in a dichotomized normal distribution, and use Pearson's correlations for association. A modification was suggested to this method using latent-level tests and pairwise correlations [134]. These methods focus on the co-primary endpoint case, where effects are required in all outcomes. To date, the case of composite endpoints where the components are measured on different scales has not been considered. Rather than requiring effects on each component, the focus here will be on the combination of the outcomes.

4.3 Aims

In this chapter we will extend the work in [133, 134] for co-primary continuous and binary endpoints to include any combination of continuous, ordinal and binary measures which require effects in all of the endpoints. We will demonstrate the application of the method on the four dimensional endpoint in the MUSE trial dataset, assuming for this purpose that it is a co-primary endpoint rather than a composite. We will then consider how we can determine the required sample size in the case of mixed outcome composite endpoints, allowing the components to be a combination of continuous, ordinal and binary endpoints and retaining the outcome of interest on the overall composite. We will investigate the power and sample size using the method developed and determine how they are affected by the correlation between components. Finally we will apply the method to the MUSE trial data to determine the sample size required in a future trial and make recommendations for future work.

4.4 Model

We set up the model as follows. Let n_T and n_C represent the number of patients in the treatment group and the control group respectively and let K be the number of outcomes measured for each patient. Let $\mathbf{Y_{Ti}} = (Y_{Ti1}, ..., Y_{TiK})^T, i = 1, ..., n_T$ be vector of K responses for patient i on the treatment arm and $\mathbf{Y_{Ci}} = (Y_{Ci1}, ..., Y_{CiK})^T, i = 1, ..., n_C$ the vector of K responses for patient i on the control arm. The first $1 \leq k \leq k_m$ elements of $\mathbf{Y_{Ti}}$ and $\mathbf{Y_{Ci}}$ are observed as continuous variables, the next $k_m < k \leq k_o$ are observed as ordinal and the remaining $k_o < k \leq K$ are observed as binary. As before we use the biserial model of association by Tate in [88], which is based on latent continuous measures manifesting as discrete variables. Formally, we say that $\mathbf{Y_{Ti}}$ and $\mathbf{Y_{Ci}}$ have latent variables $\mathbf{Y^*_{Ti}}$ and $\mathbf{Y^*_{Ci}}$ respectively, where $\mathbf{Y^*_{Ti}} \sim N_K(\boldsymbol{\mu_T}, \Sigma_T)$ and $\mathbf{Y^*_{Ci}} \sim N_K(\boldsymbol{\mu_C}, \Sigma_C)$

$$\Sigma_T = \begin{pmatrix} \sigma^2_{T1} & \cdots & \rho_{T1K}\sigma_{T1}\sigma_{TK} \\ \vdots & \ddots & \vdots \\ \rho_{T1K}\sigma_{T1}\sigma_{TK} & \cdots & \sigma^2_{TK} \end{pmatrix}, \quad \Sigma_C = \begin{pmatrix} \sigma^2_{C1} & \cdots & \rho_{C1K}\sigma_{C1}\sigma_{CK} \\ \vdots & \ddots & \vdots \\ \rho_{C1K}\sigma_{C1}\sigma_{CK} & \cdots & \sigma^2_{CK} \end{pmatrix}$$

For $k \neq k' : 1 \leq k < k' \leq K$ we let $Var(Y_{Tik}) = \sigma^2_{Tk}, Var(Y_{Cik}) = \sigma^2_{Ck}, Corr(Y_{Tik}, Y_{Tik'}) =$

$\rho_{Tkk'}, Corr(Y_{Cik}, Y_{Cik'}) = \rho_{Ckk'}$, where $\rho_{Tkk'}$ and $\rho_{Ckk'}$ are the association measures between the endpoints. Then, for:

- $1 \leq k \leq k_m : Y_{Tik} = Y^*_{Tik}$ and $Y_{Cik} = Y^*_{Cik}$
- $k_m < k \leq k_o$:

$$Y_{Tik} = \begin{cases} 0 & \text{if } \tau_{k0} \leq Y^*_{Tik} < \tau_{k1}, \\ 1 & \text{if } \tau_{k1} \leq Y^*_{Tik} < \tau_{k2}, \\ \vdots & \vdots \\ w_k & \text{if } \tau_{kw_k} \leq Y^*_{Tik} < \tau_{k(w_k+1)} \end{cases} \qquad Y_{Cik} = \begin{cases} 0 & \text{if } \tau_{k0} \leq Y^*_{Cik} < \tau_{k1}, \\ 1 & \text{if } \tau_{k1} \leq Y^*_{Cik} < \tau_{k2}, \\ \vdots & \vdots \\ w_k & \text{if } \tau_{kw_k} \leq Y^*_{Cik} < \tau_{k(w_k+1)} \end{cases}$$

- $k_o < k \leq K : Y_{Tik} = \begin{cases} 0 & \text{if } \tau_{k0} \leq Y^*_{Tik} < \tau_{k1}, \\ 1 & \text{if } \tau_{k1} \leq Y^*_{Tik} < \tau_{k2} \end{cases} \quad Y_{Cik} = \begin{cases} 0 & \text{if } \tau_{k0} \leq Y^*_{Cik} < \tau_{k1}, \\ 1 & \text{if } \tau_{k1} \leq Y^*_{Cik} < \tau_{k2} \end{cases}$

We set $\tau_{k0} = -\infty, \tau_{k(w_k+1)} = \infty$ for $k_m < k \leq k_o$ meaning that the intercept must be set to zero in order to estimate the cut-points and $\tau_{k0} = -\infty, \tau_{k1} = 0, \tau_{k2} = \infty$ for $k_o < k \leq K$ so that the intercepts can be estimated for the outcomes observed as binary. Furthermore, for $k_m < k \leq K$, $\sigma^2_{Tk} = \sigma^2_{Ck} = 1$.
Letting $\Sigma = \Sigma_T = \Sigma_C$ and partitioning $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ so that Σ is as shown in (4.1).

$$\Sigma = \begin{pmatrix} \sigma^2_1 & \cdots & \rho_{1k_m}\sigma_1\sigma_{k_m} & \rho_{1k_{m+1}}\sigma_1 & \cdots & \rho_{1K}\sigma_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1k_m}\sigma_1\sigma_{k_m} & \cdots & \sigma^2_{k_m} & \rho_{k_mk_{m+1}}\sigma_{k_m} & \cdots & \rho_{k_mK}\sigma_{k_m} \\ \rho_{1k_{m+1}}\sigma_1 & \cdots & \rho_{k_mk_{m+1}}\sigma_{k_m} & 1 & \cdots & \rho_{k_{m+1}K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\sigma_1 & \cdots & \rho_{k_mK}\sigma_{k_m} & \rho_{k_{m+1}K} & \cdots & 1 \end{pmatrix} \tag{4.1}$$

For $k_m < k \leq K$ we can define the conditional mean for outcome k as $\mu_{k|1\ldots k_m} = \mu^*_k + \Sigma_{12}\Sigma^{-1}_{22}(y_k - \mu_k)$ where μ^*_k is the latent mean.
The correlation matrix for the outcomes can then be defined using the pairwise correlations between elements of Y_i and Y^*_i as below.

$$\Gamma = \begin{pmatrix} \boldsymbol{D}^{-\frac{1}{2}}\Sigma_{11}\boldsymbol{D}^{-\frac{1}{2}} & \boldsymbol{D}^{-\frac{1}{2}}\Sigma_{12} \\ & \Sigma_{22} \end{pmatrix} \tag{4.2}$$

where $\boldsymbol{D}^{-\frac{1}{2}} = diag(\sigma_1^{-1}, ..., \sigma_{k_m}^{-1})$.

4.5 Mixed Outcome Co-Primary Endpoints

As mentioned in Section 4.2, one potential application of the latent variable model is to mixed outcome co-primary endpoints. Sozu et al. [133] propose a method to calculate the sample size for a mixture of continuous and binary endpoints which we can easily extend to ordinal outcomes, as shown below.

4.5.1 Hypothesis Testing

In many clinical trials the hypothesis of interest is based on superiority, namely that the proposed treatment will perform better than the control treatment, defined by some predefined margin. The null hypothesis is that the difference in treatment effects for the treatment arm and control arm is zero. This is straightforward to formalise in the case of one endpoint but less so when there are multiple co-primary endpoints, particularly when they are measured on different scales. Based on the work by Sozu et al. [133] we can state the hypothesis of interest as shown in (4.3).

$$\begin{aligned} H_0 &: \ \exists k \ s.t. \ \pi_{Tk} - \pi_{Ck} \leq 0 \\ H_1 &: \ \pi_{Tk} - \pi_{Ck} > 0 \ \forall k \end{aligned} \tag{4.3}$$

For $k_o < k \leq K$ we can specify $\pi_{Tik} = P(Y_{Tik} = 0) = P(Y^*_{Tik} < 0)$ and $\pi_{Cik} = P(Y_{Cik} = 0) = P(Y^*_{Cik} < 0)$ for the treatment and control group respectively. We can generalise this assumption from [133] to account for the ordinal endpoints based on the fact that for $k_m < k \leq k_o$ $\pi_{Tik} = P(Y_{Tik} = w_k) = P(\tau_{kw_k} < Y^*_{Tik} < \tau_{k(w_k+1)})$. Therefore, multiple levels in the ordinal outcomes can be considered by selecting the appropriate τ thresholds. For instance, $\pi_{Tik} = P(Y_{Tik} = 0) + P(Y_{Tik} = 1) + P(Y_{Tik} = 2) = P(-\infty < Y^*_{Tik} < \tau_{k3})$. As the latent means are estimable by maximum likelihood, $\mu_{Ti1} = \Phi^{-1}(\pi_{Ti1}), \ldots, \mu^*_{TiK} = \Phi^{-1}(\pi_{TiK})$ in the treatment group and $\mu_{Ci1} = \Phi^{-1}(\pi_{Ci1}), \ldots, \mu^*_{CiK} = \Phi^{-1}(\pi_{CiK})$ in the control group.
We can proceed by specifying the hypothesis in (4.3) holds if and only if the hypothesis in (4.4) holds [134].

$$\begin{aligned} H_0^* &: \ \exists k \ s.t. \ \delta_k^* \leq 0 \\ H_1^* &: \ \delta_k^* > 0 \ \forall k \end{aligned} \tag{4.4}$$

where $\delta_k^* = \mu_{Tk}^* - \mu_{Ck}^*$. The maximum likelihood estimates $\hat{\mu}_{Tk}^*$ and $\hat{\mu}_{Ck}^*$ can be used and the variance can be obtained using the inverse of the Fisher information matrix.

4.5.2 Overall Power

Having specified the hypothesis in Section 4.5.1 to include ordinal outcomes, the power in this case is as defined for mixed continuous and binary co-primary endpoints [134], as shown in (4.5). This can be summarised as the overall probability of the standardised Z values exceeding the standardised z values for each of the K individual hypotheses. Note that the only difference between the case for the observed continuous and latent continuous outcomes is the $\sigma_k = 1$ for $k \geq k_{m+1}$ as assumed by the model.

$$1 - \beta = P\left(\bigcap_{k=1}^{k_m}\{Z_k > z_\alpha\} \bigcap_{k_{m+1}}^{K}\{Z_k^* > z_\alpha\} \mid \boldsymbol{\delta}\right) \simeq P\left(\bigcap_{k=1}^{K}\{Z_k^\dagger > z_k^\dagger\} \mid \boldsymbol{\delta}\right) \tag{4.5}$$

for $\boldsymbol{\delta} = (\delta_1, ..., \delta_{k_m}, ..., \delta_{k_o}, ..., \delta_K)^T \neq \mathbf{0}$ and

$$Z_k^\dagger = \begin{cases} Z_k - \dfrac{\delta_k}{\sigma_k}\sqrt{\dfrac{\kappa n_T}{1+\kappa}} = \dfrac{\bar{Y}_{Tk} - \bar{Y}_{Ck} - \delta_k}{\sigma_k\sqrt{\dfrac{1+\kappa}{\kappa n_T}}}, k = 1, ..., k_m \\ \\ Z_k^* - \delta_k^*\sqrt{\dfrac{\kappa n_T}{1+\kappa}} = \dfrac{\hat{\mu}_{Tk}^* - \hat{\mu}_{Ck}^* - \delta_k^*}{\sqrt{\dfrac{1+\kappa}{\kappa n_T}}}, k = k_{m+1}, ..., K \end{cases}$$

$$z_k^\dagger = \begin{cases} z_\alpha - \dfrac{\delta_k}{\sigma_k}\sqrt{\dfrac{\kappa n_T}{1+\kappa}}, k = 1, ..., k_m \\ \\ z_\alpha - \delta_k^*\sqrt{\dfrac{\kappa n_T}{1+\kappa}}, k = k_{m+1}, ..., K \end{cases}$$

where $\delta_k = \mu_{Tk} - \mu_{Ck}$, $\delta_k^* = \mu_{Tk}^* - \mu_{Ck}^*$, $\kappa = n_C/n_T$ and z_α is the $(1-\alpha)100^{th}$ standard normal percentile.

$$1-\beta \simeq P\left(\bigcap_{k=1}^{K}\{Z_k^{\dagger} > z_k^{\dagger}\} \mid \boldsymbol{\delta}\right) = \Phi_K\left(-z_1^{\dagger}, ..., -z_K^{\dagger}; \Gamma\right) \tag{4.6}$$

Therefore the power of the K co-primary endpoints can be evaluated using the K-dimensional cumulative normal distribution function as shown in (4.6). Assuming $n_T = n_C = n$, we can input different values for n to achieve the required power.

4.5.3 Sample Size Calculation

It is possible, as discussed in [116], to rearrange (4.6) to obtain a sample size formula in terms of n.

$$1-\beta \leq \int_{z_{1-\alpha}}^{\infty} \dots \int_{z_{1-\alpha}}^{\infty} f\left(z_1, \dots, z_{k_m}, z^*_{k_{m+1}}, \dots, z^*_K; \sqrt{nk}\boldsymbol{\delta}^{\dagger}, \Gamma\right) dz_1, \dots, dz^*_K \tag{4.7}$$

where $\boldsymbol{\delta}^{\dagger} = \left(\frac{\delta_1}{\sigma_1}, \dots, \frac{\delta_K}{\sigma_K}\right)$. This can also be expressed as:

$$n = \frac{(C_k + z_{1-\alpha})^2}{k\delta_k^2} \tag{4.8}$$

where C_k is the solution of

$$1-\beta = \int_{-\infty}^{\gamma_1 C_k + z_{1-\alpha}(\gamma_1 - 1)} \dots \int_{-\infty}^{\gamma_{k-1} C_k + z_{1-\alpha}(\gamma_{k-1} - 1)} \int_{-\infty}^{C_k} f\left(z_1, \dots, z^*_K; \mathbf{0}, \Gamma\right) dz^*_K \dots dz_1$$

4.5.4 Application to MUSE Trial

We can apply the theory to the four dimensional endpoint used in the MUSE trial [112] for systemic lupus erythematosus by assuming that we require a treatment effect in each of the four endpoints, rather than forming a composite endpoint. As before Y_1 is the continuous SLEDAI outcome, Y_2 is the continuous PGA outcome, Y_3 is the observed ordinal BILAG measure assumed to come from latent Y_3^* with levels $w = 5$ and Y_4 is the binary taper variable arising from latent Y_4^*.

We can use the MUSE trial to design a future study that assumes a treatment effect in each of the outcomes is required to conclude that the treatment is beneficial. Table 4.1 shows the sample sizes required in each group, for the co-primary endpoint to obtain an overall power of at least 80% to detect a difference of $\delta_1 = 0.88$ in SLEDAI, $\delta_2 = 0.38$ in PGA, $\delta_3 = 0.24$ in BILAG and $\delta_4 = 0.40$ in the taper outcome at alpha level $\alpha = 0.025$,

Table 4.1: Sample sizes $n = n_C = n_T$ for the systemic lupus erythematosus co-primary endpoint for overall power $1 - \beta \approx 0.80$, $\alpha = 0.025$, $k_2 = 2, k_o = K = 1$. SS_1, SS_2, SS_3, SS_4 are sample sizes required per group for the individual endpoints for a power of at least $1 - \beta = 0.80$

SLEDAI		PGA		BILAG		Taper		n	SS_1	SS_2	SS_3	SS_4
δ_1	σ_1^2	δ_2	σ_2^2	(π_{T3}, π_{C3})	δ_3^*	(π_{T4}, π_{C4})	δ_4^*					
0.88	18	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	365	39	273	99
0.88	19	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	419	386	39	273	99
0.88	20	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	435	406	39	273	99
0.88	18	0.38	0.45	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	365	18	273	99
0.88	18	0.38	0.55	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	365	22	273	99
0.88	18	0.38	0.65	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	365	26	273	99

based on the values observed in the trial. We also allow for uncertainty in the variance of the continuous measures by setting $\sigma_1^2 = 18, 19, 20$ and $\sigma_2^2 = 0.35, 0.45, 0.55, 0.65$. The sample sizes required for each individual endpoint are also shown in Table 4.1 based on achieving a power of at least 80%.

Sample sizes required for the individual endpoints vary based on the different assumed treatment effects. The sample size required for SLEDAI is 365 per arm based on an effect size of 0.21, PGA would require 39 per arm to detect an effect size of 0.64, BILAG would require 273 per arm for effect size of 0.24 and powering for the taper variable would only require 99 per arm to detect an effect size of 0.40. The sample sizes shown for the co-primary endpoint range from 403 to 435 per group, based on the different variances assumed for the SLEDAI outcome. Changing the variance assumed for the PGA outcome between $\sigma_2^2 = 0.35$ and $\sigma_2^2 = 0.65$ does not change the number of patients required for the co-primary endpoint. The main factor driving the required sample size for the co-primary endpoint is the individual endpoint requiring the largest number of patients for a given treatment effect and power. Note that as the original MUSE trial, which was designed to detect differences based on the composite endpoint required n=100, would have been underpowered to show a statistical significance of the co-primary endpoints which require n=403.

Figure 4.1 shows the power for the co-primary endpoints and each of the individual endpoints for different sample sizes based on the effects in the MUSE trial. Overall for the individual endpoints, the power is largest across all sample sizes for the PGA outcome and lowest for the SLEDAI outcome, as expected from the assumed effect sizes. The sample size required for a given power is largest for the co-primary endpoints,

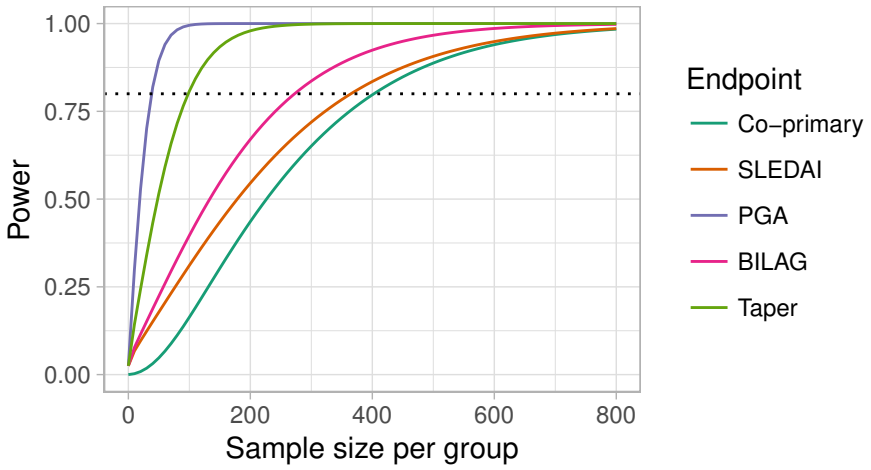


Figure 4.1: Overall power $1 - \beta$ to detect the treatment effects assumed from the MUSE trial for the systemic lupus erythematosus co-primary endpoints SLEDAI, PGA, BILAG and Taper and power for individual endpoints for different sample sizes per group $n = n_C = n_T$

which is intuitive given it is the number of patients required to show a statistical significance in all four of the outcomes. Figure 4.2 shows the resulting power for a range of sample sizes when the co-primary endpoints are PGA, BILAG and taper and when the co-primary endpoints are PGA and taper. The results agree with the four outcome scenario, where the power to detect significant differences in the co-primary endpoints from a given sample size is either slightly less than or equivalent to the power to detect the smallest effect size in the individual outcomes.
Figure 4.3 shows the resulting power to detect the stated treatment effects in the co-primary endpoints for different correlations between the endpoints. The power is lowest for a given sample size when the correlation between the endpoints is zero and rises as the correlation increases. To achieve 80% power, the required sample size ranges from approximately 385 per group to 415 per group.

4.6 Mixed Outcome Composite Endpoints

In practice, most studies using composite endpoints are aiming to detect a difference in probability in response between arms rather than show all components are different. Consequently, a sample size calculation using the latent variable model for composite endpoints will have to take this in to account. We can begin by assuming the latent variable structure. Let $\mathbf{Y_i}$ be the vector of outcomes for patient i, where the first

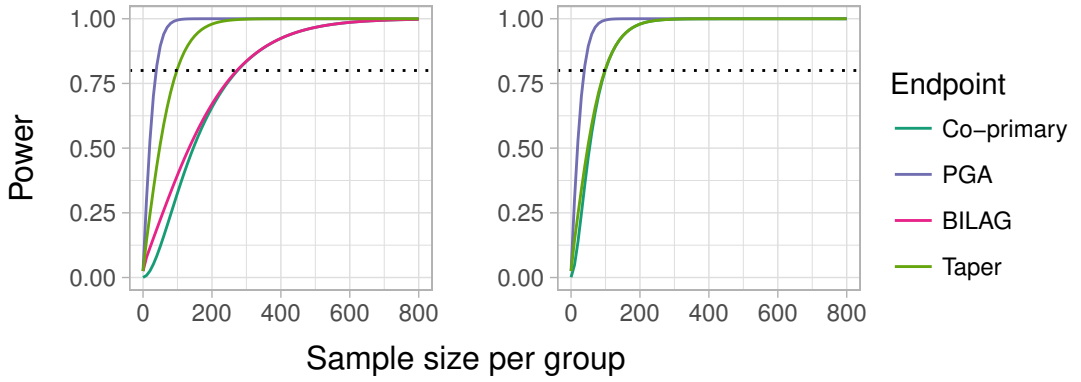


Figure 4.2: Overall power $1 - \beta$ to detect the treatment effects assumed from the MUSE trial for the systemic lupus erythematosus co-primary endpoints and individual endpoints for different sample sizes per group $n = n_C = n_T$ for co-primary endpoints PGA, BILAG and Taper (left) and co-primary endpoints PGA and Taper (right)

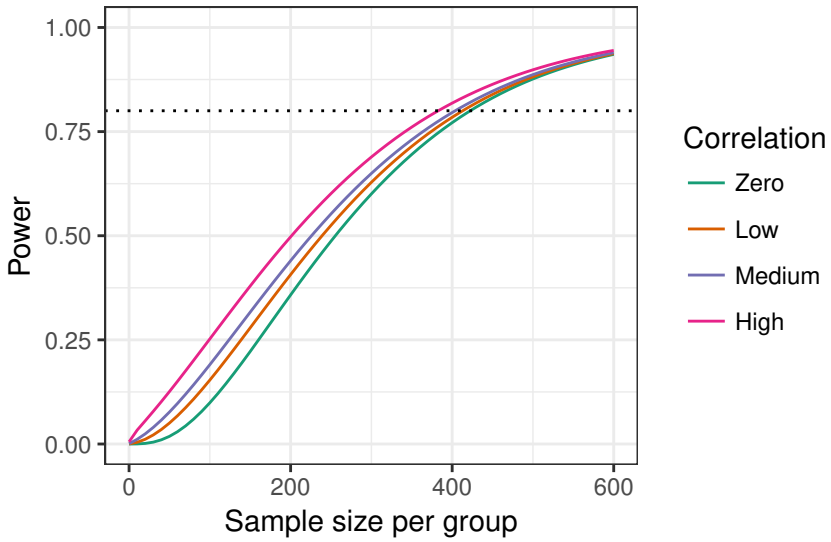


Figure 4.3: Overall power $1 - \beta$ to detect the treatment effects assumed from the MUSE trial for the systemic lupus erythematosus co-primary endpoints for different sample sizes per group $n = n_C = n_T$ and differing correlations between outcomes, where Low=0.3, Medium=0.5 and High=0.8

$1 \leq k \leq k_m$ elements are observed as continuous variables, the next $k_m < k \leq k_o$ are observed as ordinal and the remaining $k_o < k \leq K$ are observed as binary. We can specify p_{iT} and p_{iC}, the probability of response for patient i in the treatment and control arm respectively, as shown below.

$$p_{iT} = P(S_i = 1|T_i = 1) = \int_{-\infty}^{\eta_1} \dots \int_{-\infty}^{\eta_K} f_{Y_1,\dots,Y_K}(y_{i1}, \dots, y_{iK}|T_i = 1, \boldsymbol{\theta}) dy_K \dots dy_1$$

$$p_{iC} = P(S_i = 1|T_i = 0) = \int_{-\infty}^{\eta_1} \dots \int_{-\infty}^{\eta_K} f_{Y_1,\dots,Y_K}(y_{i1}, \dots, y_{iK}|T_i = 0, \boldsymbol{\theta}) dy_K \dots dy_1$$

The quantities $(\eta_1, \dots, \eta_K)$ are the predefined responder thresholds and $\boldsymbol{\theta}$ is the vector of model parameters. We can assume that $p_T \sim N(\delta_T, \sigma_{\delta_T})$ and $p_C \sim N(\delta_C, \sigma_{\delta_C})$. As in the case of co-primary endpoints, the assumptions allow us to estimate latent means $(\mu^*_{k_{m+1}}, \dots, \mu^*_K)$ for the observed discrete components using the model parameters.

4.6.1 Hypothesis Testing

An important consideration for hypothesis testing in the mixed outcome composite endpoint setting is that whilst we are exploiting the latent multivariate Gaussian structure for precision gains, we are ultimately still interested in the one dimensional endpoint. This is distinct from the co-primary endpoint case where the overall hypothesis test must be based on some union or intersection of the hypotheses for the individual outcomes. This is illustrated in Figure 4.4, which compares the different stages in analysis and hypothesis testing for the composite endpoint using the binary and latent variable methods and for co-primary endpoints using the latent variable model.

For the composite endpoint, the test statistic of interest is $\frac{\bar{p}_T - \bar{p}_C}{\sigma_\delta}$ where $\sigma_\delta = \sqrt{\frac{\sigma_{\delta_T}}{n_T} + \frac{\sigma_{\delta_C}}{n_C}}$, n_T is the number of patients in the treatment group and n_C is the number of patients in the control group. Therefore, we can formulate the hypothesis as shown in (4.9).

$$\begin{aligned} H_0 &: \ \delta = \mu_0 \\ H_1 &: \ \delta \neq \mu_0 \end{aligned} \tag{4.9}$$

In this instance we consider a risk difference however we can also state the hypothesis of interest in the form of a risk ratio or odds ratio. For sample size estimation, we require the distribution of the test statistic δ under H_1, which we can assume to be $\delta \sim N(\delta_T - \delta_C, \sigma^2_\delta)$. The delta method, based on Taylor series expansion, is a useful technique for determining approximate distributions. The theory states that if we have

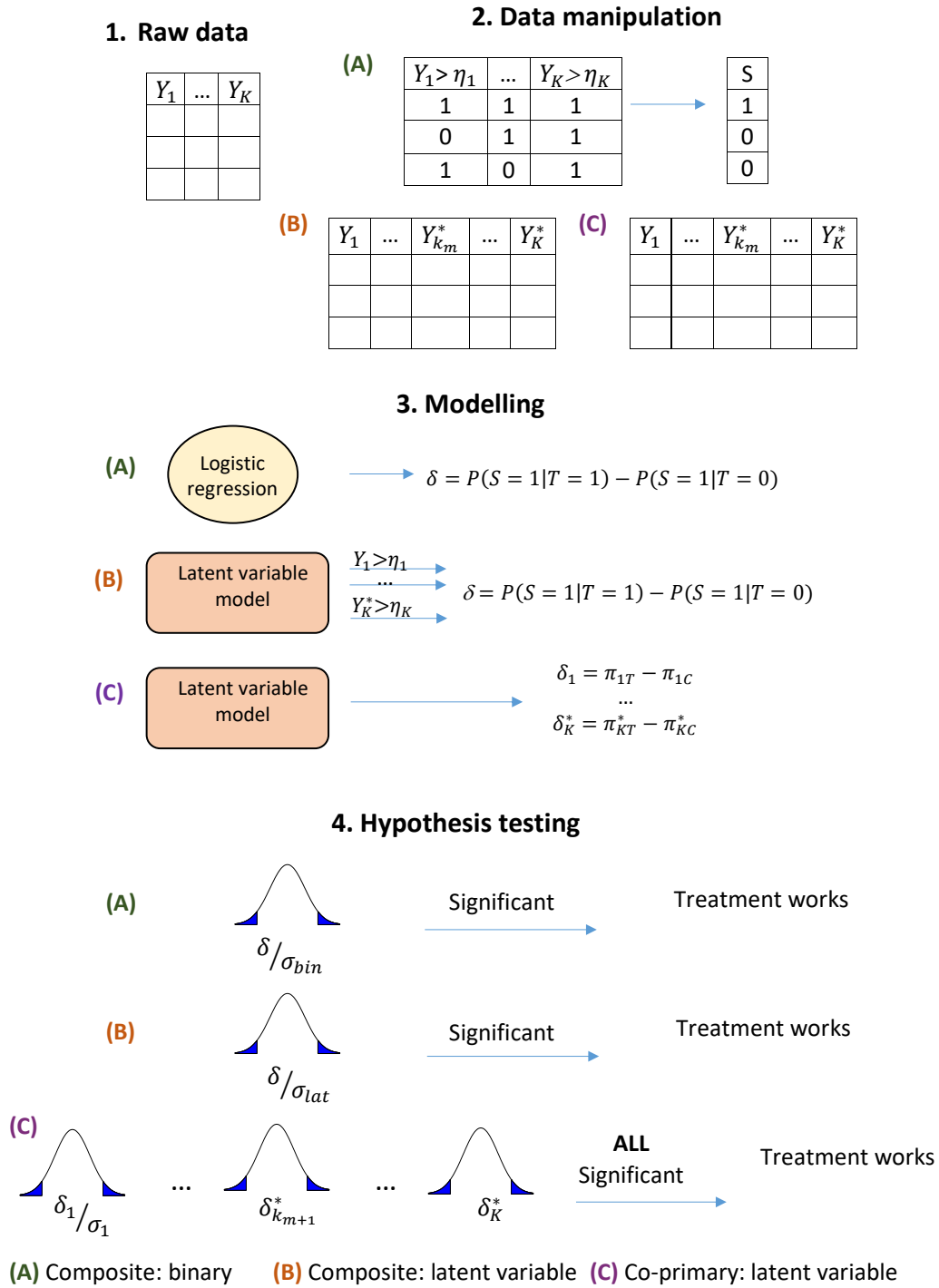


Figure 4.4: Stages in analysis and hypothesis testing for a composite using the standard binary method, a composite using the latent variable method and co-primary endpoints using the latent variable model, where $Y_1 ... Y_K$: observed outcomes, $Y^*_{k_m} ... Y^*_K$: latent outcomes, $\eta_1 ... \eta_K$: response thresholds, S: overall response, δ: treatment effect, σ_{bin}: standard error from binary method, σ_{lat}: standard error from latent variable method

a random variable $X \sim N(\mu, \sigma^2)$ and W= $g(X)$, where $g(\mu) \neq 0$, then W will have an approximate normal distribution which may be found using the usual rules for linear transformations of normals [135]. To first order:

$$\begin{aligned} E(W) \approx & g'(\mu)\mu + g(\mu) - g'(\mu)\mu \\ = & g(\mu) \end{aligned} \tag{4.10}$$

$$\begin{aligned} var(W) \approx var(g(X)) = & (g(X) - g(\mu))^2 \\ = & (g'(\mu)(X-\mu))^2 \\ = & (g'(\mu))^2(X-\mu)^2 \\ = & (g'(\mu))^2 var(X) \end{aligned} \tag{4.11}$$

Therefore, in our case we can evaluate δ_T and δ_C as shown in (4.12).

$$\begin{aligned} \delta_T = & \Phi_K(\eta_1, \cdots, \eta_K; \boldsymbol{\mu}_T, \Sigma_T) \\ \delta_C = & \Phi_K(\eta_1, \cdots, \eta_K; \boldsymbol{\mu}_C, \Sigma_C) \end{aligned} \tag{4.12}$$

where $\Phi_K(.; \boldsymbol{\mu}, \Sigma)$ is the K dimensional multivariate normal distribution function, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Estimates of the quantities can be obtained using the estimated model parameters $\hat{\boldsymbol{\theta}}$. Namely, δ_T is estimated by $\hat{\delta}_T = \Phi_K(\eta_1, \cdots, \eta_K; \hat{\boldsymbol{\mu}}_T^*, \widehat{\Sigma}_T)$ and δ_C is estimated by $\hat{\delta}_C = \Phi_K(\eta_1, \cdots, \eta_K; \hat{\boldsymbol{\mu}}_C^*, \widehat{\Sigma}_C)$, where $\boldsymbol{\mu}_{\boldsymbol{T}}^*$ is the K-dimensional vector of mean values in the treatment arm and $\boldsymbol{\mu}_{\boldsymbol{C}}^*$ is the corresponding vector for the control arm. Note that for $k \leq k_m$ the quantities are observed and for $k > k_m$ the quantities are latent.
Given that we are interested in a function of $\hat{\boldsymbol{\theta}}$, $\hat{\delta} = \Phi_K(\eta_1, \cdots, \eta_K; \hat{\boldsymbol{\mu}}_T^*, \widehat{\Sigma}_T) - \Phi_K(\eta_1, \cdots, \eta_K; \hat{\boldsymbol{\mu}}_C^*, \widehat{\Sigma}_C)$, and can obtain $Cov(\hat{\boldsymbol{\theta}})$, then we can use (4.11) to obtain the quantity σ_δ^2 as follows.

$$var(\hat{\delta}) \approx (\boldsymbol{''\delta})^T Cov(\hat{\boldsymbol{\theta}})(\boldsymbol{''\delta}) \tag{4.13}$$

which is estimated by $\widehat{var}(\hat{\delta}) = (\boldsymbol{''\delta_T})^T \widehat{Cov}(\hat{\boldsymbol{\theta}})(\boldsymbol{''\delta_T})$, where $\boldsymbol{''\delta}$ is the vector of partial derivatives of δ with respect to each of the parameter estimates. One potential difficulty for conducting sample size estimation using the latent variable model in practice is that the vector of model parameters $\boldsymbol{\theta}$ may be large depending on the number of outcomes, with certain quantities difficult to elicit such as the biserial correlation between binary outcomes.

4.6.2 Obtaining Required Quantities

An important aspect of the sample size calculation is defining the target difference. Hislop et al. [136] performed a systematic review to determine how the target difference is determined in RCT sample size calculations in the literature. They found seven different methods, employed to varying degrees, which were used for specifying this difference. A more recent literature review and Delphi study agreed with the findings [137]. We briefly summarise the methods suggested in Table 4.2.
In order to obtain the distribution of the test statistic under the alternative, we need estimates for $\boldsymbol{\theta}$ and $Cov(\boldsymbol{\theta})$. From the methods suggested in Table 4.2, we can use pilot trial data to obtain the target difference, hence also obtaining parameter estimates $\hat{\boldsymbol{\theta}}$ and an estimate for their covariance matrix $\widehat{Cov}(\hat{\boldsymbol{\theta}})$. If all model parameters and their covariance matrix could be specified, fitting the model on pilot data would not be required, however this would be difficult in practice.

4.6.3 Critical Value

To test the hypothesis in (4.9), we need to determine the critical value, cv. As the endpoint of interest is specified in terms of the overall one dimensional composite endpoint, we can do this using the formula used when employing the standard binary method and the approximation for the distribution of the test statistic δ under H_1.

$$\begin{aligned}
\alpha/2 &= P(\bar{p}_T - \bar{p}_C \geq cv \mid H_0) \\
&= 1 - P\left(\frac{\bar{p}_T - \bar{p}_C - \mu_0}{\sigma_\delta} \leq \frac{cv - \mu_0}{\sigma_\delta} \mid H_0\right) \\
&= 1 - \Phi\left(\frac{cv - \mu_0}{\sigma_\delta}\right) \\
z_{\alpha/2} &= \frac{cv - \mu_0}{\sigma_\delta} \\
cv &= \mu_0 + \sigma_\delta z_{\alpha/2}
\end{aligned}$$

4.6.4 Power

Let us assume that $\sigma_T = \sigma_C = \sigma$ and $n_T = n_C = n$, so that $\delta \sim N(\delta_T - \delta_C, 2\sigma^2/n)$. We can determine the power in the standard way by using the critical value, as demonstrated below.

Table 4.2: Methods for determining the target difference in a sample size calculation, identified in a systematic review of randomised controlled trials along with advantages and disadvantages in the application of each

Method	Definition	Advantages	Disadvantages
Anchor	Chosen using patient or professional judgement	• Could include patient & clinician perspective	• Possible recall bias and response shift
Distribution	Value based on using distributional variation	• Takes account of uncertainty	• Difficult to translate to target disease
Health Economic	Any method using the principle of economic evaluation e.g. decision theory	• Accounts for resources	• Complex to implement
Standardised effect size	Magnitude of effect is defined on standardised scale	• Easy to compare across studies and and conditions	• Difficult to establish why different effect sizes are observed
Pilot study	Obtains estimate through a smaller study	• Estimated from relevant data	• Imprecise effect size
Opinion-seeking	Eliciting plausible values from individuals e.g. patients, clinicians	• Relatively easy to implement	• Not representative of wider community
Evidence base	Uses current evidence, ideally a meta-analysis of existing RCTs	• Provides important and/or realistic difference	• May not be appropriate for the new population

$$
\begin{aligned}
1-\beta &= P\left(\bar{p}_T - \bar{p}_C \geq \mu_0 + z_{\alpha/2}\sqrt{2\sigma^2/n} \mid H_1\right) \\
&= 1 - P\left(\frac{\bar{p}_T - \bar{p}_C - \delta}{\sqrt{2\sigma^2/n}} \leq \frac{\mu_0 + z_{\alpha/2}\sqrt{2\sigma^2/n} - \delta}{\sqrt{2\sigma^2/n}} \mid H_1\right) \\
&= 1 - \Phi\left(\frac{\mu_0 + z_{\alpha/2}\sqrt{2\sigma^2/n} - \delta}{\sqrt{2\sigma^2/n}}\right) \\
&= \Phi\left(\frac{\delta - \mu_0}{\sqrt{2\sigma^2/n}} - z_{\alpha/2}\right)
\end{aligned}
$$

4.6.5 Sample Size Estimation

Note that $\frac{2\sigma^2}{n} = \sigma_\delta^2$, however for sample size estimation we will need to separate n from the variance estimate. Although obtaining the variance using the delta method gives us the quantity σ_δ^2, by fitting the model to pilot trial data we can obtain σ^2 as we know the value of n in this instance. We can get an estimate for the sample size by rearranging the standard power formula, as shown below.

$$
\begin{aligned}
1-\beta &= \Phi\left(\frac{\delta - \mu_0}{\sqrt{\frac{2\sigma^2}{n}}} - z_{\alpha/2}\right) \\
z_{1-\beta} &= \frac{\delta - \mu_0}{\sqrt{\frac{2\sigma^2}{n}}} - z_{\alpha/2} \\
\frac{2\sigma^2}{n}(z_{1-\beta} + z_{\alpha/2})^2 &= (\delta - \mu_0)^2 \\
n &= \frac{2\sigma^2(z_{1-\beta} + z_{\alpha/2})^2}{(\delta - \mu_0)^2}
\end{aligned}
$$

In summary, as we are interested in the one dimensional overall composite endpoint, the sample size formula is the same as the binary case. However, due to exploiting the latent structure the values of δ and σ_δ used are likely to be quite different to those used in the standard binary case. Therefore, the main challenge with sample size estimation using the latent variable model is the specification of the model parameters and their covariance matrix. Using existing pilot trial data or data from an earlier phase study will alleviate these problems, however this comes with the assumption that the existing data provides reasonable estimates for the future study. In practice it is perhaps wise to use conservative parameter and covariance estimates hence reducing the sample size

from that required for the binary method, along with the potential for higher power if the estimates are too conservative in reality.

4.7 Empirical Comparisons

The simulation study in Chapter 3 highlighted the importance of data structure in the potential precision gains available. One crucial element is the factors that drive response, which we will investigate in relation to sample size. Furthermore, we consider the relationship between the correlation structure of the components and required sample size for different combinations of endpoints.

4.7.1 One Continuous, One Ordinal, One Binary

We begin by considering the case where the composite is a combination of one continuous, one ordinal and one binary outcome and the components are dichotomised at their mean, so that all three drive response. We use the median variance estimate from 1000 simulated datasets to estimate σ^2. Figure 4.5 shows the change in estimated sample size per group as the overall treatment effect on the composite endpoint changes and everything else remains constant. This is shown for a range of correlations between the three endpoints. The sample size required for the latent variable method generally decreases slightly as the correlation between the endpoints increases. However, there is also a small rise in sample size required when the correlation is high. This is not true for the binary method, which requires an increasing sample size as the correlation increases. The sample size decreases exponentially as the treatment effect increases for both methods.

Figure 4.6 shows the boxplots of the estimated variance for the standard binary and latent variable methods from 1000 simulated datasets for a range of correlation between the endpoints when all components drive response. The estimated variance reported from the latent variable method is always smaller than that obtained from the standard binary method. The difference in variance from the methods is smallest when there is zero correlation between the outcomes. The variation in variance estimates is larger for the binary method. Figure 4.7 shows the corresponding boxplots in terms of the sample size required per group using the latent variable method and standard binary method for different correlations between components. Figure 4.8 shows violin plots of the sample size required per group from the standard binary and latent variable

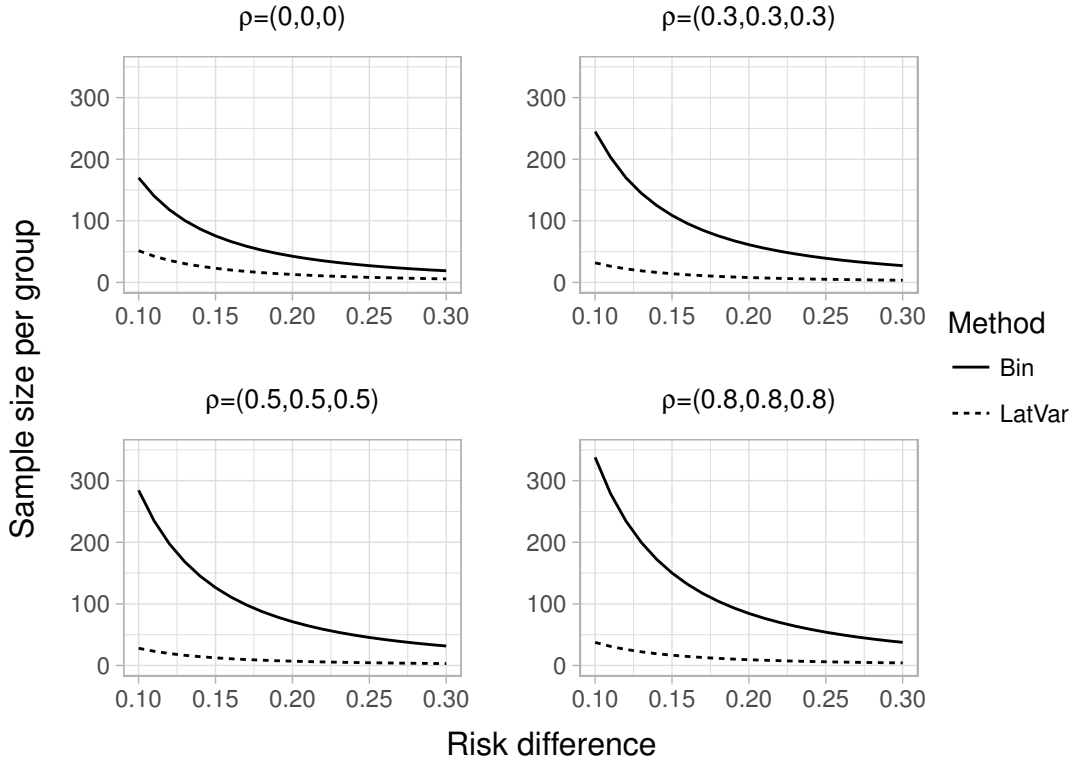


Figure 4.5: Estimated sample size per group for different values of the risk difference using the latent variable and standard binary methods when the composite endpoint is formed from one continuous, one ordinal and one binary outcome, where all components drive response and correlations between the outcomes are between 0 and 0.8, where $\rho = (\rho_{12}, \rho_{13}, \rho_{23})$

methods for different treatment effect structures in the components. For both methods the sample size does not appear to depend on the underlying treatment effect structure in the components. Instead it depends only on the overall treatment effect in the composite outcome.

Figure 4.9 shows boxplots of the estimated reduction in required sample size when employing the latent variable method rather than the standard binary method for a range of correlations between the endpoints. This is shown for the scenarios where response is driven by all three components, the continuous and binary components and the binary component only, where the difference between treatment arms is 0.08. In the case where all three components drive response and the correlation between the endpoints is zero, the latent variable method can reduce the sample size by 18-77%. However when there is correlation between the endpoints, the sample size is reduced by 80-90%. A similar pattern occurs when response is driven by the continuous and binary components, however there is a small drop in efficiency. For example, when the correlation between endpoints is low, the reduction in required sample size drops from 85% to approximately 77%, indicating that the ordinal component with 5 levels contributes to the increased precision. When the binary component is the only driver of response and there is no correlation between the endpoints, the median sample size required is the same for both methods. However, when the binary component is the only driver of response and there is correlation present between the endpoints, the latent variable method offers precision gains over the standard binary method. The magnitude of the gain depends on the strength of the correlation between the endpoints, where a higher correlation results in a larger reduction in required sample size.

Table 4.3 shows the median estimated sample size per group using the latent variable method for different treatment effect structures under a range of correlation assumptions. Generally, increasing the correlation between endpoints results in a smaller sample size estimate, however this is not true for high correlations between endpoints, where the sample size increases. The corresponding results for the binary method are shown in Table 4.4. When all three components drive response, increasing the correlation between the endpoints results in a larger sample size. When response is driven by the binary component only, the sample size is unaffected by correlation between the endpoints. Table 4.5 shows the empirical power of the method for a range of sample sizes and correlations where response is driven by (Y_1, Y_2, Y_3), (Y_1, Y_3) and (Y_3) for overall treatment effect $\delta = 0.05, 0.10$ and 0.15. This was obtained from 5000 samples

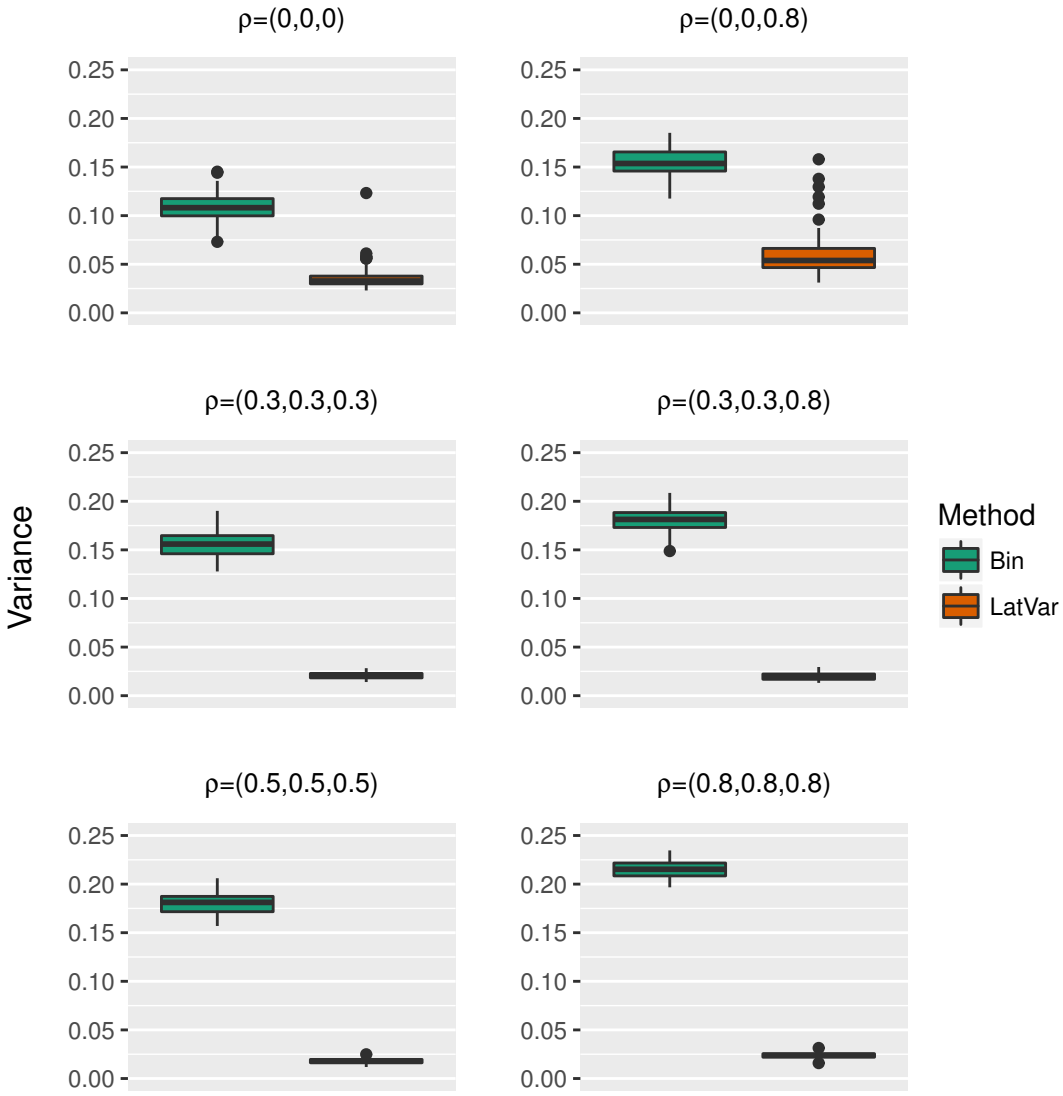


Figure 4.6: Boxplots of the estimated variance from 1000 simulated datasets for the standard binary and latent variable methods for a range of correlations between one continuous, one ordinal and one binary measure when all outcomes have equal treatment effect and drive response and $\rho = (\rho_{12}, \rho_{13}, \rho_{23})$

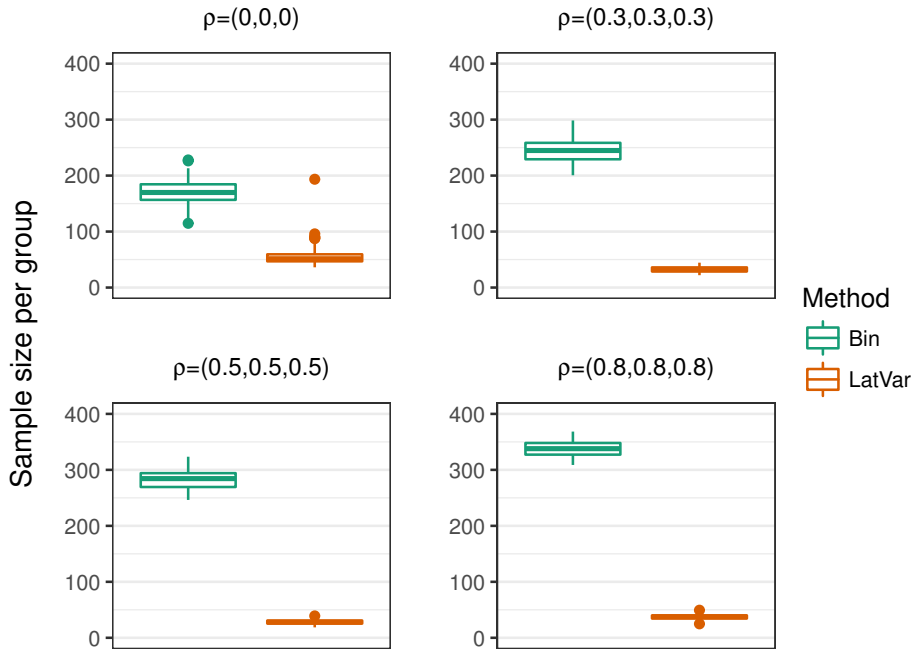


Figure 4.7: Boxplots of the estimated sample size per group from 1000 simulated datasets for the standard binary and latent variable methods for a range of correlations between one continuous, one ordinal and one binary measure when all outcomes have equal treatment effect and drive response and where $\rho = (\rho_{12}, \rho_{13}, \rho_{23})$

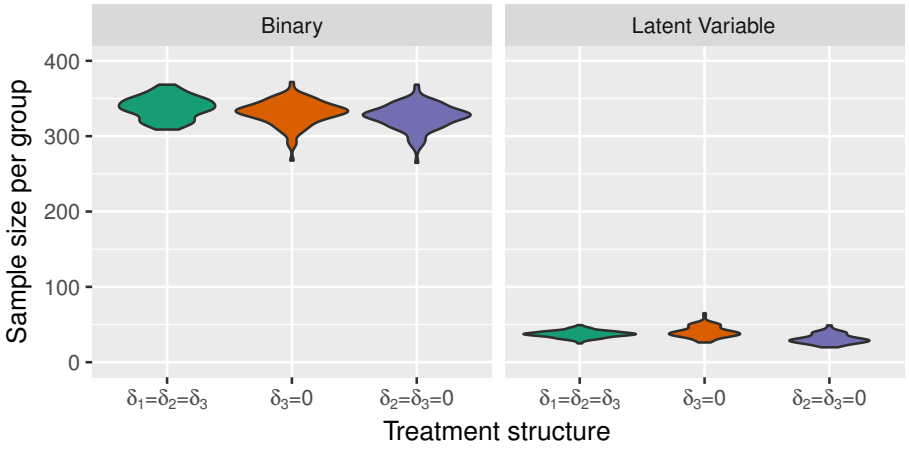


Figure 4.8: Violin plots of the estimated sample size from 1000 simulated datasets for the latent variable and standard binary methods for different treatment effect structures between one continuous, one ordinal and one binary measure when all outcomes drive response

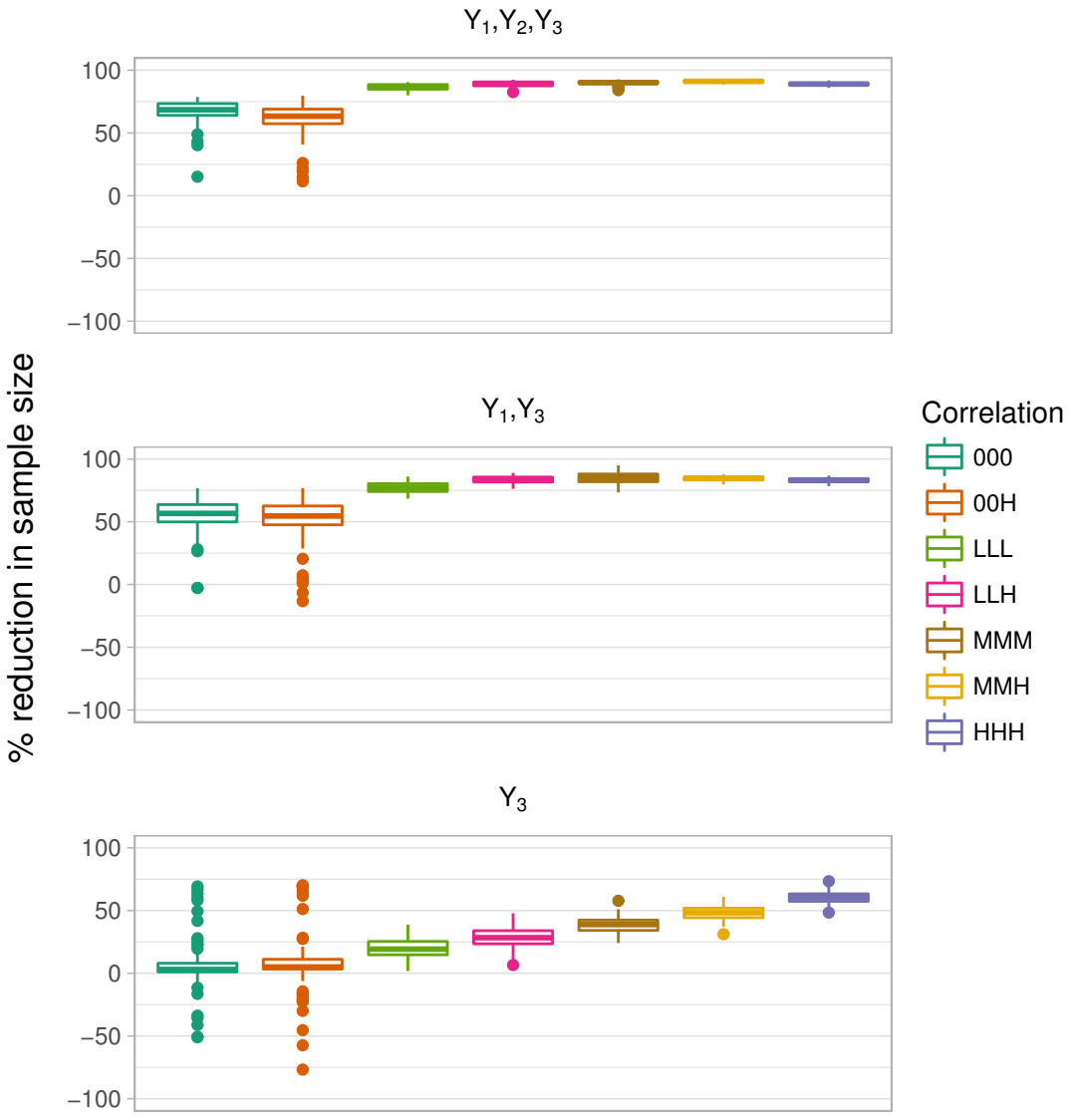


Figure 4.9: Boxplots of the estimated reduction in required sample size from employing the latent variable method instead of the standard binary method for a range of correlations between one continuous, one ordinal and one binary measure when all outcomes have equal treatment effect. Response is driven by all components in the top panel, the continuous and binary in the middle panel and only the binary in the bottom panel

Table 4.3: Median sample sizes $n = n_C = n_T$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.05$, $k_m = k_o = K = 1$, $\delta = \mu_T - \mu_C$: overall risk difference on the composite, $\delta^\dagger$: treatment effect structure in the components, for a combination of correlations 0, L=0.3, M=0.5, H=0.8 using the latent variable model when the composite is made up of one continuous, one ordinal and one binary outcome

Response	$\delta^\dagger$	δ	Correlation $(\rho_{12}, \rho_{13}, \rho_{23})$						
			000	00H	LLL	LLH	MMM	MMH	HHH
Y_1, Y_2, Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	206	339	128	119	112	108	145
		0.10	52	85	32	30	28	27	38
		0.15	23	38	15	14	13	12	17
	$\delta_3 = 0$	0.05	201	334	123	117	105	112	159
		0.10	51	84	31	30	27	28	40
		0.15	23	38	14	13	12	13	18
	$\delta_2 = \delta_3 = 0$	0.05	197	329	119	114	101	106	121
		0.10	49	81	29	29	27	21	30
		0.15	21	36	14	13	12	12	15
Y_1, Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	474	494	289	200	267	205	240
		0.10	119	124	73	50	67	52	60
		0.15	53	55	33	23	30	23	27
	$\delta_3 = 0$	0.05	468	499	286	195	264	203	248
		0.10	117	125	72	49	66	51	62
		0.15	52	56	32	22	30	23	28
	$\delta_2 = \delta_3 = 0$	0.05	470	501	287	196	264	204	249
		0.10	119	123	73	49	67	51	63
		0.15	53	55	32	23	31	24	28
Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	1493	1472	1250	1113	948	793	609
		0.10	374	368	313	279	237	199	153
		0.15	166	164	139	124	106	89	68
	$\delta_3 = 0$	0.05	1502	1468	1256	1113	960	806	622
		0.10	376	367	315	279	240	202	156
		0.15	176	164	140	124	107	90	70
	$\delta_2 = \delta_3 = 0$	0.05	1504	1465	1259	1115	963	807	624
		0.10	376	370	316	280	241	203	156
		0.15	174	164	139	126	106	90	70

Table 4.4: Median sample sizes $n = n_C = n_T$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.05$, $k_m = k_o = K = 1$, $\delta = \mu_T - \mu_C$: overall risk difference on the composite, $\delta^\dagger$: treatment effect structure in the components, for a combination of correlations ranging from 0, L=0.3, M=0.5, H=0.8 using the standard binary method when the composite is made up of one continuous, one ordinal and one binary outcome

Response	$\delta^\dagger$	δ	Correlation $(\rho_{12}, \rho_{13}, \rho_{23})$						
			000	00H	LLL	LLH	MMM	MMH	HHH
Y_1, Y_2, Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	680	965	980	1141	1138	1214	1352
		0.10	170	242	245	286	285	304	338
		0.15	76	108	109	127	127	135	151
	$\delta_3 = 0$	0.05	628	939	928	1098	1102	1183	1332
		0.10	157	235	232	275	276	296	333
		0.15	70	105	104	122	123	132	148
	$\delta_2 = \delta_3 = 0$	0.05	609	920	914	1086	1097	1171	1310
		0.10	147	231	228	270	271	290	328
		0.15	68	101	101	119	121	130	146
Y_1, Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	1127	1136	1255	1261	1334	1320	1425
		0.10	282	284	314	316	334	330	357
		0.15	126	127	140	141	149	147	159
	$\delta_3 = 0$	0.05	1078	1072	1216	1218	1298	1296	1403
		0.10	270	268	304	305	325	324	351
		0.15	120	120	136	136	145	144	156
	$\delta_2 = \delta_3 = 0$	0.05	1066	1063	1202	1209	1296	1297	1403
		0.10	263	259	300	299	319	324	351
		0.15	121	119	133	131	143	143	156
Y_3	$\delta_1 = \delta_2 = \delta_3$	0.05	1547	1548	1550	1548	1550	1550	1549
		0.10	387	387	388	387	388	388	388
		0.15	172	172	173	172	173	173	173
	$\delta_3 = 0$	0.05	1544	1545	1549	1548	1549	1548	1545
		0.10	386	387	388	387	388	387	387
		0.15	172	172	173	172	173	172	172
	$\delta_2 = \delta_3 = 0$	0.05	1544	1546	1549	1546	1551	1549	1545
		0.10	386	389	387	385	388	388	387
		0.15	173	172	173	173	172	172	172

Table 4.5: Empirical power (%) for $n = n_C = n_T$, $\alpha = 0.05$, $\delta = \mu_T - \mu_C$: overall risk difference on the composite, $\delta_1 = \delta_2 = \delta_3$, for a combination of correlations ranging from 0, L=0.3, M=0.5, H=0.8 using the latent variable method when the composite is made up of one continuous, one ordinal and one binary outcome

Response	δ	n=50			n=100			n=200		
		000	MMM	HHH	000	MMM	HHH	000	MMM	HHH
Y_1, Y_2, Y_3	0.05	79.1	80.1	80.3	80.0	80.0	80.3	80.5	80.1	79.8
	0.10	80.1	80.4	80.1	80.0	80.4	79.9	80.1	80.0	80.2
	0.15	80.8	80.9	80.4	80.2	80.5	80.0	80.3	80.2	80.4
Y_1, Y_3	0.05	79.8	80.2	80.1	79.9	80.1	80.3	80.2	79.5	80.1
	0.10	80.1	80.3	80.0	79.9	80.2	80.1	80.0	80.0	79.8
	0.15	80.2	80.4	80.3	80.1	80.2	80.7	80.3	80.1	80.0
Y_3	0.05	80.1	79.7	80.2	80.1	79.2	80.4	80.0	80.1	79.9
	0.10	80.4	80.0	80.3	79.7	80.1	80.2	80.4	80.5	80.2
	0.15	80.0	80.2	80.1	80.2	80.2	80.0	80.1	80.1	80.2

from the latent variable model by fitting the method and determining the proportion of confidence intervals that do not contain zero. The empirical power is close to the desired power of 80% across the cases investigated.

4.7.2 Two Continuous, One Ordinal, One Binary

As well as considering a composite endpoint made up of one of each type of component, it is interesting to consider the efficiency gains from an additional continuous component. In this instance Y_1 and Y_2 are continuous measures, Y_3 is ordinal and Y_4 is binary. Figure 4.10 compares the boxplots of sample sizes required from the latent variable method when the composite has one continuous, one ordinal and one binary outcome and when an additional continuous component is added. The results show that when using the latent variable method, adding a second continuous component which also drives treatment response can reduce the median required sample size by a further 46-58% for the different correlation structures investigated.

The boxplots of the estimated percentage reduction in required sample size from using the latent variable method rather than the standard binary method is shown in Figure 4.11. When the correlation between the components is zero, the median reduction in sample size is 80%. For any correlation between the outcomes, the median reduction in required sample size is approximately 94%. The reduction in required sample size is the same when the binary component is included and when it is removed. However,

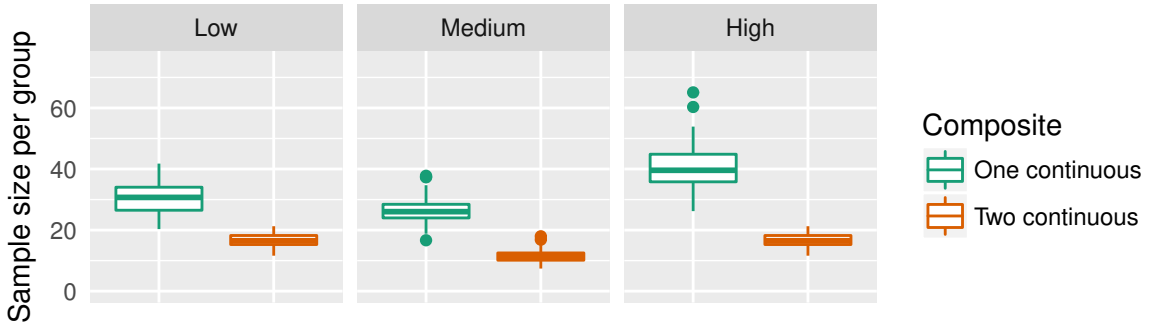


Figure 4.10: Boxplots comparing the estimated sample size from 1000 simulated datasets using the latent variable method for composites containing one continuous, one ordinal and one binary component and composites containing two continuous, one ordinal and one binary component, when all outcomes drive response. Correlations between components are low=0.3, medium=0.5 and high=0.8, the risk difference between treatment arms in 0.2 and the treatment effect is the same on all components.

this is not true for the ordinal component, as the percentage reduction in sample size is smaller when the ordinal component is removed. These results indicate that most of the efficiency gains are obtained from the continuous measures and only a very small amount of this is from the ordinal variable.

Table 4.6 and Table 4.7 show the median sample size per group required for power equal to 80% and $\alpha = 0.05$ for different combinations of correlations, treatment effects and drivers of response, when using the latent variable method and standard binary method respectively. The findings for the latent variable method from Table 4.6 are visualised in Figure 4.12. From this we can see that the sample sizes required are similar across different treatment effect structures in the components, including when the effects of components are in different directions as in $\delta_1 = -\delta_2$. Based on the theory introduced in Chapter 1 we would have expected for the treatment effect structure within the components to have had a more substantial impact on the sample size required, particularly when the treatment effect on different components are in opposite directions. The sample sizes required are similar when response is driven by (Y_1, Y_2, Y_3, Y_4) and (Y_1, Y_2, Y_3). In this setting the sample size is largest for zero correlation and reduces when the components are correlated. However, the median sample size is smaller for lower correlations between outcomes and increases slightly for larger correlations. Sample sizes are similar for all correlations when response is driven by (Y_1, Y_2, Y_4).

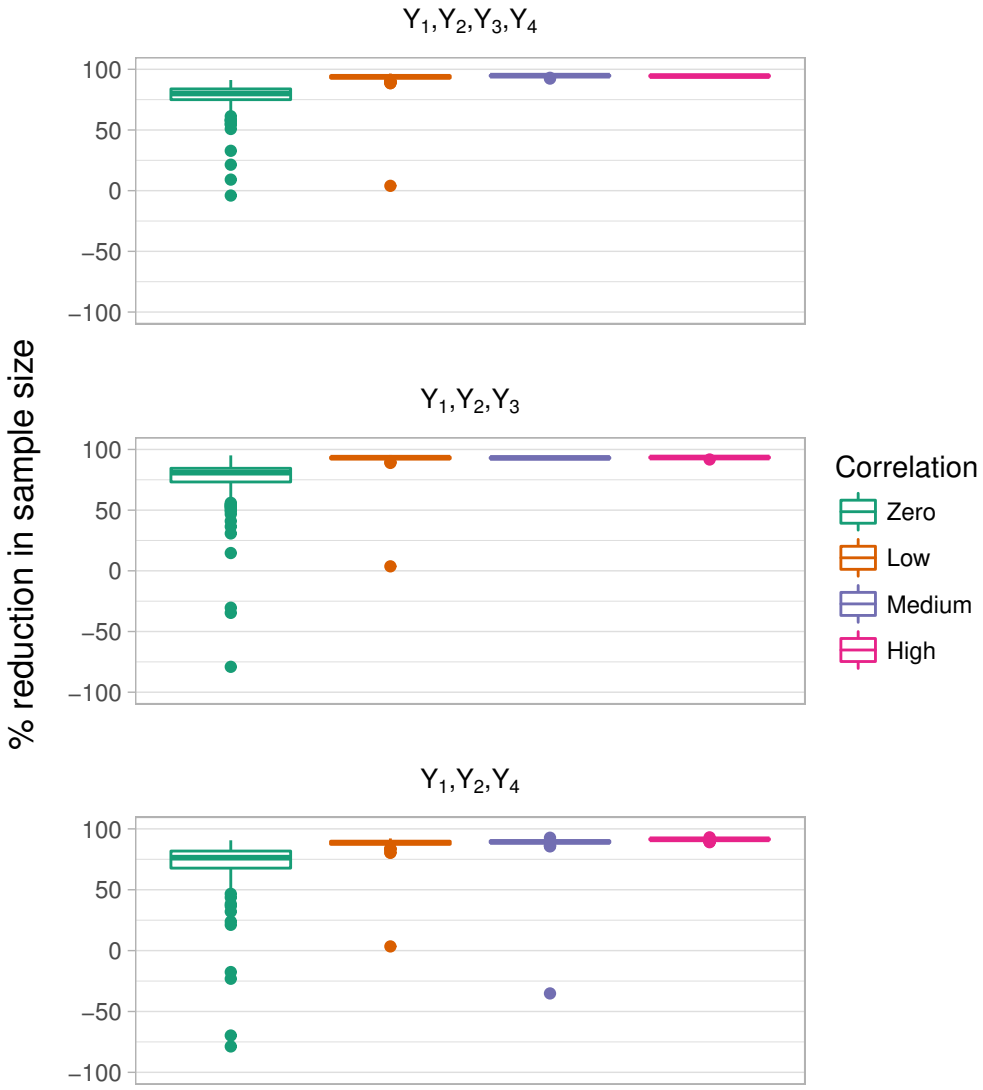


Figure 4.11: Boxplots of the estimated reduction in required sample size in 1000 simulated datasets from employing the latent variable method instead of the standard binary method for correlations of zero, low=0.3, medium=0.5 and high=0.8 between two continuous (Y_1, Y_2), one ordinal (Y_3) and one binary (Y_4) measure. Response is driven by all components in the top panel, two continuous and ordinal in the middle panel and two continuous and binary in the bottom panel

Table 4.6: Median sample sizes per group $n = n_C = n_T$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.05$, $k_m = 2, k_o = K = 1$, $\delta = \mu_T - \mu_C$: overall risk difference on the composite , $\delta^{\dagger}$: treatment effect structure in the components, for a combination of correlations 0, L=0.3, M=0.5, H=0.8 using the latent variable model when the composite is comprised of two continuous, one ordinal and one binary outcome

Response	$\delta^{\dagger}$	δ	Correlation $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$			
			000000	LLLLLL	MMMMMM	HHHHHH
Y_1, Y_2, Y_3, Y_4	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	70	51	41	81
		0.10	18	13	11	21
		0.15	8	6	5	9
	$\delta_3 = \delta_4 = 0$	0.05	62	41	47	67
		0.10	16	11	12	17
		0.15	7	5	6	8
	$\delta_1 = -\delta_2$	0.05	55	34	45	72
		0.10	14	9	12	18
		0.15	7	4	5	8
Y_1, Y_2, Y_3	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	139	71	63	105
		0.10	35	18	16	27
		0.15	16	8	7	12
	$\delta_3 = \delta_4 = 0$	0.05	120	63	75	85
		0.10	30	16	19	22
		0.15	13	7	9	10
	$\delta_1 = -\delta_2$	0.05	105	50	76	99
		0.10	27	13	19	25
		0.15	12	6	9	11
Y_1, Y_2, Y_4	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	166	106	105	112
		0.10	42	27	27	28
		0.15	19	12	12	13
	$\delta_3 = \delta_4 = 0$	0.05	147	105	113	111
		0.10	37	27	29	28
		0.15	17	12	13	13
	$\delta_1 = -\delta_2$	0.05	132	78	88	86
		0.10	33	20	22	22
		0.15	15	9	10	10

Table 4.7: Median sample sizes per group $n = n_C = n_T$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.05$, $k_m = 2, k_o = K = 1$, $\delta = \mu_T - \mu_C$: overall risk difference on the composite, $\delta^\dagger$: treatment effect structure in the components, for a combination of correlations 0, L=0.3, M=0.5, H=0.8 using the standard binary method when the composite is comprised of two continuous, one ordinal and one binary outcome

Response	$\delta^\dagger$	δ	Correlation $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$			
			000000	LLLLLL	MMMMMM	HHHHHH
Y_1, Y_2, Y_3, Y_4	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	386	739	665	1240
		0.10	97	185	167	310
		0.15	43	83	74	138
	$\delta_3 = \delta_4 = 0$	0.05	324	665	867	1205
		0.10	81	167	217	302
		0.15	36	74	97	134
	$\delta_1 = -\delta_2$	0.05	331	666	858	1169
		0.10	83	167	215	293
		0.15	37	74	96	130
Y_1, Y_2, Y_3	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	690	956	912	1300
		0.10	173	239	228	325
		0.15	77	107	102	145
	$\delta_3 = \delta_4 = 0$	0.05	650	912	1053	1283
		0.10	163	228	264	321
		0.15	73	102	117	143
	$\delta_1 = -\delta_2$	0.05	605	866	1017	1232
		0.10	152	217	255	308
		0.15	68	97	113	137
Y_1, Y_2, Y_4	$\delta_1 = \delta_2 = \delta_3 = \delta_4$	0.05	690	962	919	1298
		0.10	173	241	230	325
		0.15	77	107	103	145
	$\delta_3 = \delta_4 = 0$	0.05	642	919	1058	1281
		0.10	161	230	265	321
		0.15	72	103	118	143
	$\delta_1 = -\delta_2$	0.05	610	876	1007	1225
		0.10	153	219	252	307
		0.15	68	98	112	137

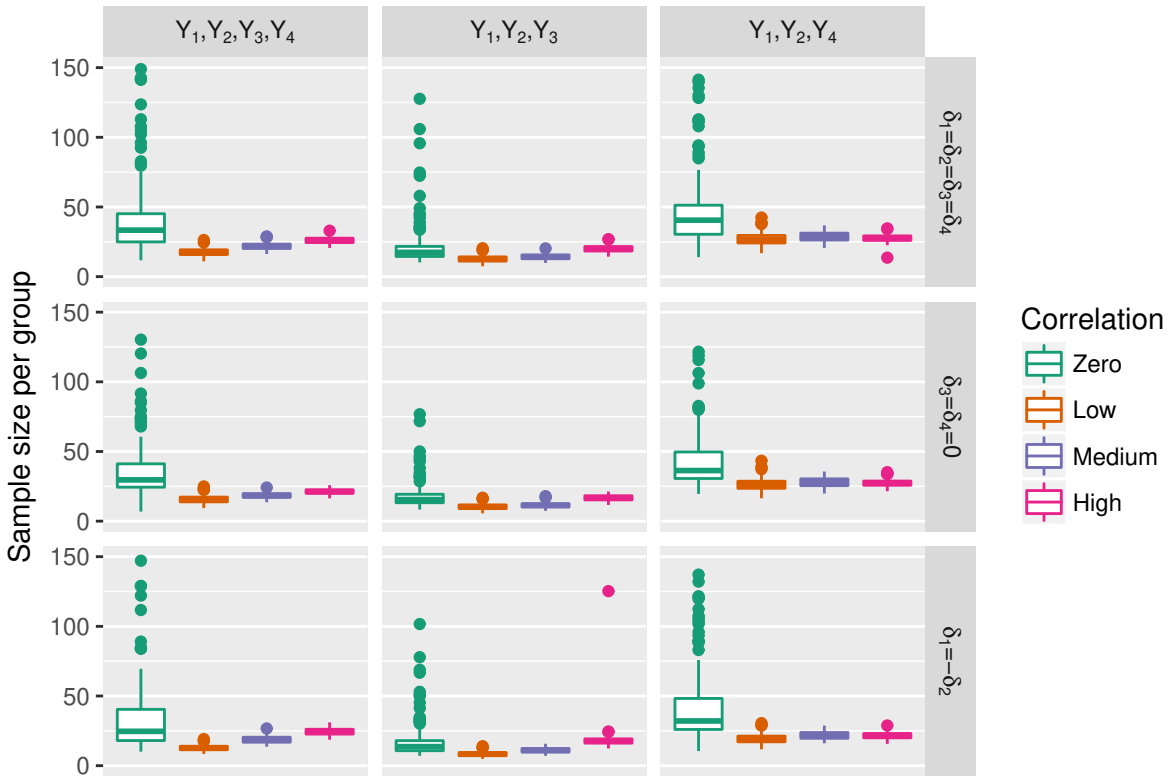


Figure 4.12: Boxplots of the estimated sample size per group from 1000 simulated datasets using the latent variable method for composites containing two continuous, one ordinal and one binary component and correlation between endpoints is zero, low=0.3, medium=0.5 or high=0.8. These are shown when response is driven by (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3) or (Y_1, Y_2, Y_4) and the treatment effect structure in the components is $\delta_1 = \delta_2 = \delta_3 = \delta_4$, $\delta_3 = \delta_4 = 0$ or $\delta_1 = -\delta_2$

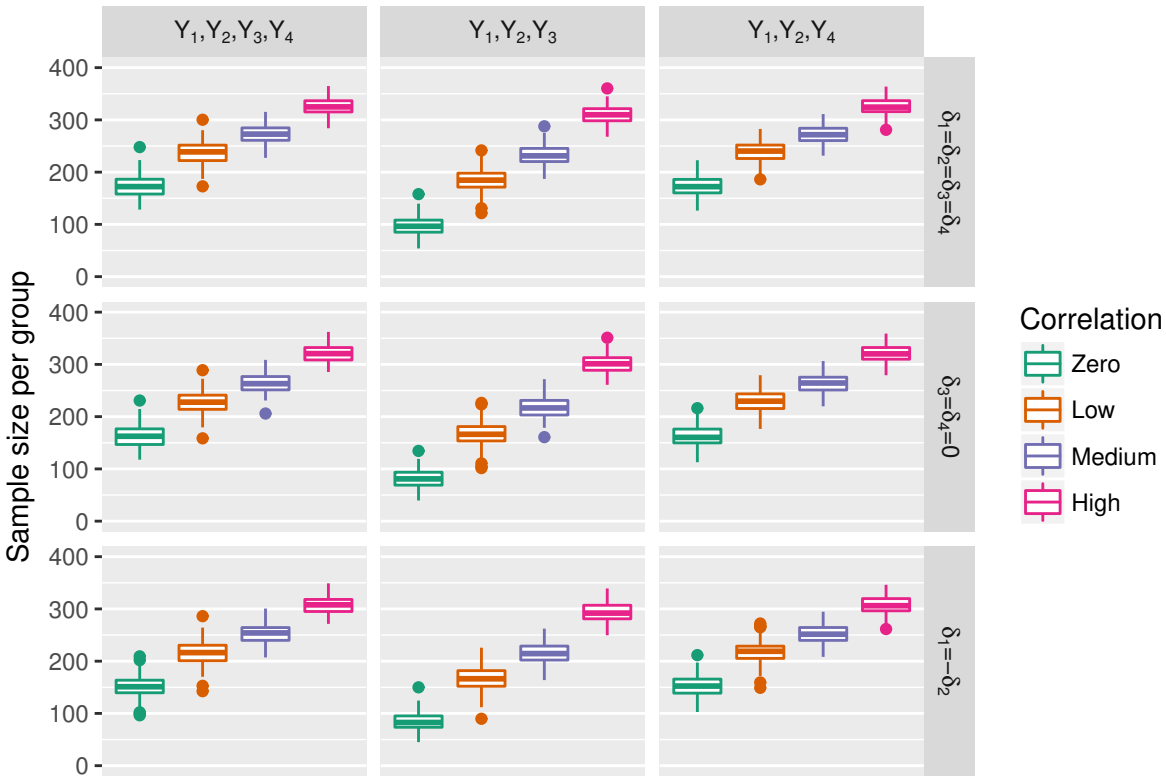


Figure 4.13: Boxplots of the estimated sample size per group from 1000 simulated datasets using the standard binary method for composites containing two continuous, one ordinal and one binary component where correlation between endpoints is zero, low=0.3, medium=0.5 or high=0.8. These are shown when response is driven by (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3) or (Y_1, Y_2, Y_4) and the treatment effect structure in the components is $\delta_1 = \delta_2 = \delta_3 = \delta_4$, $\delta_3 = \delta_4 = 0$ or $\delta_1 = -\delta_2$

Figure 4.13 is the corresponding figure for the binary method. In this instance the sample sizes estimated are the same across different treatment effect structures. The sample sizes are smallest when response is driven by Y_1, Y_2 and Y_3. As the correlation between the endpoints increases, the required sample size increases almost linearly.

4.8 Application: MUSE Trial

We illustrate the method for composite endpoint sample size determination on the MUSE trial [112]. The primary end point was the percentage of patients achieving an SRI response at week 24 with sustained reduction of oral corticosteroids (<10mg/day and less than or equal to the dose at week 1 from week 12 through 24). The study had a target sample size of 100 patients per group based on providing 88% power at the 0.10 alpha level, to detect at least 20% absolute improvement in SRI(4) response rate at week 24 for anifrolumab relative to placebo. The investigators assumed a 40% placebo response rate.

Table 4.8 shows the sample size required per group from the latent variable method, allowing for uncertainty in σ_δ. The estimated variance for the risk difference from the trial dataset is $\sigma_\delta = 0.048$ with correlation parameters $\rho_{12} = 0.448, \rho_{13} = 0.521, \rho_{14} = 0.003, \rho_{23} = 0.448, \rho_{24} = -0.031, \rho_{34} = 0.066$. For a risk difference of 0.2, the required sample size per group is 20, compared to 100 for 88% power in the standard binary method. However, the observed binary variance is lower than that assumed in the original sample size calculation so we allow for variation in the latent variable treatment effect variance. Allowing $\sigma_\delta = 0.10$ would increase the required sample size per group to 40, which is a more conservative estimate for use in practice. If the method were to be employed for increased power, rather than a decrease in required sample size, the estimated power of the latent variable method is over 99.99% for sample sizes giving 88% power at the 0.10 alpha level in the binary method. The empirical power is shown for the latent variable method in 1000 simulated datasets, which is approximately 88% for each sample size, as required.

Figure 4.14 shows the power from the latent variable method to detect a risk difference between the anifrolumab 300mg arm and the placebo arm of 0.05 to 0.20. This is shown for treatment effect variance σ_δ between 0.05 and 0.10. The latent variable method has the required 88% power to detect a risk difference of 0.125 when the variance is 0.10 and a difference of 0.08 when the variance is 0.05. This is in contrast with the

Table 4.8: Sample sizes $n = n_C = n_T$ from the latent variable method for overall power $1 - \beta \approx 88\%$, $\alpha = 0.10$, $k_2 = 2, k_o = K = 1$ to detect a response risk difference of 0.2, 0.18 and 0.16 as in the original MUSE trial sample size determination. Estimated power is shown from the latent variable method for the sample size required by the standard binary method

Risk difference	σ_δ	n.latent	Empirical power (%)	n.binary	Power (%)
0.20	0.05	20	88.05	100	99.99
0.20	0.06	24	87.01	100	99.99
0.20	0.07	28	87.62	100	99.98
0.20	0.08	32	87.04	100	99.96
0.20	0.09	36	87.83	100	99.89
0.20	0.10	40	88.12	100	99.89

binary method, which has 88% power to detect a difference of 0.20 based on the values assumed in the MUSE trial [112].

4.9 Discussion

The work in this chapter aimed to develop a method for determining the sample size required when using the latent variable model for mixed multiple outcomes. We extended work by Sozu et al. [133] to determine the sample size required in multiple co-primary continuous, ordinal and binary endpoints. We applied the method to the MUSE trial with two continuous, one ordinal and one binary outcome to demonstrate how to calculate the sample size in a future study requiring significant improvement in all outcomes. Furthermore, we provided a method for determining the sample size for mixed outcome composite endpoints, which involves using data to obtain maximum likelihood parameter estimates and their covariance matrix and using these to approximate the distribution of the risk difference under the alternative hypothesis. We found that the sample size required depended only on the overall treatment effect in the composite and not the treatment structure in the components. Sample sizes varied depending on the components driving response and the correlation between outcomes. We found that the magnitude of the increase in power offered by the latent variable method is smallest when the components are uncorrelated.

Our results show that the sample sizes required from the standard binary method increase as the correlation between the components increase. These results are unexpected as sample size typically decreases with increasing correlation. This is possibly

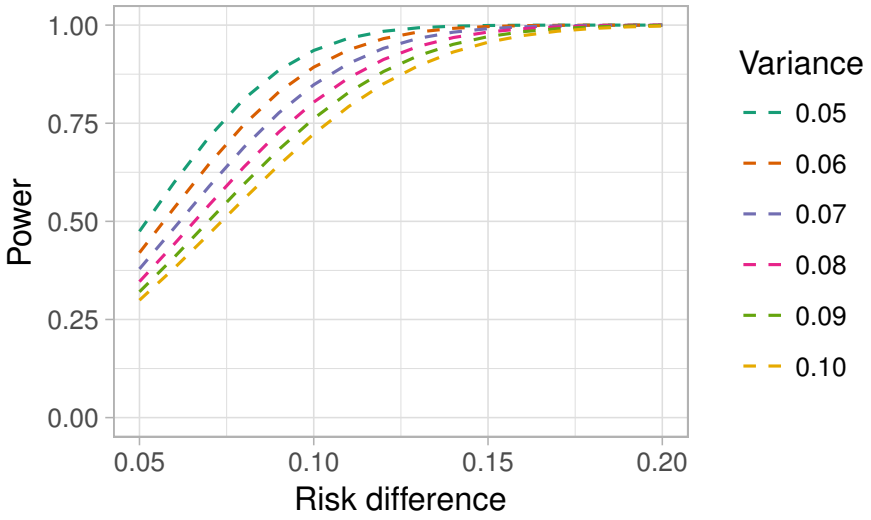


Figure 4.14: Power of the latent variable method in the MUSE trial dataset to detect a risk difference between the anifrolumab 300mg and placebo arms of 0.05 to 0.20 for σ_δ between 0.05 and 0.10 based on alpha level 0.10 and sample size 100 per group

due to the fact that as the correlation increases a patient is more likely to be a responder in both components if they are a responder in one and therefore more likely to be a responder overall. This would result in a smaller proportion of the patients being labelled as non-responders, hence requiring a larger sample size. For the latent variable method the sample size is largest for zero correlation, as we would expect. However, the sample size required is smaller for low correlation between the components than for medium and high correlation between outcomes. This ambiguity in how the correlation structure affects sample size is problematic for determining the sample size using this method in practice. One possible conservative solution is to allow for uncertainty in the correlations and use the maximum required sample size, which will still offer an improvement over the binary method.

One important result from the work in this chapter is quantifying the efficiency gain from adding a second continuous component to the composite, provided both components drive response. We found the median required sample size is reduced by 46-58% by including the additional continuous component. The results showed that the inclusion of the ordinal component with five levels is only responsible for a very small proportion of the precision gains. Given that the inclusion of the ordinal component substantially increases complexity and computational demand, it may be the case that it is sufficient to combine any ordinal components with the binary outcome. It is likely that the precision gains will be larger for ordinal variables with a larger number of

categories however this will greatly increase computation time, as discussed in Chapter 3. Ordinal outcomes with a large number of levels may be included as continuous components.

In order to determine the sample size for the co-primary endpoints, we require the parameter estimates for the latent variable method. It may be possible to determine these in multiple ways, such as from pilot trial data or by eliciting the values from experts. However, to determine the sample size using the latent variable model for composite endpoints we must fit the method to data. This is due to using the delta method to obtain the variance of the risk difference, requiring the covariance matrix of the parameter values. Basing the calculation on pilot data is potentially challenging and restrictive for a number of reasons. Firstly, it requires that a pilot or earlier phase trial must have already taken place in order to apply the method in a certain disease area. This is particularly undesirable in the case of rare diseases which would benefit most from the increased efficiency but where trials are run very infrequently. Furthermore, the pilot data could be fundamentally different to the future trial and observed effects may be imprecise. Therefore, placing too much emphasis on the existing data may lead to problems in the main trial. In theory, it is possible to elicit the required covariance parameters without data. In practice this would be difficult but allowing for uncertainty in the elicited quantities and choosing conservative values should provide an appropriate sample size estimate. An alternative when there is no data available is to apply the method using the sample size required to achieve 80% power for the binary method. Applying the latent variable method would then result in the study having a power much larger than this. We could extend this approach to use adaptive sample size re-estimation, or an internal pilot to allow for reductions in the required sample size in the trial as we collect more information about the treatment effect variability. Future work could focus on developing the method further to obtain an exact distribution for the test statistic rather that the approximation obtained using the delta method. The result would be that the covariance matrix of the parameter values and hence pilot data would not be required.

A further limitation of this work is that we have only applied the method to one trial dataset. In order to understand more about how the method works in real data we should apply this to multiple datasets, with a range of composite structures. In particular, it would be beneficial to apply the method to endpoints where all components drive response and investigate the empirical power in this case.

Chapter 5

Discussion

5.1 Summary

Composite endpoints are widely used in medical research studies for a number of reasons, discussed in Chapter 1. This thesis has proposed methods for analysing them more efficiently hence making better use of available resources. Due to the fact that composite endpoints are frequently recommended in studies of rare diseases, one of our objectives was to understand the most efficient way to model these endpoints for application in rare disease or small population clinical trials. In Chapter 2 we compared the standard binary method often used with the augmented binary method, a novel joint modelling approach shown to improve efficiency in larger samples. Given that statistical methods are most sensitive to assumptions in small samples we implemented both GLS and GEE to estimate the parameters in the longitudinal continuous model. We introduced small sample corrections and compared the corrected methods with the uncorrected methods for total sample sizes between 30 and 80 patients. We found that small sample adjustments are required to correct the type I error rate in the augmented binary method and in some scenarios in the standard binary method. We identified the small sample corrected augmented binary method using GLS as the most appropriate method, as the type I error rate is controlled and it offers substantial power gains over the standard analysis. We found that for the same statistical power the augmented binary method could reduce the required sample size in a rare disease trial using these endpoints by 32% [58].

Based on the fact that the augmented binary method performs well in the analysis of composite endpoints with one continuous and one binary outcome, we hypothesised that we could achieve even larger gains in efficiency if we employed a joint modelling

framework in composites with multiple continuous and ordinal outcomes. Motivated by a four dimensional endpoint in SLE containing two continuous, one ordinal and one binary outcome, in Chapter 3 we proposed a latent variable framework which assumes that discrete variables are manifestations of latent continuous variables. The results showed that retaining the information in multiple continuous and ordinal components greatly improved efficiency, supporting our theory that the augmented binary method could be improved upon in this setting. However, the magnitude of these gains depends on which components drive response (i.e. divide patients into responders and non-responders). We implemented the latent variable, augmented binary and standard binary methods in a phase IIb trial in patients with moderate to severe SLE and found that the treatment effect was reported 2.5 times more precisely using the latent variable model compared with logistic regression. This translates to a 60% reduction in required sample size [138]. As the simulation study showed that bias was introduced in to the treatment effect reported by the latent variable method when the joint normality assumptions were not satisfied, we introduced a bootstrap procedure to correct for this. We implemented a novel method to assess goodness of fit which showed that the latent variable method explained the data well.

As the methods development in this thesis is strongly motivated by practice, another objective of this work was to develop a method to calculate the sample size required should the latent variable model be used as a primary analysis method. In Chapter 4, we built on the work of [133] on sample size estimation for mixed continuous and binary co-primary endpoints, applying it to the SLE endpoint. We developed a method for calculating the sample size using the latent variable method for composite endpoints. This involved using the delta method to estimate the distribution of the test statistic under the alternative, which can subsequently be used for power and sample size calculation. We investigated the effect of correlation and component structure on a general endpoint containing one continuous, one ordinal and one binary outcome and through simulation found only a small variability in the sample sizes suggested by the method when the components are correlated. We quantified the effect of adding an additional continuous component driving response as reducing the sample size by an additional 46-58% when using the latent variable method. As the work in this chapter covered endpoints with differing numbers of components, different response profiles and different treatment effect structures for different correlation patterns, it is highly generalisable to clinical trials using any composite endpoint with mixed continuous, ordinal and binary components.

Overall, we were able to address the existing limitations that we had identified and are confident that this work could substantially change the practice of wasting large amounts of information in composite endpoint trials. However there exist limitations, which are highlighted below.

5.2 Limitations

Although the adjusted augmented binary method performed well in small samples, the generalisability of this work is potentially limited. Evaluating the performance in the ACR50 and ACR70 endpoints indicated the effect of differing response rates, however as the primary investigation of its behaviour was based on re-sampling from an existing trial in rheumatoid arthritis, our findings do not necessarily apply to the structures present in other diseases. We verified our findings using a simulated example however this does not ensure the applicability of the method in all rare disease trials. Another limitation in applying this in practice is that we have not focused on sample size estimation using the augmented binary method. Therefore applying the method as a secondary analysis measure in rare disease trials where it can offer additional power, rather than reduce the sample size, may be more realistic. This is still highly advantageous in small studies, however in practice there may be instances with very rare diseases that would benefit more from a reduction in required sample size. One option is to explore the small sample properties of the latent variable method, combining some components to increase events in each if necessary. Given the promising performance of the latent variable model for more common diseases, investigating its application in small samples would follow on aptly from the work in this thesis.

One problem with realising the efficiency gains from the latent variable method in practice is the computational burden of the method. To apply the latent variable method in a real trial, we would require some form of preliminary data to inform the parameter values. When the data is collected, the analysis would be performed and the outcome of interest obtained. If there are concerns about joint normality of the components, which could not be tested, then we would advise implementing the bootstrap procedure as shown in Chapter 3. This is substantially more demanding than performing a logistic regression. It is also worth noting that the methods developed in chapters 3 and 4 assume one follow-up time. Adding additional follow-up times would further increase the computational burden. Furthermore, as we have not yet developed a general package for application in composite endpoints, anyone performing

the analysis would have to edit the code for endpoints with different structures to that of the SLE endpoint.

A limitation of the work presented in Chapter 4 is the requirement for some form of existing data for sample size calculation. Even if this data exists, there may be concerns about how relevant the data is and how well it will estimate parameters for the new trial population. Although this is a serious consideration that must be taken in to account, this is a general concern for many trials using pilot data to inform parameter estimates. Alternatively all of the parameters and their covariance matrix could be elicited, although this may not be the best approach should data exist. In practice, a conservative choice of estimates could be selected based on those suggested by the data.

Finally, throughout the thesis we focused on the application of the methods to drug trials. This is not a necessary requirement and the methods could be more widely employed in studies with alternative interventions. However, in some cases this may require adaptations to the methods, for instance if problems were to arise with unbalanced data in longitudinal studies, if the data was collected continuously through wearable devices or if complex interventions were administered such as those typical in mental health trials.

5.3 Recommendations

Based on our findings we have recommendations for how this work can effect change. The methods in Chapter 2, 3 and 4 should initially be implemented retrospectively in relevant existing trials to ensure that they have been applied across a range of diseases and endpoints, and that treatment effect estimates that arise from the new method are broadly consistent with the existing. Should the analysis methods in Chapter 2 and 3 prove to be generalisable then they could be implemented as secondary analysis methods, allowing studies to avail of the increased power. One implication for this application is that the interpretations of results from each analysis method is decided prospectively. This is in order to avoid scenarios where investigators could use the method for ad-hoc justifications of efficacy, when the confidence interval for the treatment effect reported from the binary method includes the null and that from the augmented binary or latent variable method does not. Eventually, we recommend that the methods are employed as primary analysis measures in clinical trials. Table 5.1 shows the methods that we recommend in various scenarios based on the research

Table 5.1: Summary of the analysis methods recommended in a range of scenarios with different structures of composite endpoints and how to determine the sample size required in each scenario based on the research conducted to date. GLS refers to Generalised Least Squares estimation and the Firth correction is the penalised maximum likelihood method proposed by Firth

Scenario	Analysis method	Sample size determination
Rare diseases	• Aug Bin method • Continuous measure - GLS • Binary measure - Firth correction	Calculate using binary method: • Use 30% reduction *or* • Increase power
1 continuous, 1 binary	• Aug Bin method with Box-Cox *or* • Lat Var method with bootstrap	• Reduce standard by 35% • Using method in Section 4.6.5
>1 continuous, 1 binary	• Lat Var with bootstrap	• Using method in Section 4.6.5
>1 continuous, 1 or more ordinal	• Lat Var with bootstrap • If ordinal levels ≤ 5, combine with binary component • If ordinal levels > 5, retain	• Using method in Section 4.6.5

conducted to date. One restriction in applying the small sample corrected augmented binary method as a primary analysis measure is that we still do not have a method for calculating the required sample size. Consequently, for application in rare disease trials an approximation of the sample size could be used based on a conservative estimate of the reduction from the standard binary method. Although this is not ideal, it is often the case in the rare disease setting that resources are so limited that regulators are more willing to accept novel design and analysis methods than in more common diseases. Therefore it may be the case that the augmented binary method is adopted in practice quicker than the latent variable model. Overall we feel that given the potential for the methods to improve practice, every effort should be made to ensure implementation.

5.4 Future Work

We have identified key areas of possible future research to improve the scope for application, which are considered briefly below.

5.4.1 Multiple Time Points

For many trials using composite responder endpoints the investigators may be interested in response at multiple time points, as was the case for the rheumatoid arthritis endpoint. The latent variable method proposed in this thesis requires extension for application in this setting. As identified in the literature review for Chapter 3, there are different ways this could be approached. One possibility is to adopt the set-up in [85] by including a latent variable in the mean structure of the model, as shown in (5.1). In this scenario Y_{ijk} is the observed continuous measure k for patient i at time point j, where Y^*_{ijk} indicates a latent continuous measure as assumed in Chapter 3 and Chapter 4.

$$\begin{aligned} Y_{ij1} &= \boldsymbol{x}_{ij1}^T\boldsymbol{\beta_1} + \boldsymbol{z}_{ij1}^T\boldsymbol{b}_{i1} + \varepsilon_{ij1} \\ &\vdots \\ Y^*_{ijK} &= \boldsymbol{x}_{ijK}^T\boldsymbol{\beta_1} + \boldsymbol{z}_{ijK}^T\boldsymbol{b}_{iK} + \varepsilon^*_{ijK} \end{aligned} \tag{5.1}$$

where $\boldsymbol{x}_{ijk}$ and $\boldsymbol{z}_{ijk}$ are known vectors of outcome k for patient i measured at time point j and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown parameter vectors. The random effects and random errors are normally distributed as shown below.

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_{i1} \\ \vdots \\ \boldsymbol{b}_{iK} \end{pmatrix} \sim N\left(\boldsymbol{0}, \Sigma\right) = N\left(\begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1K} \\ \vdots & \ddots & \vdots \\ \Sigma_{K1} & \cdots & \Sigma_{KK} \end{bmatrix}\right) \tag{5.2}$$

$$\boldsymbol{\varepsilon}_{ij} = \begin{pmatrix} \varepsilon_{ij1} \\ \vdots \\ \varepsilon_{ijK} \end{pmatrix} \sim N\left(\boldsymbol{0}, \Sigma_e\right) = N\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1K}\sigma_1 \\ \vdots & \ddots & \vdots \\ \rho_{1K}\sigma_1 & \cdots & 1 \end{bmatrix}\right) \tag{5.3}$$

We can specify the distribution in terms of the joint distribution of observed and latent outcomes $\boldsymbol{Y}^*$ given the random effects $\boldsymbol{b}_i$ and the joint distribution of the random effects. To obtain the distribution of interest we can integrate over the random effects as demonstrated in (5.4).

$$f(Y_1, \cdots, Y_K^*; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(Y_1, \cdots, Y_K^* | \boldsymbol{b}; \boldsymbol{\theta}) f(\boldsymbol{b}; \boldsymbol{\theta}) d\boldsymbol{b} \tag{5.4}$$

A closed form solution to the likelihood can be found by exploiting properties of the skew-normal distribution [139]. This could then be used to determine the probability of response and its standard error can be obtained using the delta method, as shown

in Chapter 3.
Note that in the case of multiple time points, it is not necessary to make the assumption that the discrete outcomes are manifestations of latent outcomes and instead correlation could be accounted for using additional random effects. Exploring the possible formulation of the model for mixed composite responder endpoints measured at multiple time points is an important next step. This would ensure the efficiency gains we have found for a single time point can be translated to longitudinally measured outcomes.

5.4.2 Estimation Methods

Another aspect that could be potentially improved upon to increase the uptake of the latent variable method in the case of the SLE endpoint is estimation. The current computational time is approximately 75 minutes however now that we know that the method offers large efficiency gains, our future work could explore and compare various approaches to estimation in an attempt to speed up the process.
The estimation procedure employed in Chapter 3 is a quasi-Newton method available under the nlminb option in the optimx package in R. One advantage of this method is the high convergence rate, however it is the slowest of those available in optimx package. An alternative quasi-Newton method is the Fletcher-Powell algorithm used by Poon and Lee [83] for maximum likelihood estimates of multivariate polyserial and polychoric correlation coefficients. Another possibility would be to explore the use of GEE as an alternative to maximum likelihood estimation as considered by Catalano [91] and Regan and Catalano [94], however our results in Chapter 2 showed that this may not be the most appropriate choice when the sample size is small. Given that the EM algorithm is a natural choice in the presence of unobserved data, such as the τ-thresholds, it is an important estimation technique to consider. This approach was shown to perform well for mixed discrete and continuous outcomes by Sammel et al. [77]. Some concerns have been raised about using the EM algorithm when transforming τ from latent to observable data, where the support of $y^*|y$ depends on parameter values $\boldsymbol{\theta}$, as this may violate regularity conditions [115]. The Parameter Expanded EM algorithm was proposed to address this limitation [96]. Another alternative is the Monte Carlo ECM algorithm applied by Chib and Greenberg [81], which they compare to an MCMC algorithm for posterior estimation. Other suggestions include Gaussian quadrature, adaptive Gaussian quadrature and marginal modelling [79]. An

extensive comparison of these methods would be an important contribution to the joint modelling literature.

5.4.3 Other Outcome Types

An important extension of the work in this thesis will be to introduce methodology to model additional outcome types within the composite responder endpoint. We examine possible developments that could be implemented to accommodate nominal, count and time-to-event components.

Nominal

To model a composite containing a nominal outcome, in addition to continuous and ordinal outcomes, we could build on the work of de Leon and Carrière [140]. They define a general mixed-data model which reduces to a CGCM in the absence of nominal outcomes. Suppose that we have a composite endpoint containing one nominal binary outcome Y_1, one continuous outcome Y_2 and one ordinal outcome Y_3. The joint distribution can be factorised as shown in (5.5).

$$f(Y_1, Y_2, Y_3) = f(Y_1)f(Y_2|Y_1)f(Y_3|Y_1, Y_2) \tag{5.5}$$

As before we can introduce a threshold model for Y_3 based on Y_3^* so that,

$$f(Y_1, Y_2, Y_3^*) = f(Y_1)f(Y_2, Y_3^*|Y_1) \tag{5.6}$$

The conditional distribution of Y_2 and Y_3^* can be assumed to be multivariate normal with mean and covariance derived from conditional normality rules, as shown in Chapter 3. We can introduce an $M \times 1$ vector $\boldsymbol{x} = (X_1, ..., X_M)^T$ where X_m is either 0 or 1 depending on which of M states Y_1 is in. Then, $\boldsymbol{x}_{(m)}$ is the vector $\boldsymbol{x}$ with $X_m = 1$ and $X_{m'} = 0$ for $m' \neq m$ and $\sum_{m=1}^{M} X_m = 1$. We can then model $\boldsymbol{x}$ using a product multinomial distribution so that $f(\boldsymbol{x}; \boldsymbol{\pi}) = \prod_{m=1}^{M} \pi_m^{x_m}$ where $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)^T$ is the vector of state probabilities and x_m is the observed value of X_m. Having defined the joint distribution we could proceed as outlined previously in chapters 2 and 3 to determine the probability of response. Other approaches for including nominal components could also be explored and compared.

Count

In order to model count data within a mixed outcome composite endpoint we may have to move away from the CGCM framework and consider an alternative application of latent variables. McCulloch [79] provides an overview of some of the techniques which could be considered. One possibility is to use shared random effects, which include the same random effect in the model for each observed outcome, however as the variance is a function of the mean in a Poisson distribution this may result in problems such as connections between the overdispersion and the observed correlation. An alternative is to introduce correlated random effects. Let us assume the composite outcome of interest is made up of one count outcome Y_1, one continuous outcome Y_2 and one ordinal outcome Y_3. We can assume, as shown in (5.7), that the count variable conditioned on a random effect b_i is Poisson distributed and that the continuous variable conditioned on a random effect is normally distributed.

$$\begin{aligned} Y_{i1}|b_i &\sim indep.Poisson(\mu_{i1}) \\ log(\mu_{i1}) &= \alpha_0 + \alpha_1 x_i + b_{i1} \\ Y_{i2}|b_i &\sim indep.N(\mu_{i2}, \sigma^2) \\ \mu_{i2} &= \beta_0 + \beta_1 x_i + b_{i2} \end{aligned} \tag{5.7}$$

By assuming the ordinal variable Y_3 comes from latent Y_3^* and including a random effect b_i means we can model Y_3 as shown below.

$$\begin{aligned} Y_{i3}^*|b_i &\sim N(\mu_{i3}^*, 1) \\ \mu_{i3}^* &= \gamma_1 x_i + b_{i3} \end{aligned} \tag{5.8}$$

The correlation between the outcomes is accounted for by assuming,

$$\boldsymbol{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \end{pmatrix} \sim N(\mathbf{0}, \Sigma_b), \Sigma_b = \begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1}\sigma_{b2}\rho_{12} & \sigma_{b1}\sigma_{b3}\rho_{13} \\ \sigma_{b1}\sigma_{b2}\rho_{12} & \sigma_{b2}^2 & \sigma_{b2}\sigma_{b3}\rho_{13} \\ \sigma_{b1}\sigma_{b3}\rho_{13} & \sigma_{b2}\sigma_{b3}\rho_{13} & \sigma_{b3}^2 \end{pmatrix} \tag{5.9}$$

A binary outcome could be included in the same way or modelled directly using a logistic regression. Joint modelling the components in this way provides a flexible framework for different outcome types, however it may become infeasible with a large number of outcomes due to integrating out the random effects and then integrating over the joint distribution to obtain the probability of response. Determining the

behaviour of these models for the composite problem poses an interesting avenue for future research.

Time-to-event

The joint modelling of time-to-event and longitudinal continuous outcomes has received much consideration in the literature. Modelling these with discrete outcomes has also been studied, although to a lesser extent. Let us assume in this scenario that a composite endpoint of interest is a combination of a time-to-event outcome Y_1 and a continuous outcome Y_2 with covariate vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively. One approach is to introduce a latent trajectory function Y_2^* and assume that $\boldsymbol{X}_2$ affects Y_2 only through Y_2^* and that Y_1 and Y_2 are conditionally independent given Y_2^* [141]. Therefore, the joint distribution can be specified as shown below.

$$f(Y_1, Y_2|\boldsymbol{X}_1, \boldsymbol{X}_2) = \int_{-\infty}^{\infty} f(Y_1|\boldsymbol{X}_1, Y_1^*) f(Y_2|Y_2^*) f(Y_2^*|\boldsymbol{X}_2) dY_2^* \tag{5.10}$$

The time-to-event component $f(Y_1|\boldsymbol{X}_1, Y_1^*)$ can be modelled using a survival model and the $f(Y_2|Y_2^*)$ and $f(Y_2^*|\boldsymbol{X}_2)$ components can be analysed within the generalised linear mixed modelling framework. Other methods to account for dependency between time-to-event and longitudinal continuous or discrete data are discussed in [142]. Although there is much existing methodology for joint modelling time-to-event and other outcome data, it would be important to compare these methods to determine the most appropriate in the composite endpoint setting.

5.4.4 Further Research Directions

In addition to the research areas discussed above, there are other important future work topics to be addressed. One focus could be on assuming a different distribution for the latent variable model in Chapter 3, such as the multivariate t distribution. Modelling the endpoint using a different distribution may make the method more robust to the assumptions, therefore eliminating the need for the bootstrap procedure when applying it to real data. Another crucial consideration is to explore the methods performance under different patterns of missing data and highlight the most appropriate techniques for handling this. Finally, the development of packages to implement the methods in various software should be prioritised to ensure uptake.

5.5 Conclusion

The research undertaken in this thesis has made an important contribution to the composite endpoint and joint modelling literature. We have shown that large gains in efficiency can be achieved using the data originally collected, by modelling the composite in a way that reflects its true structure. For patients to benefit from the efficiency gains found in this work it will be crucial to focus on the dissemination of the methods.

Bibliography

[1] National Research Council, *The Prevention and Treatment of Missing Data in Clinical Trials.*, ch. Appendix A, Clinical Trials: Overview and Terminology. Washington (DC): National Academies Press (US), 2010. https://doi.org/10.17226/12955.

[2] D. G. Altman, *Practical Statistics for Medical Research.* London: Chapman and Hall, 1991.

[3] H. Moses, E. Dorsey, D. Matheson, and S. Thier, "Financial anatomy of biomedical research," *JAMA*, vol. 294, no. 11, pp. 1333–1342, 2005. doi:10.1001/jama.294.11.1333.

[4] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics*, vol. 22, no. 2, pp. 151–185, 2003. https://doi.org/10.1016/S0167-6296(02)00126-1.

[5] D. M. Dilts, "Robbing peter to pay paul: Financing clinical trial follow-up," *Journal of Clinical Oncology*, vol. 30, no. 13, pp. 1404–1405, 2012. doi:10.1200/JCO.2011.41.3120.

[6] R. Collier, "Rapidly rising clinical trial costs worry researchers," *CMAJ*, vol. 180, pp. 277–8, 2009. doi:10.1503/cmaj.082041.

[7] J. P. A. Ioannidis, S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz, and R. Tibshirani, "Increasing value and reducing waste in research design, conduct, and analysis," *The Lancet*, vol. 383, no. 9912, pp. 166–175, 2014.

[8] S. Ross, "Composite outcomes in randomized clinical trials: arguments for and against," *Am J Obstet gynecol*, vol. 196, no. 2, pp. 199e1–6, 2007. doi:10.1016/j.ajog.2006.10.903.

[9] N. Freemantle, M. Calvert, J. Wood, J. Eastaugh, and C. Griffin, "Composite outcomes in randomized trials: greater precision but with greater uncertainty?," *JAMA*, vol. 289, pp. 2554–2559, 2003. doi:10.1001/jama.289.19.2554.

[10] V. Montori, G. Permanyer-Miralda, I. Ferreira-Gonzalez, J. Busse, V. Pacheco-Huergo, and D. B. et al., "Validity of composite endpoints in clinical trials," *BMJ*, vol. 330, pp. 594–596, 2005. doi:10.1136/bmj.330.7491.594.

[11] *Multiple Analyses in Clinical Trials*, ch. 7: Introduction to composite endpoints. Statistics for Biology and Health, Springer, 2003. doi:10.1007/b97513.

[12] J.-M. Ahn, J.-H. Roh, Y.-H. Kim, D.-W. Park, S.-C. Yun, P. H. Lee, and et al., "Randomized trial of stents versus bypass surgery for left main coronary artery disease: 5-year outcomes of the precombat study," *Journal of the American College of Cardiology*, vol. 65, no. 20, pp. 2198 – 2206, 2015. https://doi.org/10.1016/j.jacc.2015.03.033.

[13] I. Ferreira-González, G. Permanyer-Miralda, J. W. Busse, D. M. Bryant, V. M. Montori, P. Alonso-Coello, and et al., "Methodological discussions for using and interpreting composite endpoints are limited, but still identify major concerns," *Journal of Clinical Epidemiology*, vol. 60, no. 7, pp. 651–657, 2007. doi:10.1016/j.jclinepi.2006.10.020.

[14] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, and et al., "New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1)," *European Journal of Cancer*, vol. 45, pp. 228–247, 2009. doi:10.1016/j.ejca.2008.10.026.

[15] I. Ferreira-Gonzalez, J. Busse, D. Heels-Ansdell, V. Montori, E. Akl, D. Bryant, and et al., "Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials," *BMJ*, vol. 334, no. 7597, p. 786, 2007. doi:10.1136/bmj.39136.682083.AE.

[16] N. Freemantle and M. Calvert, "Weighing the pros and cons for composite outcomes in clinical trials," *J Clin Epidemiol*, vol. 60, pp. 658–659, 2007. doi:10.1016/j.jclinepi.2006.10.024.

[17] G. Tomlinson and A. Detsky, "Composite endpoints in randomized trials: There is no free lunch," *JAMA*, vol. 303, no. 3, pp. 267–268, 2010. doi:10.1001/jama.2009.2017.

[18] J. A. Lewis, "Statistical principles for clinical trials (ich e9): an introductory note on an international guideline," *Stat Med*, vol. 18, pp. 1903–1942, 1999.

[19] C. P. Cannon, "Clinical perspectives on the use of composite endpoints," *Control Clin Trials*, vol. 18, pp. 517–529, 1997. https://doi.org/10.1016/S0197-2456(97)00005-6.

[20] G. F. Gensini and A. A. Conti, "The evaluation of the results of clinical trials: surrogate end points and composite endpoints," *Minerva Med*, vol. 95, pp. 71–75, 2004.

[21] A. V. Carneiro, "Composite outcomes in clinical trials: uses and problems," *Rev Port Cardiol*, vol. 22, pp. 1253–1263, 2003.

[22] M. Lauer, "Clinical trials - multiple treatments, multiple endpoints and multiple lessons," *JAMA*, vol. 289, pp. 2575–2577, 2003. doi:10.1001/jama.289.19.2575.

[23] N. Freemantle and M. Calvert, "Interpreting composite outcomes in trials. editorial," *BMJ*, vol. 341, p. c3529, 2010. https://doi.org/10.1136/bmj.c3529.

[24] G. Y. Chi, "Some issues with composite endpoints in clinical trials," *Fundam Clin Pharmacol*, vol. 19, pp. 609–619, 2005. doi:10.1111/j.1472-8206.2005.00370.x.

[25] J. Lubsen and B. Kirwan, "Combined endpoints: can we use them?," *Stat Med*, vol. 21, pp. 2959–2970, 2002. doi:10.1002/sim.1300.

[26] L. Wittkop, C. Smith, Z. Fox, C. Sabin, L. Richert, J. Aboulker, and et al., "Methodological issues in the use of composite endpoints in clinical trials: examples from the hiv field.," *Clin trials*, vol. 7, no. 1, pp. 19–35, 2010. doi:10.1177/1740774509356117.

[27] M. Gent, "Some issues in the construction and use of clusters of outcome events," *Control Clin Trials*, vol. 18, pp. 546–549, 1999. https://doi.org/10.1016/S0197-2456(97)82056-9.

[28] S. J. Pocock, "Clinical trials with multiple outcomes: a statistical perspective on their design, analysis and interpretation," *Control Clin Trials*, vol. 18, pp. 530–545, 1997.

[29] P. M. Rothwell, "External validity of randomised controlled trials: "to whom do the results of this trial apply?"," *Lancet*, vol. 365, pp. 82–93, 2005. doi:10.1016/S0140-6736(04)17670-8.

[30] G. Cordoba, L. Schwartz, S. Woloshin, H. Bae, and P. Gøtzsche, "Definition, reporting, and interpretation of composite outcomes in clinical trials - systematic review," *BMJ*, vol. 341, p. 3920, 2010. https://doi.org/10.1136/bmj.c3920.

[31] FDA, "Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance," *BMC Health and Quality of Life Outcomes*, vol. 4, p. 79, 2006. doi:10.1186/1477-7525-4-79.

[32] EUnetHTA, "Endpoints used for relative effectiveness assessment of pharmaceuticals - composite endpoints," health technology assessment joint action 1, work package 5, European network for Health Technology Assessment, 2013.

[33] S. Senn, "Disappointing dichotomies," *Pharmaceutical Statistics*, vol. 2, pp. 239–240, 2003. https://doi.org/10.1002/pst.90.

[34] S. Suissa, "Binary methods for continuous outcomes: a parametric alternative," *Journal of Clinical Epidemiology*, vol. 44, no. 3, pp. 241–248, 1991.

[35] J. Wason and S. R. Seaman, "Using continuous data on tumour measurements to improve inference in phase ii cancer studies," *Statistics in Medicine*, vol. 32, no. 26, pp. 4639–4650, 2013. doi:10.1002/sim.5867.

[36] J. Wason and M. Jenkins, "Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method," *Rheumatology*, vol. 55, no. 10, pp. 1796–1802, 2016. doi: 10.1093/rheumatology/kew263.

[37] C.-J. Lin and J. M. Wason, "Improving phase ii oncology trials using best observed recist response as an endpoint by modelling continuous tumour measurements," *Statistics in Medicine*, vol. 36, no. 29, pp. 4616–4626, 2017. doi: 10.1002/sim.7453.

[38] J. Wason, M. McMenamin, and S. Dodd, "Analysis of responder-based endpoints: improving power through utilising continuous components." PREPRINT (Version 1) available at Research Square +https://doi.org/10.21203/rs.2.11783/v1+, 22 July 2019.

[39] R. Griggs, M. Batshaw, M. Dunkle, R. Gopal-Srivastava, E. Kaye, and J. Krischer, "Clinical research for rare disease: Opportunities, challenges and solutions," *Molecular Genetics and Metabolism*, vol. 91, no. 1, pp. 20–26, 2009. doi: 10.1016/j.ymgme.2008.10.003.

[40] R. Joppi, V. Bertele, and S. Garattini, "Orphan drug development is progressing too slowly," *British Journal of Clinical Pharamacology*, vol. 61, pp. 355–360, 2006. doi: 10.1111/j.1365-2125.2006.02579.x.

[41] R. Hilgers, K. Roes, and N. Stallard, "Directions for new developments on statistical design and analysis of small population group trials," *Orphanet Journal of Rare Diseases*, vol. 11, no. 1, p. 78, 2016. doi: 10.1186/s13023-016-0464-5.

[42] M. Dunkle, W. Pines, and P. Saltonstall, "Advocacy groups and their role in rare diseases research," *Advances in experimental medicine and biology*, vol. 686, pp. 514–525, 2010. doi: 10.1007/978-90-481-9485-828.

[43] A. Hall and M. Carlson, "The current status of orphan drug development in europe and the us," *Intractable Rare Disease Research*, vol. 3, no. 1, pp. 1–7, 2014. doi: 10.5582/irdr.3.1.

[44] Z. Saleh and T. Arayssi, "Update on the therapy of behçet disease," *Therapeutic Advances in Chronic Disease*, vol. 5, no. 3, pp. 112–134, 2014. doi: 10.1177/2040622314523062.

[45] A. Albert and J. Anderson, "On the existence of maximum likelihood methods in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984.

[46] M. Webb, J. Wilson, and J. Chong, "An analysis of quasi-complete binary data with logistic models: Applications to alcohol abuse data," *Journal of Data Science*, vol. 2, pp. 273–285, 2004. doi:10.6339/JDS.2004.02(3).155.

[47] D. Firth, "Bias reduction of maximum likelihood estimation," *Biometrika*, vol. 80, pp. 27–38, 1993.

[48] S. Cole, H. Chu, and S. Greenland, "Maximum likelihood, profile likelihood, and penalized likelihood: A primer," *Am J Epidemiol*, vol. 179, no. 2, pp. 252–260, 2014.

[49] C. Rainey and K. McCaskey, "Estimating logit models with small samples," *Political Science Research and Methods*, In Press.

[50] *brglm: Bias reduction in binomial-response Generalized Linear Models.* http://www.ucl.ac.uk/ ucakiko/software.html, 2013.

[51] J. C. Gardiner, L. Zhehui, and L. A. Roman, "Fixed effects, random effects and gee: What are the differences?," *Statistics in Medicine*, vol. 28, pp. 221–239, 2009. https://doi.org/10.1002/sim.3478.

[52] M. Wang, L. Kong, Z. Li, and L. Zhang, "Covariance estimators for generalized estimating equations (gee) in longitudinal analysis with small samples," *Statistics in Medicine*, vol. 35, no. 10, pp. 1706–1721, 2015. doi: 10.1002/sim.6817.

[53] J. Morel, M. Bokossa, and N. Neerchal, "Small sample correction for the variance of gee estimators," *Biometrical Journal*, vol. 45, pp. 395–409, 2003. https://doi.org/10.1002/bimj.200390021.

[54] J. Rao and A. Scott, "The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 221–230, 1981. doi: 10.2307/2287815.

[55] M. Wang, *geesmv: Modified Variance Estimators for Generalized Estimating Equations*. https://CRAN.R-project.org/package=geesmv, r package version 1.3 ed., 2015.

[56] M. Weinblatt, M. Genovese, M. Ho, and et al., "Effects of fostamatinib, an oral spleen tyrosine kinase inhibitor, in rheumatoid arthritis patients with an inadequate response to methotrexate: results from a phase iii, multicenter, randomized, double-blind, placebo-controlled, parallel-group study," *Arthritis Rheumatology*, vol. 66, pp. 3255–3264, 2014.

[57] G. Imbens and M. Kolesar, "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, vol. 98, no. 4, pp. 701–712, 2016.

[58] M. McMenamin, A. Berglind, and J. Wason, "Improving the analysis of composite endpoints in rare disease trials," *Orphanet Journal of Rare Diseases*, vol. 13, p. 81, 2018. https://doi.org/10.1186/s13023-018-0819-1.

[59] M. Parmar, M. Sydes, and T. Morris, "How do you design randomised trials for smaller populations? a framework," *BMC Medicine*, vol. 14, p. 183, 2016.

[60] P. Peduzzi, J. Concato, E. Kemper, T. Holford, and A. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, pp. 1373–1379, 1996.

[61] L. Lisnevskaia, G. Murphy, and D. Isenberg, "Systemic lupus erythematosus," *The Lancet*, vol. 384, no. 9957, pp. 1878 – 1888, 2014. https://doi.org/10.1016/S0140-6736(14)60128-8.

[62] D. P. D'Cruz, M. A. Khamashta, and G. R. Hughes, "Systemic lupus erythematosus," *The Lancet*, vol. 369, no. 9561, pp. 587 – 596, 2007. https://doi.org/10.1016/S0140-6736(07)60279-7.

[63] K. Corapi, M. Dooely, and W. Pendergraft, "Comparison and evaluation of lupus nephritis response criteria in lupus activity indices and clinical trials.," *Arthritis Research and Therapy*, vol. 17, no. 1, p. 110, 2015. doi: 10.1186/s13075-015-0621-6.

[64] K. Luijten, J. Tekstra, J. Bijlsma, and M. Bijl, "The systemic lupus erythematosus responder index (sri); a new sle disease activity assessment," *Autoimmunity reviews*, vol. 11, no. 5, pp. 326–329, 2012. doi: 10.1016/j.autrev.2011.06.011.

[65] D. Gladman, D. Ibañez, and M. Urowitz, "Systemic lupus erythematosus disease activity index 2000," *The Journal of rheumatology*, vol. 29, no. 2, pp. 288–291, 2002.

[66] R. Nelsen, *An Introduction to Copulas: Definitions and Basic Properties*. New York, NY: Springer New York, 1999. doi: 10.1007/978-1-4757-3076-0.

[67] A. de Leon and K. Carriere, eds., *Analysis of Mixed Data Methods and Applications*. Chapman and Hall/CRC, 2013.

[68] G. Verbeke, S. Fieuws, and G. Molenberghs, "The analysis of multivariate longitudinal data: A review," *Statistical Methods in Medical Research*, vol. 23, no. 1, pp. 42–59, 2014. doi: 10.1177/0962280212445834.

[69] A. de Leon and B. Wu, "Copula based regression models for a bivariate mixed discrete and continuous outcome," *Statistics in Medicine*, vol. 30, pp. 175–185, 2010. https://doi.org/10.1002/sim.4087.

[70] B. Wu and de Leon. A.R., "Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology," *JABES*, vol. 19, no. 1, pp. 39–56, 2014. https://doi.org/10.1007/s13253-013-0155-9.

[71] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. Wiley and Sons, 1990. https://doi.org/10.1007/BF02618478.

[72] S. Lauritzen and N. Wermuth, "Graphical models for association between variables, some of which are qualitative and some quantitative," *Annals of Statistics*, vol. 17, pp. 31–54, 1989. doi:10.1214/aos/1176347003.

[73] I. Olkin and R. Tate, "Multivariate correction models with mixed discrete and continuous variables," *Annals of Mathematical Statistics*, vol. 32, pp. 448–465, 1961. doi:10.1214/aoms/1177705052.

[74] G. Fitzmaurice and N. Laird, "Regression models for a bivariate discrete and continuous outcome with clustering," *Journal of the American Statistical Association*, vol. 90, pp. 845–852, 1995. doi: 10.2307/2291318.

[75] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC, 2004.

[76] M. Sammel and L. Ryan, "Effects of covariance misspecification in a latent variable model for multiple outcomes," *Statistica Sinica*, vol. 12, pp. 1207–1222, 2002.

[77] M. Sammel, L. Ryan, and J. Leger, "Latent variable models for mixed discrete and continuous outcomes," *J. R. Statist. Soc. B*, vol. 59, no. 3, pp. 667–678, 1997. https://doi.org/10.1111/1467-9868.00090.

[78] D. Dunson, "Bayesian latent variable models for clustered mixed outcomes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 2, pp. 355–366, 2000. doi: 10.1111/1467-9868.00236.

[79] C. McCulloch, "Joint modelling of mixed outcome types using latent variables," *Statistical Methods in Medical Research*, vol. 17, pp. 53–73, 2008. doi: 10.1177/0962280207081240.

[80] J. Ashford and R. Sowden, "Multivariate probit analysis," *Biometrics*, vol. 26, pp. 535–46, 1970. doi: 10.2307/2529107.

[81] S. Chib and E. Greenberg, "Analysis of multivariate probit models," *Biometrika*, vol. 85, no. 2, pp. 347–361, 1998. https://doi.org/10.1093/biomet/85.2.347.

[82] K. Pearson, "Mathematical contributions to the theory of evolution. xii. on a generalised theory of alternative inheritance, with special reference to mendel's laws," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 203, pp. 53–86, 1904.

[83] W. Poon and S. Lee, "Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients," *Psychometrika*, vol. 52, no. 3, pp. 409–430, 1987. https://doi.org/10.1007/BF02294364.

[84] A. De Leon, "Pairwise likelihood approach to grouped continuous model and its extension," *Statistics & probability letters*, vol. 75, no. 1, pp. 49–57, 2005. https://doi.org/10.1016/j.spl.2005.05.017.

[85] R. Gueorguieva and A. Agresti, "A correlated probit model for joint modeling of clustered binary and continuous responses," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1102–1112, 2001. https://doi.org/10.1198/016214501753208762.

[86] W. J. Boscardin, X. Zhang, and T. R. Belin, "Modeling a mixture of ordinal and continuous repeated measures," *Journal of Statistical Computation and Simulation*, vol. 78, no. 10, pp. 873–886, 2008. https://doi.org/10.1080/00949650701480259.

[87] A. R. De Leon and K. C. Carriègre, "General mixed-data model: Extension of general location and grouped continuous models," *Canadian Journal of Statistics*, vol. 35, no. 4, pp. 533–548, 2007.

[88] R. Tate, "The theory of correlation between two continuous variables when one is dichotomised," *Biometrika*, vol. 42, no. 1-2, pp. 205–216, 1955. doi: 10.2307/2333437.

[89] D. Cox and N. Wermuth, "Response models for mixed binary and quantitative variables," *Biometrika*, vol. 79, no. 3, pp. 441–61, 1992. https://doi.org/10.1093/biomet/79.3.441.

[90] P. Catalano and L. Ryan, "Bivariate latent variable models for clustered discrete and continuous outcomes," *Journal of the American Statistical Association*, vol. 87, no. 419, pp. 651–658, 1992. doi: 10.2307/2290200.

[91] P. Catalano, "Bivariate modelling of clustered continuous and ordered categorical outcomes," *Statistics in Medicine*, vol. 16, pp. 883–900, 1997.

[92] E. Samani and M. Ganjali, "A multivariate latent variable model for mixed continuous and ordinal responses," *World Applied Sciences Journal*, vol. 3, no. 2, pp. 294–299, 2008.

[93] G. Arminger and U. Kusters, *Latent trait and latent class models*, ch. Latent trait models with indicators of mixed measurement level, pp. 51–73. Plenum, 1988. https://doi.org/10.1007/978-1-4757-5644-9.

[94] M. Regan and P. Catalano, "Regression models and risk estimation for mixed discrete and continuous outcomes in developmental toxicology," *Risk Analysis*, vol. 20, pp. 363–376, 2000. doi: 10.1111/0272-4332.203035.

[95] C. Faes, H. Geys, M. Aerts, P. Catalano, and G. Molenberghs, "Modelling combined continuous and ordinal outcomes from developmental toxicity studies," in *In Proceedings of the 17th International Workshop on Statistical Modelling* (M. Stasinopoulos and G. Touloumi, eds.), (Chania, Crete), 2002. https://doi.org/10.1198/108571104X16349.

[96] H. Zhang, D. Liu, J. Zhao, and X. Bi, "Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine co-dependence," *The Annals of Applied Statistics*, vol. 12, no. 4, pp. 2359–2378, 2018. doi:10.1214/18-AOAS1156.

[97] T. Sozu, T. Sugimoto, and T. Hamisaki, "Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables," *Biometrical Journal*, vol. 54, no. 5, pp. 716–729, 2012. doi: 10.1002/bimj.201100221.

[98] T. Sozu, T. Sugimoto, and T. Hamasaki, "Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints," *J Biopharm Stat*, vol. 21, pp. 650–668, 2011. https://doi.org/10.1002/sim.3972.

[99] E. Lessafre and G. Molenberghs, "Multivariate probit analysis: A neglected procedure in medical statistics," *Statistics in Medicine*, vol. 10, no. 9, pp. 1391–1403, 1991. https://doi.org/10.1002/sim.4780100907.

[100] R. Gueorguieva and G. Sanacora, "A latent variable model for joint analysis of repeatedly measured ordinal and continuous outcomes," in *In Proceedings of the 18th International Workshop on Statistical Modelling* (G. Verbeke, G. Molenberghs, M. Aerts, and S. Fiews, eds.), (Katholieke Universiteit Leuven: Leuven), pp. 171–176, 2003.

[101] R. Gueorguieva and G. Sanacora, "Joint analysis of repeatedly observed continuous and ordinal measures of disease severity," *Statistics in Medicine*, vol. 25, no. 8, pp. 1307–1322, 2006. doi: 10.1002/sim.2270.

[102] A. Genz, "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, vol. 1, pp. 141–150, 1992. doi: 10.2307/1390838.

[103] J. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965. https://doi.org/10.1093/comjnl/7.4.308.

[104] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, 1970. https://doi.org/10.1093/comjnl/13.3.317.

[105] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal of Scientific Computing*, vol. 16, pp. 1190–1208, 9 1995. https://doi.org/10.1137/0916069.

[106] L. F. Richardson, "The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam," *Philosophical Transactions of the Royal Society A*, vol. 210, pp. 307–357, 1911. doi:10.1098/rsta.1911.0009.

[107] N. J. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002. https://doi.org/10.1093/imanum/22.3.329.

[108] A. Teixeira-Pinto and S.-L. Normand, "Correlated bivariate continuous and binary outcomes: Issues and applications," *Statistics in medicine*, vol. 28, no. 13, pp. 1753–1773, 2009. doi: 10.1002/sim.3588.

[109] M. Gilli and E. Schumann, "A note on 'good starting values' in numerical optimisation," *Numerical Optimisation*, June 2010. http://dx.doi.org/10.2139/ssrn.1620083.

[110] T. Morris, I. White, and M. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, 2019. https://doi.org/10.1002/sim.8086.

[111] A. Azzalini and A. Dalla Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, pp. 715–726, 1996.

[112] R. Furie, M. Khamashta, J. Merrill, V. Werth, K. Kalunian, P. Brohawn, and et al., "Anifrolumab, an anti interferon alpha receptor monoclonal antibody, in moderate-to-severe systemic lupus erythematosus," *Arthritis and Rheumatology*, vol. 69, no. 2, pp. 376–386, 2017. doi: 10.1002/art.39962.

[113] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. doi: 10.2307/2958830.

[114] K. C. Kalunian, M. B. Urowitz, D. Isenberg, J. T. Merrill, M. Petri, R. A. Furie, and et al., "Clinical trial parameters that influence outcomes in lupus trials that use the systemic lupus erythematosus responder index," *Rheumatology*, vol. 57, no. 1, pp. 125–133, 2018. doi: 10.1093/rheumatology/kex368.

[115] P. A. Ruud, "Extensions of estimation methods using the em algorithm," *Journal of Econometrics*, vol. 49, pp. 305–341, 1991. https://doi.org/10.1016/0304-4076(91)90001-T.

[116] T. Sozu, T. Sugimoto, T. Hamasaki, and S. Evans, *Sample Size Determination in Clinical Trials with Multiple Endpoints*. Springer, 2015.

[117] C. Xiong, K. Yu, F. Gao, Y. Yan, and Z. Zhang, "Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to alzheimer's treatment trial," *Clin Trials*, vol. 2, pp. 387–393, 2005. doi: 10.1191/1740774505cn112oa.

[118] T. Sugimoto, T. Sozu, and T. Hamasaki, "A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints," *Pharm Stat*, vol. 11, pp. 118–128, 2012a. doi: 10.1002/pst.505.

[119] M. Eaton and R. Muirhead, "On a multiple endpoints testing problem," *J Stat Plan Infer*, vol. 137, pp. 3416–3429, 2007. https://doi.org/10.1016/j.jspi.2007.03.021.

[120] S. Julious and N. McIntyre, "Sample sizes for trials involving multiple correlated must-win comparisons," *Pharm Stat*, vol. 11, pp. 177–185, 2012. doi: 10.1002/pst.515.

[121] C. Chuang-Stein, P. Stryszak, A. Dmitrienko, and W. Offen, "Challenge of multiple co-primary endpoints: a new approach.," *Stat Med*, vol. 26, pp. 1181–1192, 2007. https://doi.org/10.1002/sim.2604.

[122] G. Kordzakhia, O. Siddiqui, and M. Huque, "Method of balanced adjustment in testing co-primary endpoints," *Stat Med*, vol. 29, pp. 2055–2066, 2010. https://doi.org/10.1002/sim.3950.

[123] H. Hung and S. Wang, "Some controversial multiple testing problems in regulatory applications," *J Biopharm Stat*, vol. 19, pp. 1–11, 2009. doi: 10.1080/10543400802541693.

[124] A. Dmitrienko, A. Tamhane, and F. Bretz, *Multiple testing problems in pharmaceutical statistics*. Boca Raton: Chapman Hall/CRC, 2010.

[125] J. Gong, J. Pinheiro, and D. DeMets, "Estimating significance level and power comparisons for testing multiple endpoints in clinical trials," *Control Clin Trials*, vol. 21, pp. 323–329, 2000. https://doi.org/10.1016/S0197-2456(00)00049-0.

[126] T. Sozu, T. Sugimoto, and T. Hamasaki, "Sample size determination in clinical trials with multiple co-primary binary endpoints," *Stat Med*, vol. 29, pp. 2169–2179, 2010. https://doi.org/10.1002/sim.3972.

[127] T. Hamasaki, S. Evans, T. Sugimoto, and T. Sozu, "Power and sample size determination for clinical trials with two correlated binary relative risks," tech. rep., In: ENAR Spring Meeting, Washington DC, USA, 2012.

[128] J. Song, "Sample size for simulataneous testing of rate differences in non-inferiority trials with multiple endpoints," *Comput Stat Data Anal*, vol. 53, pp. 1201–1207, 2009. https://doi.org/10.1016/j.csda.2008.10.028.

[129] T. Hamasaki, T. Sugimoto, S. Evans, and T. Sozu, "Sample size determination for clinical trials with co-primary outcomes: exponential event-times," *Pharma Stat*, vol. 12, pp. 28–34, 2013. doi: 10.1002/pst.1545.

[130] T. Sugimoto, T. Hamasaki, and T. Sozu, "Sample size determination in clinical trials with two correlated co-primary time-to-event endpoints," in *In: The 7th international conference on multiple comparison procedures*, (Washington DC, USA), 29 Aug - 1 Sept 2011.

[131] T. Sugimoto, T. Sozu, T. Hamasaki, and S. Evans, "A logrank test-based method for sizing clinical trials with two co-primary time-to-events endpoints," *Biostatistics*, vol. 14, pp. 409–421, 2013. doi: 10.1093/biostatistics/kxs057.

[132] T. Sugimoto, T. Hamasaki, T. Sozu, and S. Evans, "Sample size determination in clinical trials with two correlated time-to-event endpoints as primary contrast," in *In: The 6th FDA-DIA forum*, (Washington DC, USA), 22-25 April 2012b.

[133] T. Sozu, T. Sugimoto, and T. Hamasaki, "Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables," *Biometrical J*, vol. 54, pp. 716–729, 2012. doi: 10.1002/bimj.201100221.

[134] B. Wu and A. de Leon, "Letter to the editor re: "sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables", by t sozu, t. sugimoto, t. hamasaki," *Biometrical Journal*, pp. 807–802, 2013. doi: 10.1002/bimj.201200254.

[135] R. Dorfman, "A note on the delta-method for finding variance formulae," *The Biometric Bulletin*, vol. 1, pp. 129–137, 1938.

[136] J. Hislop, T. E. Adewuyi, L. D. Vale, K. Harrild, C. Fraser, T. Gurung, and et al., "Methods for specifying the target difference in a randomised controlled trial: The difference elicitation in trials (delta) systematic review," *PLOS Medicine*, vol. 11, pp. 1–16, 05 2014. doi: 10.1371/journal.pmed.1001645.

[137] J. A. Cook, S. A. Julious, W. Sones, L. V. Hampson, C. Hewitt, J. A. Berlin, D. Ashby, R. Emsley, D. A. Fergusson, S. J. Walters, E. C. F. Wilson, G. MacLennan, N. Stallard, J. C. Rothwell, M. Bland, L. Brown, C. R. Ramsay, A. Cook, D. Armstrong, D. Altman, and L. D. Vale, "Delta2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial," *BMJ*, vol. 363, 2018.

[138] M. McMenamin, J. K. Barrett, A. Berglind, and J. M. Wason, "Employing latent variable models to improve efficiency in composite endpoint analysis." arXiv:1902.07037, February 2019.

[139] B. Arnold, "Flexible univariate and multivariate models based on hidden truncation," *Journal of Statistical Planning and Inference*, vol. 139, pp. 3741–3749, 2009. https://doi.org/10.1016/j.jspi.2009.05.013.

[140] A. R. de Leon and K. Carrière, "General mixed-data model: extension of general location and grouped continuous models," *The Canadian Journal of Statistics*, vol. 35, no. 4, pp. 533–548, 2007. https://doi.org/10.1002/cjs.5550350405.

[141] Y.-Y. Chi and J. G. Ibrahim, "Joint models for multivariate longitudinal and multivariate survival data," *Biometrics*, vol. 62, pp. 432–445, 2006. doi: 10.1111/j.1541-0420.2005.00448.x.

[142] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kollamonage-Dona, "Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues," *BMC Medical Research Methodology*, vol. 16, no. 1, p. 117, 2016. doi: 10.1186/s12874-016-0212-5.

Appendix A

Preprint: Scope for Application

Analysis of responder-based endpoints: improving power through utilising continuous components

James Wason[1,2*], Martina McMenamin[2], Susanna Dodd[3]

[1] Institute of Health and Society, Newcastle University, Baddiley-Clark Building Newcastle upon Tyne NE2 4BN, United Kingdom. Email: james.wason@newcastle.ac.uk

[2] MRC Biostatistics Unit, University of Cambridge, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, United Kingdom. Email: martina.mcmenamin@mrc-bsu.cam.ac.uk

[3] Department of Biostatistics, University of Liverpool, Block F Waterhouse Building,
1-5 Brownlow Street, Liverpool, L69 3GL. Email: S.R.Dodd@liverpool.ac.uk

*Corresponding author

Abstract

Background

Clinical trials and other studies commonly assess effectiveness of an intervention through use of responder-based endpoints. These classify patients based on whether they meet a number of criteria which often involve continuous variables categorised as being above or below a threshold. The proportion of patients who are responders is estimated and, where relevant, compared between groups. An alternative method called the augmented binary method keeps the definition of the endpoint the same but utilises information contained within the continuous component to increase the power considerably (equivalent to increasing the sample size by >30%). In this article we summarise the method and investigate the variety of clinical conditions that use endpoints to which it could be applied.

Methods

We reviewed a database of physiological and mortality trial endpoints recommended for collection in clinical trials of different disorders. We identified responder-based endpoints where the augmented binary method would be useful for increasing power.

Results

We identified 68 new clinical areas where endpoints were used that would be more efficiently analysed using the augmented binary method.

Conclusions

The augmented binary method can potentially provide large benefits in a vast array of clinical areas. Further methodological development is needed to account for some types of endpoint.
Keywords: Augmented binary method; composite endpoint, efficiency, responder analysis, statistical analysis

Background

In clinical trials gathering evidence about the effectiveness of a medical intervention, it is necessary to specify a primary endpoint. An endpoint should represent how patients respond after being given the treatment; it should be expected that the distribution of the endpoint will be more favourable if a treatment is effective than if it is ineffective. In many disorders it is difficult to specify just one specific endpoint, as an intervention may have a variety of effects that cannot be adequately measured through one measurement. For this reason, it is common in many conditions to combine multiple distinct endpoints (which we will refer to as components) into a composite endpoint.

Composite endpoints have been recommended when there is large variability in the disease manifestation, e.g. complex multisystem diseases, allowing multiple equally relevant outcomes to be considered without the need to correct for multiplicity. They have also been advocated for rare diseases, where they might improve the power by increasing the number of events observed. On the other hand, they have been criticised for making trial results more difficult to interpret (1).

One specific type of composite endpoint is a composite responder endpoint, which divides patients into responders and non-responders on the basis of the set of components. Some of these components may be binary (present or absent), some may be continuous. In the case of continuous components, some dichotomisation is necessary, so that patients are responders only if the continuous component is above or below a specified threshold. In Table 1, we provide examples of some commonly used responder-based endpoints and their definitions. In some cases (such as tumour response in Table 1), a patient must meet all the criteria to be a responder; in other cases (such as Rheumatoid Arthritis in Table 1) a patient must meet a set number. Some responder endpoints are not composite and are just formed from a single dichotomised continuous endpoint.

Responder endpoints are appealing as they simplify several (potentially complex) pieces of information into one responder/non-responder variable. The proportion of patients who are responders serves as an easy to interpret measurement of the effectiveness of a treatment.

From a statistical point of view, however, this appealing simplicity comes at a non-appealing cost when one or more component is continuous. Dichotomising continuous variables loses information, a point which has been made many times (e.g. (2)). This means that if considering one continuous endpoint, it is substantially more efficient to analyse it as a continuous variable rather than dichotomise it and test as a binary variable. As a rule of thumb, the minimum cost of dichotomisation is

requiring a 35% higher sample size for the same level of statistical precision(2).

Assuming that avoiding dichotomisation is desirable, it is not obvious how this is possible when the responder endpoint consists of a mix of continuous and binary components. Even in the case of a single continuous component, there may be compelling clinical reasons to keep a responder endpoint dichotomised (3): ease of interpretation to researchers and patients, wide acceptance as important, meaningful clinical diagnosis (e.g. diabetes or hypertension).

This motivates statistical methods that can be used to keep to what is clinically relevant by inferring the proportion of patients who are responders, but utilise information contained in continuous components to improve the efficiency. For the single-component responder, this idea dates back to the 90s, where Suissa and Blais (4,5) proposed methods for doing this for a single continuous component case. To our knowledge, this method rarely is applied in practice despite its advantages over analysing the endpoint as binary. More recently, an approach known as the *augmented binary* method has been developed that allows composite responder endpoints (that consist of at least one continuous component) to be analysed in a more efficient way, whilst maintaining the definition of the endpoint.

In this paper (and associated supplementary material) we first describe the augmented binary method, focusing on its advantages and drawbacks. We next present a review that identifies new clinical areas where trial efficiency can be improved through use of the augmented binary method. Finally, we discuss some further developments to the method that are motivated by the review.

The augmented binary method - intuition, benefits and drawbacks

The augmented binary method extends previous work focused on a single dichotomised continuous endpoint (4,5) to composite responder endpoints with a mixture of continuous and binary endpoints. The original motivation was solid-tumour oncology (6,7), but subsequent papers have focused on developing the methodology for rheumatology(8) and rare diseases using composite endpoints (9).

For simplicity we focus on the case of a composite responder endpoint that combines a dichotomised continuous component with a binary component. For example, response in solid-tumour oncology consists of the sum of target lesion diameters shrinking by at least 30% from a baseline scan (dichotomised continuous) and no new tumour lesions appearing on a scan (binary). The traditional, binary analysis would work with the data on whether or not each patient is a responder or not. If a patient meets the criteria they are a responder, otherwise not. If analysing a randomised controlled trial (RCT), then one might test for a difference between arms in the proportion of patients who are responders with an

established method (e.g., logistic regression if there are baseline covariates to adjust for, Pearson chi-squared or Fisher exact test if not).

A detailed description of how to fit the method is provided in the supplementary material. The main intuition behind the method is to first fit a more sophisticated model to the data from the different components, and second to use this model to estimate a probability of response and test for a difference between arms. The second step can be thought of weighting the different patients as a proportion of a response with this proportion depending on how close the continuous component was to the threshold. This is demonstrated in Figure 1, where patients are measured on a continuous and binary component. The continuous measurement must be above 1 for the patient to be a responder, however patients must also meet the additional binary criteria. The binary method treats the information as 0s and 1s whereas the augmented binary method uses a ‘response weight’ which is determined from the underlying model and is higher as the continuous component increases. The supplementary material contains a link to an R package that can be used to fit the model. The benefit of the method is primarily the increased power. By better using the available information, the proportion of patients who respond (and therefore any differences between arms in a RCT) can be estimated more precisely. In more statistical language, the variance of the estimate is lower and the width of the confidence interval (CI) is narrower. Simulation studies presented in (6) found that the average gain in power was equivalent to increasing the sample size by between 30-50%

depending on the scenario. This theoretical gain in power has been confirmed in analysis of a real RCT in rheumatoid arthritis, which showed the reduction in CI width was equivalent to an increase in sample size of >50%. It should be emphasised that this gain in power does not rely on additional data being collected – it just comes from using the existing data more efficiently.

There are some additional benefits of the approach. First, due to the underlying model being fitted, it better allows for missing data. This is especially true when there is the possibility of some components having more missing data than others. Second, it may also help address issues of misclassification due to measurement error: if a patient is truly close to the responder threshold then a measurement error will have a potentially very large impact on the binary method, but a small impact on the augmented binary method.

There are also drawbacks. First, it is undoubtedly more complex to apply the method compared with standard binary approaches. Some code is available (see supplementary material) for applying the method in specific cases but a more generic implementation in different commonly used statistical programs is a high priority for the future. Second, the method makes more assumptions, for instance that the distribution of the continuous components is normal. This means that it is necessary to check this prior to analysing the data and use a suitable correction if assumptions are not met, such as applying a Box-Cox transformation (10)

to ensure the continuous component is normally distributed. Third, if the number of components or number of timepoints at which the endpoint is measured is large, applying the method can require a large amount of computational time. This is generally not an issue for an analysis of a single trial; however assessing the performance of the method on a large number of computer simulations can become infeasible.

Up to now, the method has been applied to datasets in solid tumour oncology(6,7), rheumatoid arthritis(8) and systemic lupus erythematosus (SLE)(11). Based on personal experience of peer-reviewing clinical trial papers and discussion with a wider group of clinicians, we hypothesised that there might be a much greater number of diseases where the augmented binary method could be useful. We decided that a more systematic attempt to identify these clinical areas was warranted.

Materials and methods

We made use of the COMET (Core Outcome Measures in Effectiveness Trials) database (http://www.comet-initiative.org/resources), which lists completed and ongoing projects in core outcome set (COS) development. COS represent the minimum that should be measured and reported in all clinical trials of a specific condition(12,13).

We reviewed physiological and mortality trial outcomes (categorised according to (14)) recorded within all core outcome sets (COS) in the

COMET database that were published before 2016. These were split amongst the three authors (JW MM and SD) to review. Each core outcome set paper was reviewed to determine if any responder (composite or categorised continuous) endpoints were recommended for reporting in all clinical trials within that condition. In some cases, a potentially relevant endpoint was not clearly described in the core outcome paper. In this case, we examined RCTs that had used the endpoint to determine whether it was a suitable responder endpoint.

Results

This process allowed us to identify 45 clinical areas (additional to solid tumour oncology, rheumatoid arthritis, and SLE) where the augmented binary method could be utilised to gain efficiency. An additional 23 clinical areas had used responder endpoints formed from a single categorised/dichotomised continuous variable. Table 2 breaks down the number by clinical classification. A full listing of these clinical areas is given in supplementary material. These are given by clinical classification in supplementary tables 1a-1m.

The clinical classifications that the method appears most useful to in terms of number of endpoints are rheumatology (11 found), non-solid tumour oncology (10 found), neurology (9 found), and cardiovascular (8 found).

We note that this was not a systematic review and represents a likely substantial underestimate of the number of clinical areas where suitable

endpoints are used, as our review only covered clinical areas which were covered by a COS published by 2016. As an example, table 1 mentions type II diabetes and shows diabetes remission would be a suitable endpoint; however, since there was no associated COS published by 2016, it does not appear in the identified clinical areas (although gestational diabetes does).

Discussion

In this paper we have shown how a more efficient analysis approach called the augmented binary method can be used to improve analysis of composite responder outcomes. The method allows retention of clinically relevant endpoints whilst improving the power of analyses by an amount equivalent to a considerable increase in sample size.

Through our review of core outcome sets we have found a great deal of new disease areas where the augmented binary method could be applied to gain power. We acknowledge that many of the core outcome sets were developed prior to best-practice guidance(15) existed and therefore the quality of them may differ.

Although the results indicate the widespread utility of the method, there are several areas where further methodological research is required to fully realise the possible benefits.

There are several endpoints which are typically analysed using time-to-event methods. Many progression, remission and relapse endpoints are used and the time until such a negative event occurs is the quantity of interest. Although the augmented binary method is well developed for composite responder outcomes that are analysed at a single timepoint or longitudinally, further work is needed to apply it to time-to-event outcomes.

In some cases, the composite responder outcomes are particularly complex with more than two components and with response being defined as meeting some, but not all, of the criteria. Recent work in this area (11) shows the potential efficiency gain is even larger in this case. In addition, the method, with some modification to the underlying latent variable model, could be applied in the case of a categorised responder endpoint with more than 2 levels.

We have focused on how the method can improve power of trials. An alternative way to use this improved power would be to reduce the sample size needed for a target power level. A barrier to widespread use of the method in this way is sample size estimation. In all cases that we have ever come across, the augmented binary method improves the power compared to a traditional binary analysis. However, the extent of the power gain is variable. It is therefore difficult to determine how to power the trial if the augmented binary method is to be used as a primary analysis. We are working on new methods to determine suitable sample size so that the trial is not inappropriately under- or over-powered.

Conclusions

In this paper we have shown that responder composite outcomes are used as primary clinical trial endpoints in many diseases. Analysing data from these trials using the augmented binary approach would improve power equivalent to increasing the sample size by at least 35%. Further methods research is needed to improve time-to-event analyses using these outcomes as events.

List of Abbreviations: CI – confidence interval; COMET - Core Outcome Measures in Effectiveness Trials; COS – core outcome set; RCT – randomized controlled trial; SLE - systemic lupus erythematosus

Declarations:

Ethics approval and consent to participate: not applicable

Consent for publication: not applicable

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: not applicable

Funding: MMM and JW are supported by funding from the MRC (grant code MC_UU_00002/6). JW is also supported by Cancer Research UK (C48553/A1811). None of the funding bodies had a role in the design of the study, analysis, interpretation of data or writing the manuscript.

Authors contributions: All authors contributed to the design of the manuscript and interpretation of the data. JW developed the first draft,

MMM and SD critically revised the manuscript and approved the final version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Acknowledgements: not applicable

References

1. Ross S. Composite outcomes in randomized clinical trials: arguments for and against. Am J Obstet Gynecol. 2007 Feb 1;196(2):119.e1-119.e6.
2. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332:1080.
3. DeCoster J, Iselin AR, Gallucci M. A conceptual and empirical examination of justifications for dichotomization. Psychol Methods. 2009;14:349–66.
4. Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol. 1991;44:241–8.
5. Suissa S, Blais L. Binary regression with continuous outcomes. Stat Med. 1995;14:247–55.
6. Wason JMS, Seaman SR. Using continuous data on tumour measurements to improve inference in phase II cancer studies. Stat Med. 2013;32(26):4639–50.
7. Lin C-J, Wason JMS. Improving phase II oncology trials using best observed RECIST response as an endpoint by modelling continuous

tumour measurements. Stat Med. 2017 Dec 20;36(29):4616–26.

8. Wason JMS, Jenkins M. Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. Rheumatology. 2016;55(10):1796–802.
9. McMenamin M, Berglind A, Wason JMS. Improving the analysis of composite endpoints in rare disease trials. Orphanet J Rare Dis. 2018;13(1):81.
10. Box GEP, Cox DR. An Analysis of Transformations. Vol. 26, Source: Journal of the Royal Statistical Society. Series B (Methodological). 1964.
11. McMenamin M, Barrett JK, Berglind A, Wason JMS. Employing latent variable models to improve efficiency in composite endpoint analysis. 2019 Feb 19;
12. Gargon E, Gurung B, Medley N, Altman DG, Blazeby JM, Clarke M, et al. Choosing Important Health Outcomes for Comparative Effectiveness Research: A Systematic Review. Scherer RW, editor. PLoS One. 2014 Jun 16;9(6):e99111.
13. Gargon E, Gorst SL, Harman NL, Smith V, Matvienko-Sikar K, Williamson PR. Choosing important health outcomes for comparative effectiveness research: 4th annual update to a systematic review of core outcome sets for research. Gillies K, editor. PLoS One. 2018 Dec 28;13(12):e0209869.
14. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to

help improve knowledge discovery. J Clin Epidemiol. 2018 Apr;96:84–92.

15. Kirkham JJ, Davis K, Altman DG, Blazeby JM, Clarke M, Tunis S, et al. Core Outcome Set-STAndards for Development: The COS-STAD recommendations. PLOS Med. 2017 Nov 16;14(11):e1002447.

Table 1 – examples of responder endpoints used in different areas of medicine; italicised components denote continuous dichotomisations.

Clinical area	Endpoint	Components and definition
Oncology	Tumour response	1. *Sum of longest diameter of target tumour lesions >=30% shrinkage from baseline* 2. No new tumour lesions
Rheumatology	ACR20	1. *Swollen joint count >=20% improvement* 2. *Tender joint count >=20% improvement* 3. 20% improvement in at least three of: a. *patient assessment* b. *physician assessment* c. *pain scale* d. *disability/functional questionnaire*

		e. acute phase reactant (ESR or CRP) *4.* No rescue therapy given.
Type II diabetes	Diabetes remission	*1. Glycated haemoglobin A_{1c} concentration <=6.5%* *2. Fasting glucose concentration <=5.6 mmol/L* 3. No non-study pharmacological treatment given

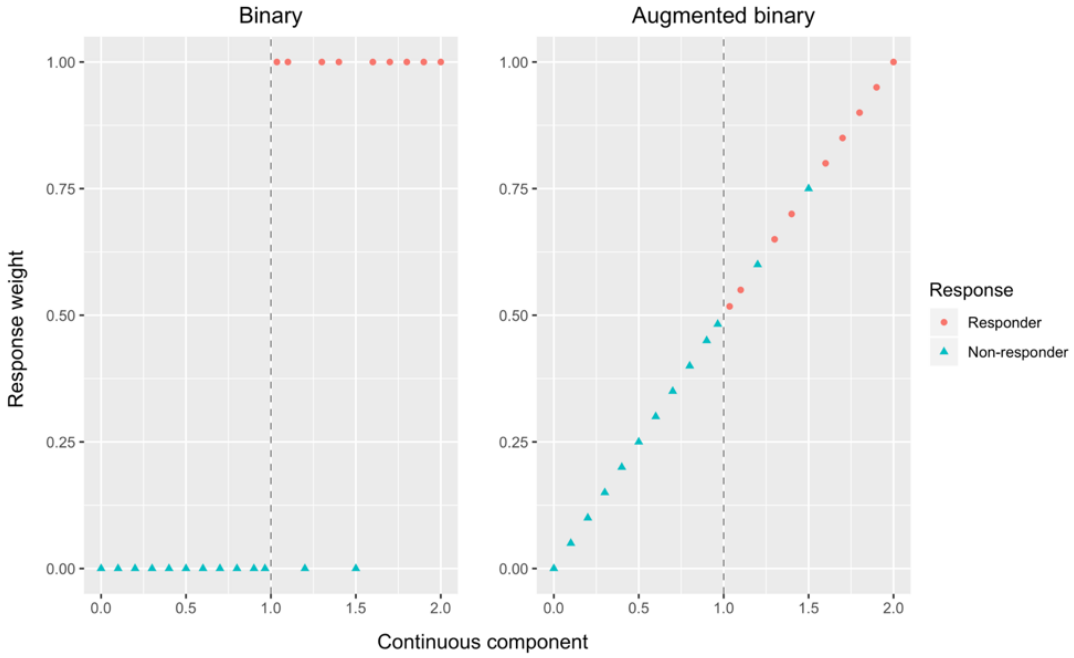
Table 2 – number of new clinical areas identified by classification; full list provided in supplementary material; if a condition had both composite

and non-composite responder endpoints identified, they were only included in the composite column. *excludes oncology

Classification	Number of conditions with suitable composite responder endpoints	Number of conditions with single-variable responder endpoints
Bleeding and Transfusion	2	1
Cancer*	8	2
Cardiovascular and circulation	5	3
Dentistry and vision	2	1
Gastroenterology	3	2
Infectious diseases	3	0
Lungs and airways	0	2
Mental health and addiction	3	1
Neurology	4	5
Orthopaedics and trauma	3	0
Renal and urology	2	1
Rheumatology	8	3
Unclassified	2	1

Figure 1 - demonstration of how information from patients is weighted by

the two different methods. Non-responders are made up of those in whom the continuous component is below 1 and those who do not respond according to another binary criterion. Underlying the augmented binary method is a joint model that is fitted to the continuous and binary data and yields fitted ‘response weights’ for each patient, which can then be compared between arms.



Appendix B

Orphanet Paper

McMenamin *et al. Orphanet Journal of Rare Diseases* (2018) 13:81
https://doi.org/10.1186/s13023-018-0819-1

Orphanet Journal of Rare Diseases

RESEARCH Open Access

Improving the analysis of composite endpoints in rare disease trials

Martina McMenamin[1*], Anna Berglind[2] and James M. S. Wason[1,3]

Abstract

Background: Composite endpoints are recommended in rare diseases to increase power and/or to sufficiently capture complexity. Often, they are in the form of responder indices which contain a mixture of continuous and binary components. Analyses of these outcomes typically treat them as binary, thus only using the dichotomisations of continuous components. The augmented binary method offers a more efficient alternative and is therefore especially useful for rare diseases. Previous work has indicated the method may have poorer statistical properties when the sample size is small. Here we investigate small sample properties and implement small sample corrections.

Methods: We re-sample from a previous trial with sample sizes varying from 30 to 80. We apply the standard binary and augmented binary methods and determine the power, type I error rate, coverage and average confidence interval width for each of the estimators. We implement Firth's adjustment for the binary component models and a small sample variance correction for the generalized estimating equations, applying the small sample adjusted methods to each sub-sample as before for comparison.

Results: For the log-odds treatment effect the power of the augmented binary method is 20-55% compared to 12-20% for the standard binary method. Both methods have approximately nominal type I error rates. The difference in response probabilities exhibit similar power but both unadjusted methods demonstrate type I error rates of 6–8%. The small sample corrected methods have approximately nominal type I error rates. On both scales, the reduction in average confidence interval width when using the adjusted augmented binary method is 17–18%. This is equivalent to requiring a 32% smaller sample size to achieve the same statistical power.

Conclusions: The augmented binary method with small sample corrections provides a substantial improvement for rare disease trials using composite endpoints. We recommend the use of the method for the primary analysis in relevant rare disease trials. We emphasise that the method should be used alongside other efforts in improving the quality of evidence generated from rare disease trials rather than replace them.

Keywords: Responder analysis, Composite endpoints, Improving efficiency

Background

For stakeholders in rare disease communities, it is imperative to keep in mind that rare diseases are far from 'rare' for those whose lives they consume. The last few decades have seen a societal shift which recognises this and has resulted in a much greater focus on rare disease research. This is characterised by a surge in patient advocacy groups, a shift in regulation and incentives, increased government funding of rare disease research and advances in technologies to improve international communication between rare disease experts and patients [1]. Despite this, for most rare diseases if treatment options even exist many of them have been approved with very limited evidence. Novel statistical design and analysis methods are needed to make the best use of information provided by studies in rare diseases [2, 3].

One way to maximise information from rare disease trials is to use composite endpoints [4]. These are endpoints which combine a number of individual outcomes in order to assess the effectiveness or efficacy of a treatment. They are typically used in situations where it is difficult

*Correspondence: martina.mcmenamin@mrc-bsu.cam.ac.uk
[1] MRC Biostatistics Unit, University of Cambridge, Forvie Site, Cambridge, UK
Full list of author information is available at the end of the article

to identify a single relevant endpoint to sufficiently capture the change in disease status incited by the treatment. Furthermore, if the components are appropriately chosen, composite endpoints that require an event in only one of the components (a or b or c etc.) may have the ability to improve the power to show a given treatment effect due to the increased number of events [5–7]. These characteristics appeal to rare diseases where many realisations of the diseases are highly variable and availability of the population may be a binding constraint.

Many composite endpoints take the form of responder indices where a binary indicator is formed based on whether the patient has experienced a predefined change in each of the components or not. In particular, in many disease areas the composite is a mixture of continuous and binary components. These endpoints frequently feature in rare autoimmune diseases and rare cancers. Examples of these are presented in Table 1, one of which is the chronic inflammatory disorder Behçet disease. A review of the research performed in this area concludes that evidence continues to be generated from anecdotal case reports rather than randomised trials [8]. As well as those shown in Table 1, any rare cancers using RECIST criteria (Response Evaluation Criteria In Solid Tumors) to define responders and non-responders use endpoints which assume this structure [9].

Table 1 Examples of rare diseases which could make use of the augmented binary method

Disease	Example responder endpoint
Primary biliary cholongitis (PBC)	• ALP $<$ 1.67 $\times$ ULN • Total bilirubin $<$ ULN • ALP decrease $\geq$ 15%
Behçets disease	• Length of principal intestinal ulcer compared to size at baseline (%) • No new lesions
Lupus Nephritis	• eGFR no more than 10% below preflare value or normal • Proteinuria UPC ratio $<$ 0.5 • Urine sediment: Inactive • No rescue therapy
Neuroblastoma	• $<$ 10mm residual soft tissue at primary site • Complete resolution of MIBG of FDG-PET uptake (for MIBG non avid tumours) at primary site
Advanced hepatocellular carcinoma	• $<$ 20% increase in the sum of the longest diameters of target lesions • No new lesions

ALP alkaline phosphatase, *ULN* upper limits of normal, *eGdFR* estimated glomerular filtration rate, *UPC* urinary protein to creatinine, *MIBG* metaiodobenzylguanidine, *FDG-PET* 18-fluorodeoxyglucose positron emission tomography

Analyses of these outcomes typically treat them as binary, thus only using the dichotomisations of continuous components. An alternative in these circumstances is the augmented binary method [10]. This involves jointly modelling the continuous component with the binary component in order to improve the efficiency of estimates by making use of how close patients were to being responders in the continuous component. For a fixed sample size, the method was shown to provide a substantial increase in the power over the standard binary method currently in use, whilst still making inference on the outcome of interest to clinicians. This was illustrated in both solid tumour cancer and rheumatoid arthritis data [10, 11].

Although the method provides substantially more power it also uses more parameters. Some evidence has suggested that it may not be suitable for trials with small samples, perhaps due to issues with asymptotics [10]. We will explore the properties of the augmented binary method in small samples and introduce and implement two small sample corrections from the literature to determine whether we can improve the performance.

If the gains provided by the augmented binary method in common diseases can be realised in smaller samples, this may allow us to gain information from randomised trials that would otherwise not have been possible. This could greatly improve outcomes for many rare disease patients. Further small sample applications of the method include earlier phase 2 trials, or when more doses are of interest but the number of patients are limited.

Methods

Data

In order to investigate the properties of the methods in small samples, we will use data from the OSKIRA-1 trial [12]. The trial was a phase III, multicentre, randomised, double-blind, placebo-controlled, parallel-group study investigating the use of fostamatinib in patients with active rheumatoid arthritis. For the purpose of investigating the small sample properties of the methods, we will only make use of two of the three arms in the trial, namely the fostamatinib 100 mg bid for 52 wks arm and the placebo arm.

A common responder endpoint used in rheumatoid arthritis is the ACR20, in which patients demonstrate clinical response if they achieve a 20% improvement from baseline, as measured by a continuous ACR-N (American College of Rheumatology) score. It is worth noting that the ACR-N score is a percentage change from baseline which is itself a composite combining 7 components but in what

follows we will treat this as a single measure, as is the case in practice. The structure of the endpoint is shown in Fig. 1.

A benefit of responder analyses is that we can easily incorporate additional information in the response definition. In the case of rheumatoid arthritis it is common to assign patients to being non-responders in the ACR20 endpoint if they require medications restricted by the protocol or withdraw from the study. Therefore, in order to be a responder to treatment a patient needs to tolerate treatment, must not receive restricted medications and they must demonstrate clinical response. This allows discontinuations of treatment for lack of efficacy or for adverse events to provide meaningful information on the drug effect and translates to estimating the effect of a combination of continuous and binary components.

Other endpoints of interest in rheumatoid arthritis are the ACR50 and ACR70 which dichotomise the ACR-N score at 50% and 70% respectively. We will discuss the findings and conclusions for the ACR20 endpoint in what follows, as this was the primary endpoint in the trial and is the endpoint that is generally used to formally evaluate benefit in the regulatory setting. Results for both the ACR50 and ACR70 endpoints are detailed in the supplementary material (see Additional file 1). These endpoints further characterise the benefit of a treatment by considering different levels of improvement from baseline. Furthermore, they will demonstrate how the methods perform with different response rates.

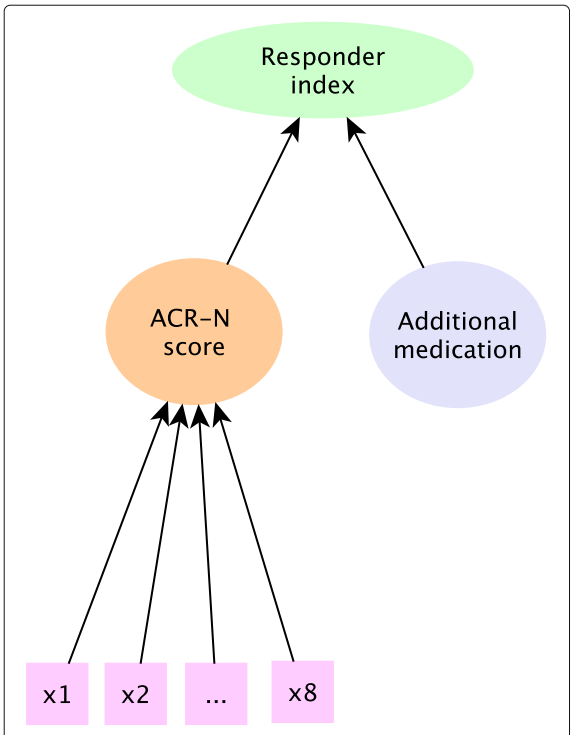


Fig. 1 Structure of the responder endpoint in rheumatoid arthritis. For the ACR20 endpoint, the continuous ACR-N score is dichotomised at 20% and combined with the rescue medication indicator to form a binary responder index. X1...X8 denotes the disease activity measures which are combined to form the continuous ACR-N score

Standard binary method

The method currently employed to analyse these endpoints in trials is a logistic regression on the binary indicator for response. We refer to this as the standard binary method.

The odds ratio and confidence interval are obtained directly. We can also obtain predicted probabilities for each patient as if they were treated $\tilde{p}_{i1}$ and not treated $\tilde{p}_{i0}$. This allows us to construct both the difference in response probabilities and the risk ratio. Their corresponding confidence intervals are obtained through the delta method, details of which are provided in the supplementary material (see Additional file 2).

Augmented binary method

The augmented binary method models the joint distribution of the continuous and binary components at multiple time points by employing factorisation techniques to model each of the components separately. We can then combine these to obtain predicted probabilities for each patient as if they were treated $\tilde{p}_{i1}$ and untreated $\tilde{p}_{i0}$ [10]. It follows that we can obtain the difference in response probabilities, the odds ratio and the risk ratio, as well as their confidence intervals as before (see Additional file 2).

Figure 2 shows a schematic for both the standard binary and augmented binary methods. From this it is clear that the augmented binary method models the components of the composite endpoint directly whereas the standard binary method throws away information before the analysis stage.

Note that we fit the repeated measures models for the continuous component in the augmented binary method using both generalised least squares (GLS) and generalised estimating equations (GEE).

Binary component adjustment

Albert and Anderson show that when fitting a logistic regression model to small samples, that although the likelihood converges, at least one parameter estimate may be theoretically infinite [13]. This phenomenon is commonly termed 'perfect separation' and occurs if the model can perfectly predict the response or if there are more parameters in the model than can be estimated because the data are sparse [14]. Firth provides an alternative to maximum

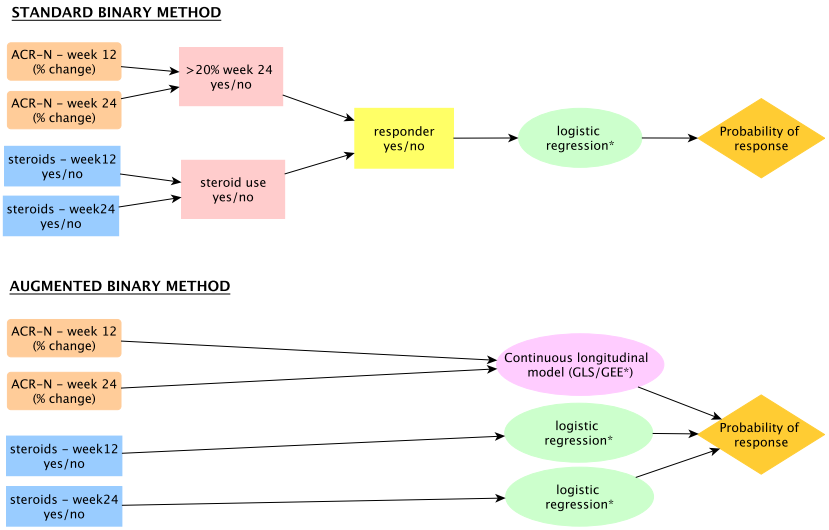


Fig. 2 Schematic comparing the stages involved in employing the standard binary and augmented binary methods (*small sample adjustments implemented)

likelihood estimation (MLE) in these circumstances [15]. This involves using penalised maximum likelihood (PML) to correct the mechanism producing the estimate, namely the score equation, rather than the estimate itself.

As maximum likelihood estimates are always biased away from zero in this setting, bias correction therefore involves some degree of shrinkage of the estimate towards this point. This results in the method also reducing the variance, so that bias reduction does not necessarily lead to a substantial loss in power. We will make these adjustments to both the standard binary method and the logit models in the augmented binary method. This can be easily implemented in R using the brglm package [16].

Continuous component adjustment

It is recognised that when using these methods when the number of clusters, in our case patients, is small that the robust standard error estimates are subject to downward bias, leading to inflated type I errors. We will implement a correction by Morel, Bokossa and Neerchal to inflate the variance estimate when modelling the continuous component using GEE methods [17]. We implement this in R using a modification of the code provided in the geesmv package [18].

The technical details for the models and adjustments are available in the supplementary material (see Additional file 2). The code to implement these in R is also available (see Additional file 3).

Assessing small sample properties

In order to determine the performance of the unadjusted and adjusted methods, we re-sample from the OSKIRA-1 trial. Employing re-sampling techniques allows us to investigate the properties of the methods under a realistic data structure.

To determine the power we re-sample 5000 replicates for each total sample size between 30 and 80 in increments of 10, which gives a Monte Carlo standard error of 0.3%. To ensure balance we randomly sample half of the total sample size we are interested in from the placebo arm and the other half from the 100 mg arm of the trial. We apply all methods to each sub-sample and record the treatment effect and 95% confidence interval. We do this for both the difference in response probabilities and log-odds estimates of the treatment effect. An estimate of the power is then the proportion of confidence intervals that do not contain zero. By re-sampling, rather than simulating from a known distribution, thinking of this quantity as power implicitly assumes the treatment effect in the trial to be the true treatment effect in the population. To ensure these results agree with the conventional power results, we present the power from a simulated example in the supplementary material (see Additional file 4).

To determine the type I error rate, we permute the treatment labels in the sub-samples in order to remove the association between treatment and outcome. An estimate of the type I error rate is then the proportion of confidence intervals that do not contain zero. The coverage is estimated as the complement of this. Again, to ensure

these results agree with when we have simulated under the null, we present an additional simulated example in the supplementary material. The median width of the confidence intervals and the average treatment effect for both methods are also presented in the supplementary material (see Additional file 5).

The unadjusted methods to be applied are the standard binary method, the augmented binary method with GLS and the augmented binary method with GEE. The adjustments refer to the standard binary method fitted with PML, the augmented binary method with GLS and PML and the augmented binary method with the GEE variance correction and PML.

Results

Log-odds scale

The power and type I error rates for the unadjusted and adjusted methods are shown in Fig. 3. The unadjusted augmented binary method provides higher power than the standard binary method for all sample sizes. The type I error rate of both methods is approximately 5%. Implementing the firth adjustment in the augmented binary method with GLS makes negligible difference to the power or type I error rate. In the adjusted augmented binary method with GEE, the type I error rate drops to 3–4%. Differences between the GLS and GEE estimators diminish with increasing sample size. The standard binary method experiences a substantial drop in type I error rate when the Firth correction is implemented.

Probability scale

Figure 4 shows the power and type I error rates for the difference in response probabilities. The power is similar to the log-odds case however both methods experience an inflation in type I error rate. Implementing the correction in the GLS augmented binary method results in a small improvement in the type I error rate with no power lost. GEE adjustments result in an average reduction in type I error of approximately 2.5% but the power drops to below that of the adjusted GLS. Again, differences in GLS and GEE diminish as the sample size increases. The adjustment for the standard binary reduces the type I error rate from 7% to approximately 5% however the power is below 20% for all sample sizes investigated.

Table 2 shows the average reduction in confidence interval width for the adjusted methods on both scales. We compare the standard binary with both implementations of the augmented binary method. We see from this that the augmented binary method with GLS offers the most precision. This translates to the adjusted augmented binary method requiring a 32% smaller sample size than what would be required for the adjusted standard binary method.

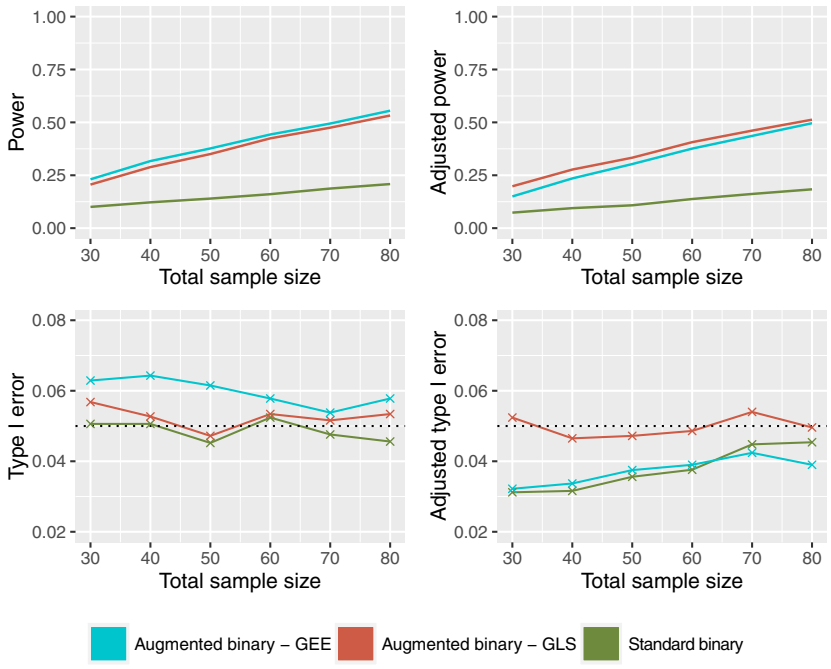


Fig. 3 Operating characteristics of the unadjusted (left) and adjusted (right) standard binary and augmented binary methods on the log-odds scale

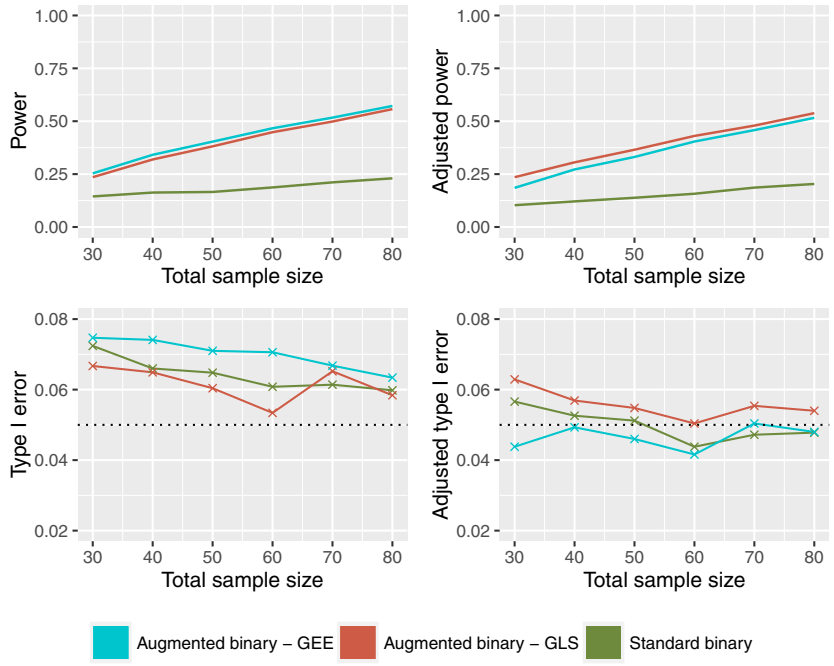


Fig. 4 Operating characteristics of the unadjusted (left) and adjusted (right) standard binary and augmented binary methods on the probability scale

To better understand the benefit of the small sample corrections it is useful to interpret the proportion of cases experiencing perfect separation alongside the average width of the confidence intervals. Table 3 shows the percentage of the 5000 sub-samples with confidence intervals for the difference in response probabilities larger than 1. This is shown for each method at each sample size. This would suggest that the corrections are most beneficial when N<60.

Table 2 Comparison in average confidence interval width for the small sample adjusted methods on the log-odds and probability scales

Comparison	Average reduction in C.I. width (%)	Reduction in required sample size (%)
Log-odds		
Standard binary vs		
Augmented binary (GLS)	17.4	31.8
Standard binary vs		
Augmented binary (GEE)	11.2	21.1
Difference in response probabilities		
Standard binary vs		
Augmented binary (GLS)	17.6	32.1
Standard binary vs		
Augmented binary (GEE)	12.3	23.1

C.I. confidence interval

Simulated example

Although re-sampling is beneficial as it details performance information under realistic data structures, the findings may be enriched by considering an example from a known distribution. We firstly set the probability of response equal to 0.470 in the treatment arm and 0.336 in the placebo arm, similar to the OSKIRA-1 study. Secondly, we simulate under the null where the probability of response equals 0.336 in both arms. We investigate power, type I error rate, average treatment effect estimates and average confidence interval width for the small sample adjusted binary and augmented binary methods. The results are presented in the supplementary material (see Additional file 4).

In summary, our comparative findings from the re-sampling are supported, in that the augmented binary method offers higher power and precision with a reduction in required sample size of approximately 38%. The augmented binary method has nominal type I error rate, which is consistent with the re-sampling results. However, the type I error for the adjusted standard binary method is 6.8–8.1%, which is higher than the type I error rates found from re-sampling. The absolute power estimates for both

Table 3 Percentage of cases experiencing extremely large variance due to perfect separation on probability scale (confidence interval for difference > 1)

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.00	0.00	10.9	0.00	10.6	0.24
40	0.00	0.00	3.71	0.00	3.81	0.02
50	0.00	0.00	1.11	0.00	1.20	0.00
60	0.00	0.00	0.24	0.00	0.30	0.00
70	0.00	0.00	0.04	0.00	0.08	0.00
80	0.00	0.00	0.00	0.00	0.00	0.00

methods also differ from those in the re-sampling results, however the comparative conclusions are the same. The methods provide approximately equal treatment effect estimates. A simulated example dataset is included should readers wish to fit the models (see Additional file 6).

Discussion

In this paper we have explored the small sample properties of the standard binary and augmented binary methods and proposed adjustments to improve them. It would appear that the increased efficiency of the augmented binary method does indeed translate to a small sample setting. The method performs better on the log-odds scale, where normality assumptions made when employing the delta method are best satisfied. These assumptions are more questionable when working with samples of this size on the probability scale, which is partly reflected in the differences in inflation present.

As rare disease trials are restricted in their capacity to detect treatment effects both because of small studies and few studies running in any given disease, it follows that maximising power within a single study is perhaps even more crucial than in more common diseases. Consequently, we recommend the use of the augmented binary method as the primary analysis method in trials of rare diseases using these endpoints.

When implementing the augmented binary method in rare disease trials, we recommend the use of the Firth adjustment for the logit models as it reduces the bias and variance of the estimates. This is especially valuable in this setting due to the restrictive nature of sample size. For the continuous component, we recommend the GLS estimator. As well as offering the largest power and precision, GLS methods make more realistic assumptions about the mechanism for missing responses, namely that they are missing at random rather than missing completely at random. Moreover they experience fewer convergence issues in very small samples.

We have previously acknowledged the potential utility of composite endpoints in rare diseases, however guidance must be followed in order to ensure valid and meaningful implementation in clinical trials [5]. Composite endpoints should be coherent, in that the components are measuring the same underlying pathophysiologic process. However, the components should not be so closely related that the patient is likely to experience all of them, hence making the combined endpoint redundant [19]. The magnitude of the gains from adopting a composite endpoint depends on the correlation between components, the direction of treatment effect in each component and hence the patient responder rates. It is therefore crucial for interpretation that effects are reported on individual components as secondary results. Binary components of the composite can be analysed with standard binary methods. Dichotomised continuous components of the composite may be analysed with standard binary methods, a continuous test or by testing the dichotomised component whilst making use of the continuous information, a technique similar to the augmented binary approach and originally proposed by Suissa [20]. This may be preferable to maintain the clinical definition of the component whilst improving the power.

It is useful to consider further the role of response rate in the composite endpoint on the operating characteristics of interest. The ACR50 and ACR70 results presented in the supplementary material indicate that power and type I error are highly dependant on responder rates and treatment effect scales (see Additional file 1). For the standard binary method, the results show deflations in the type I error rate on the log-odds scale and inflations on the probability scale, with type I error rates ranging from 0 to 10%. This is likely to be due to logistic regression methods having poorly estimated standard errors when there are few events per parameter, as is the case for the ACR50 and ACR70 endpoints [21]. Overall, the augmented binary method shows fewer deviations from nominal type I error rates whilst exhibiting increased power over the standard binary method in every scenario investigated.

The findings from the simulated example in the supplementary material further reiterate these problems with type I error rate control in the standard binary method. As the type I error rate is more stable for the augmented

binary method both in the re-sampling and the simulated example, we would suggest that it is perhaps more robust in the rare disease setting than logistic regression methods.

Although it is recognised that novel methods developed for use in rare diseases may be of more immediate utility than in common diseases, some resistance to implementing the augmented binary method in real rare disease trials may be experienced due to its increased complexity. To assist with this we supply full R code for all unadjusted and adjusted versions of the method. It is of paramount importance that the efficiency gains provided by this method are not used as a substitute for other important efforts and considerations undertaken when running rare disease trials. That is, the method should be used to complement efforts in establishing international, multi-centre trials with maximum feasible enrolment periods, alongside other achievable strategies to increase sample size; not to replace them.

There are some limitations in what we have presented. We have only investigated the performance of the method in small samples in relation to the rheumatoid arthritis endpoint. Similar procedures may be carried out in other data sets and the methods applied directly to rare disease data, to ensure these gains are always experienced across a range of responder indices and response rates. Moreover, due to the increased number of parameters, the augmented binary method starts to experience some problems when we reduce the total sample size to n=20. This is unlikely to be a problem in practice, as a randomised trial as small as that would be unusual. If required, it may be possible to make further assumptions in order to reduce the number of parameters to be estimated.

Our future work aims to improve the uptake of the augmented binary method in rare disease trials by developing methods for performing power calculations. This would overcome using the approximation that, for fixed power, the average gains equate to reducing the required sample size by at least 32%. A further extension on which we are currently working is developing joint modelling methods for the instance when the composite is a more complicated combination of outcomes, namely multiple continuous, ordinal and binary components. We expect these methods to exhibit even larger efficiency gains due to using information in multiple continuous and ordinal components. This will provide the potential to improve even further the frequency and quality of evidence generated in many rare disease areas.

Conclusion

In rare diseases where there are few or no available treatments and limited opportunity to test emerging new treatments, the power to detect an effective treatment is of critical importance. The augmented binary method with small sample adjustments offers a substantial improvement for trials in these populations over methods currently being used, which throw away valuable information. We recommend the use of the augmented binary method in relevant rare disease trials using composite endpoints and supply R code to assist with the implementation.

Additional files

Additional file 1: Results for the power and type I error rates of the ACR50 and ACR70 endpoints. (PDF 23 kb)

Additional file 2: Notation, models and small sample corrections. Technical detail for the models and small sample corrections. (PDF 92 kb)

Additional file 3: R code to fit methods and small sample adjustments. (ZIP 7 kb)

Additional file 4: Results for simulated examples from a known distribution. (PDF 52 kb)

Additional file 5: Median confidence interval widths and average treatment effect estimates from re-sampling. (PDF 23 kb)

Additional file 6: Simulated example dataset. (CSV 7 kb)

Abbreviations

ACR: American college of rheumatology; ALP: Alkaline phosphatase; eGFR: Estimated glomerular filtration rate; FDG-PET: 18-fluorodeoxyglucose positron emission tomography; GEE: Generalised estimating equations; GLS: Generalised least squares; MIBG: Metaiodobenzylguanidine; MLE: Maximum likelihood estimation; RECIST: Response evaluation criteria in solid tumors; PML: Penalised maximum likelihood; ULN: Upper limits of normal; UPC: Urinary protein to creatinine

Acknowledgements

We thank the reviewers for their very useful comments that helped improve the paper.

Funding

MMM and JW are supported by funding from the MRC (grant code MC_UU_00002/6). JW is also supported by Cancer Research UK (C48553/A1811). None of the funding bodies had a role in the design of the study, analysis, interpretation of data or writing the manuscript.

Availability of data and materials

As we are not owners of the trial data that is used for re-sampling, we are unable to make it available.

Authors' contributions

MMM contributed to the conception of the paper, conducted the analysis and drafted the paper. AB acquired the data, contributed to the interpretation of the results and contributed to the drafting and revision of the paper. JW contributed to the conception of the paper, inpretation of the results and drafting and revision of the paper. All authors have given final approval of the version to be published.

Ethics approval and consent to participate

The original trial was conducted in accordance with the ethics principles of the Declaration of Helsinki and was approved by the appropriate institutional review boards at each participating investigational center. All patients provided written informed consent. Our study uses anonymised samples of this study so no specific ethics approval is required.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]MRC Biostatistics Unit, University of Cambridge, Forvie Site, Cambridge, UK. [2]Global Medicines Development, Biometrics and Information Sciences, AstraZeneca, Gothenburg, Sweden. [3]Institute of Health and Society, Newcastle University, Newcastle, UK.

Received: 13 November 2017 Accepted: 1 May 2018
Published online: 22 May 2018

References

1. Griggs R, Batshaw M, Dunkle M, Gopal-Srivastava R, Kaye E, Krischer J. Clinical research for rare disease: Opportunities, challenges and solutions. Mol Genet Metab. 2009;91(1):20–6.
2. Joppi R, Bertele V, Garattini S. Orphan drug development is progressing too slowly. Br J Clin Pharamacol. 2006;61:355–60.
3. Hilgers R, Roes K, Stallard N. Directions for new developments on statistical design and analysis of small population group trials. Orphanet J Rare Dis. 2016;11(1):78.
4. Jonker A, Mills A, Lau L, Ando Y, Baroldi P, Bretz F, et al. Small population clinical trials: Challenges in the field of rare diseases. Technical report. 2016.
5. Ross S. Composite outcomes in randomized clinical trials: arguments for and against. Am J Obstet Gynecol. 2007;196(2):199–16.
6. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289:2554–9.
7. Montori V, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse J, Pacheco-Huergo V, Bryant D, et al. Validity of composite endpoints in clinical trials. BMJ. 2005;330:594–6.
8. Saleh Z, Arayssi T. Update on the therapy of behçet disease. Ther Adv Chronic Dis. 2014;5(3):112–34.
9. Eisenhauer E, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). Eur J Cancer. 2009;45:228–47.
10. Wason J, Seaman SR. Using continuous data on tumour measurements to improve inference in phase ii cancer studies. Stat Med. 2013;32(26):228–47. https://doi.org/10.1002/sim.5867.
11. Wason J, Jenkins M. Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. Rheumatology. 2016;55(10):1796–802.
12. Weinblatt ME, Genovese MC, Ho M, et al. Effects of fostamatinib, an oral spleen tyrosine kinase inhibitor, in rheumatoid arthritis patients with an inadequate response to methotrexate: results from a phase iii, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. Arthritis Rheumatol. 2014;66:3255–64.
13. Albert A, Anderson J. On the existence of maximum likelihood methods in logistic regression models. Biometrika. 1984;71(1):1–10.
14. Webb M, Wilson J, Chong J. An analysis of quasi-complete binary data with logistic models: Applications to alcohol abuse data. J Data Sci. 2004;2:273–85.
15. Firth D. Bias reduction of maximum likelihood estimation. Biometrika. 1993;80:27–38.
16. Brglm: Bias Reduction in Binomial-response Generalized Linear Models. 2013. https://cran.r-project.org/web/packages/brglm/index.html.
17. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of gee estimators. Biom J. 2003;45:395–409.
18. Wang M. Geesmv: Modified Variance Estimators for Generalized Estimating Equations, R package version 1.3 edn. 2015. https://CRAN.R-project.org/package=geesmv.
19. Multiple Analyses in Clinical Trials. Statistics for Biology and Health. Chap. 7: Introduction to composite endpoints. New York: Springer-Verlag; 2003.
20. Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol. 1991;44:241–8.
21. Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373–9.

Appendix C

Orphanet Paper: Supplementary Materials

C.1 Models and Small Sample Adjustments

Notation

Suppose we have i patients, with $i \in \{1, ..., n\}$. Let $T_i \in \{1, 2\}$ indicate the treatment arm of patient i. The baseline ACR-N score is y_{i0}, with Y_{i1}, Y_{i2} denoting the continuous ACR-N scores at the week 12 visit and week 24 visit respectively. F_{i1} is an indicator variable taking a value equal to 1 if the patient discontinues treatment or requires rescue medication before the week 12 visit. F_{i2} is the corresponding indicator for the period between the week 12 and week 24 visit. S_i is then a binary variable indicating whether or not patient i was a responder. For the ACR20 endpoint, $S_i = 1$ if $Y_{i2} \geqslant 20$ and $F_{i1} = F_{i2} = 0$.

Standard Binary Method

The standard binary method is a logistic regression on the binary indicator S_i.

$$logit\left(P\left(S_i = 1 | T_i, y_{i0}\right)\right) = \alpha + \beta T_i + \gamma y_{i0} \tag{C.1}$$

This provides us with maximum likelihood estimates $\hat{\theta}_{SB} = \{\alpha, \beta, \gamma\}$ and $Cov(\hat{\theta_{SB}})$. From this we can obtain a fitted probability of response for each patient i as if they were treated with the experimental treatment $\tilde{p}_{i1}$ and the control treatment $\tilde{p}_{i0}$.

From this we can then can construct various quantities of interest:

1. Difference in Response Probabilities

$$\tilde{\delta}_1 = \frac{\sum_{i=1}^{n} \tilde{p}_{i1} - \sum_{i=1}^{n} \tilde{p}_{i0}}{n} \tag{C.2}$$

2. Risk ratio

$$\tilde{\delta}_2 = \frac{\sum_{i=1}^{n} \tilde{p}_{i1}}{\sum_{i=1}^{n} \tilde{p}_{i0}} \tag{C.3}$$

3. Odds ratio

$$\tilde{\delta}_3 = \frac{\left(\frac{\sum_{i=1}^{n} \tilde{p}_{i1}}{n - \sum_{i=1}^{n} \tilde{p}_{i1}}\right)}{\left(\frac{\sum_{i=1}^{n} \tilde{p}_{i0}}{n - \sum_{i=1}^{n} \tilde{p}_{i0}}\right)} \tag{C.4}$$

Confidence intervals for these treatment effect estimates can be constructed by obtaining standard error estimates through the delta method. This requires the covariance matrix of the maximum likelihood estimates $\text{Cov}(\hat{\theta}_{SB})$ and the vector of partial derivatives of $\tilde{\delta}$ with respect to each of the parameter estimates, $''\tilde{\delta}$.
For example, the variance of $\tilde{\delta}_1$:

$$Var(\tilde{\delta}_1) = (''\tilde{\delta}_1)^T Cov(\hat{\theta}_{SB})(''\tilde{\delta}_1) \tag{C.5}$$

Augmented Binary Method

The augmented binary method models the joint distribution of $(Y_1, Y_2, F_1, F_2)|T, Y_0$ by employing factorisation techniques to model each of the components separately, as shown by the equations below.

$$\begin{aligned} Y_{ij} &= \alpha + \beta_1 T_i I\{j = 1\} + \beta_2 T_i I\{j = 2\} + \gamma y_{i0} + \delta_j + \varepsilon_{ij} \\ (\varepsilon_{i1}, \varepsilon_{i2}|T_i, y_{i0}) &\sim N\left((0,0), \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \end{aligned} \tag{C.6}$$

$$logit\left(P(F_{i1} = 1|T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \alpha_{F1} + \beta_{F1} T_i + \gamma_{F1} y_{i0} \tag{C.7}$$

$$logit\left(P(F_{i2} = 1|F_{i1} = 0, T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \alpha_{F2} + \beta_{F2} T_i + \gamma_{F2} Y_{i1} \tag{C.8}$$

We fit repeated measures models using both generalised least squares (GLS) and generalised estimating equations (GEE) to the continuous component. GLS estimates the variance-covariance matrix using restricted maximum likelihood methods and GEE makes use of robust variance estimation techniques.
After fitting these models and obtaining maximum likelihood estimates $\hat{\theta}_{AB} = \{\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}, \hat{\delta}_1, \hat{\delta}_2, \hat{\alpha_{F1}}, \hat{\beta_{F1}}, \hat{\gamma_{F1}}, \hat{\alpha_{F2}}, \hat{\beta_{F2}}, \hat{\gamma_{F2}}\}$, we can obtain the overall probability in response in each arm. For patient i, the probability of response in the ACR20 endpoint is:

$$
\begin{aligned}
&P(Y_{i2} \geqslant 20, F_{i1} = F_{i2} = 0 | T_i, y_{i0}) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(Y_{i2} \geqslant 20, F_{i1} = F_{i2} = 0 | T_i, y_{i0}, Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) f(y_{i1}, y_{i2}; T_i, y_{i0}) dy_2 dy_1 \\
&= \int_{-\infty}^{\infty} \int_{20}^{\infty} P(F_{i1} = F_{i2} = 0 | T_i, y_{i0}, Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) f(y_{i1}, y_{i2}; T_i, y_{i0}) dy_2 dy_1 \\
&= \int_{-\infty}^{\infty} \int_{20}^{\infty} P(F_{i2} = 0 | F_{i1} = 0, T_i, y_{i0}, Y_{i1} = y_{i1}) P(F_{i1} = 0 | T_i, y_{i0}, Y_{i1} = y_{i1}) f(y_{i1}, y_{i2}; T_i, y_{i0}) dy_2 dy_1
\end{aligned}
$$

Again, we can obtain a fitted probability of response for each patient i as if they were treated with the experimental treatment $\tilde{p}_{i1}$ and the control treatment $\tilde{p}_{i0}$. Treatment effect estimates and confidence intervals are constructed as before, where $Cov(\hat{\theta}_{AB})$ is as shown in equation (C.9).

$$
Cov(\hat{\theta}_{AB}) = \begin{pmatrix} Cov(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}, \hat{\delta}_1, \hat{\delta}_2) & 0 & 0 \\ 0 & Cov(\hat{\alpha_{F1}}, \hat{\beta_{F1}}, \hat{\gamma_{F1}}) & 0 \\ 0 & 0 & Cov(\hat{\alpha_{F2}}, \hat{\beta_{F2}}, \hat{\gamma_{F2}}) \end{pmatrix} \tag{C.9}
$$

Binary Component Adjustment

The penalised likelihood is shown below, where $L(\theta)$ is the usual likelihood function for a logit model and $I(\theta)$ is the information matrix.

$$
L^*(\theta) = L(\theta)|I(\theta)|^{\frac{1}{2}} \tag{C.10}
$$

Continuous Component Adjustment

The standard robust sandwich covariance estimator is shown in equation C.11.

$$V_{sandwich} = (\sum_{i=1}^{n} D_i V_i^{-1} D_i)^{-1} (\sum_{i=1}^{n} D_i V_i^{-1} \widehat{Cov}(Y_i) V_i^{-1} D_i) (\sum_{i=1}^{n} D_i V_i^{-1} D_i)^{-1} \quad \text{(C.11)}$$

where:

$D_i = \frac{\partial \mu_i}{\partial \beta}$

μ_i is the vector of mean responses

β the parameter vector

V_i is the working variance-covariance matrix for Y_i

$\widehat{Cov}(Y_i) = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.

The small sample adjusted variance estimator is shown in equation C.12.

$$V_{MBN} = \left(\sum_{i=1}^{n} D_i V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i V_i^{-1} \left(k\widehat{Cov}(Y_i) + \delta_m \xi V_i\right) V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i V_i^{-1} D_i\right)^{-1} \quad \text{(C.12)}$$

where:

$k = \frac{N-1}{N-p} \frac{n}{n-1}$

p is the number of parameters

N is the total number of observations

$$\delta_m = \begin{cases} \frac{p}{n-p}, & \text{if n>3p} \\ \frac{1}{2}, & \text{otherwise} \end{cases}$$

$$\xi = max\left(1, \frac{trace\left(\left(\sum_{i=1}^{n} D_i V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i V_i^{-1} Cov(Y_i) V_i^{-1} D_i\right)\right)}{p}\right)$$

C.2 Supplementary Results: ACR20

Table C.1: Median width of confidence intervals of the standard binary and augmented binary methods for the log-odds treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	3.019 (431.8)	3.000 (0.236)	2.509 (1390)	2.477 (0.212)	2.458 (1413.6)	2.739 (0.384)
40	2.602 (136.7)	2.592 (0.163)	2.155 (170.4)	2.145 (0.150)	2.134 (183.2)	2.325 (0.175)
50	2.320 (0.135)	2.314 (0.112)	1.924 (61.75)	1.919 (0.117)	1.916 (81.20)	2.053 (0.123)
60	2.117 (0.103)	2.112 (0.089)	1.755 (16.656)	1.753 (0.096)	1.755 (22.07)	1.861 (0.099)
70	1.959 (0.081)	1.954 (0.072)	1.624 (0.588)	1.862 (0.218)	1.630 (1.318)	1.712 (0.081)
80	1.832 (0.069)	1.828 (0.063)	1.521 (0.071)	1.378 (0.129)	1.531 (0.066)	1.598 (0.069)

Results shown on logarithmic scale

GLS generalised least squares, GEE generalised estimating equations

Table C.2: Median width of confidence intervals of the standard binary and augmented binary methods for the difference in response probabilities treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.675 (0.044)	0.686 (0.037)	0.567 (248.8)	0.554 (0.045)	0.554 (245.1)	0.614 (0.142)
40	0.592 (0.030)	0.597 (0.026)	0.492 (36.558)	0.464 (0.062)	0.488 (39.049)	0.528 (0.062)
50	0.536 (0.023)	0.539 (0.021)	0.442 (14.154)	0.470 (0.059)	0.442 (18.558)	0.472 (0.033)
60	0.490 (0.018)	0.492 (0.017)	0.406 (3.985)	0.403 (0.021)	0.407 (5.229)	0.429 (0.024)
70	0.455 (0.015)	0.456 (0.014)	0.377 (0.142)	0.375 (0.018)	0.379 (0.319)	0.396 (0.020)
80	0.426 (0.013)	0.428 (0.012)	0.354 (0.016)	0.351 (0.016)	0.357 (0.017)	0.371 (0.017)

Results shown on probability scale

GLS generalised least squares, GEE generalised estimating equations

Table C.3: Average treatment effect in subsamples using the standard binary and augmented binary methods for the log-odds treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.606 (1.096)	0.526 (0.737)	0.783 (0.649)	0.747 (0.619)	0.812 (0.678)	0.773 (0.646)
40	0.595 (0.745)	0.543 (0.634)	0.794 (0.544)	0.765 (0.525)	0.828 (0.565)	0.796 (0.545)
50	0.572 (0.587)	0.536 (0.547)	0.790 (0.478)	0.767 (0.465)	0.821 (0.495)	0.795 (0.480)
60	0.570 (0.541)	0.540 (0.510)	0.788 (0.435)	0.770 (0.425)	0.816 (0.449)	0.795 (0.438)
70	0.577 (0.476)	0.551 (0.453)	0.794 (0.394)	0.902 (0.456)	0.821 (0.406)	0.802 (0.398)
80	0.568 (0.455)	0.546 (0.436)	0.790 (0.367)	0.707 (0.333)	0.817 (0.377)	0.801 (0.370)

Results shown on logarithmic scale

GLS generalised least squares, GEE generalised estimating equations

Table C.4: Average treatment effect in subsamples using the standard binary and augmented binary methods for the difference in response probabilities treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.129 (0.178)	0.118 (0.163)	0.180 (0.145)	0.170 (0.137)	0.179 (0.153)	0.168 (0.143)
40	0.133 (0.153)	0.124 (0.142)	0.185 (0.124)	0.171 (0.116)	0.191 (0.128)	0.182 (0.123)
50	0.131 (0.132)	0.124 (0.125)	0.186 (0.110)	0.195 (0.117)	0.193 (0.113)	0.185 (0.109)
60	0.131 (0.123)	0.125 (0.117)	0.187 (0.101)	0.181 (0.098)	0.193 (0.104)	0.186 (0.100)
70	0.134 (0.109)	0.129 (0.104)	0.189 (0.091)	0.184 (0.089)	0.195 (0.094)	0.190 (0.091)
80	0.133 (0.104)	0.128 (0.101)	0.189 (0.085)	0.184 (0.083)	0.195 (0.087)	0.190 (0.085)

Results shown on probability scale

GLS generalised least squares, GEE generalised estimating equations

Table C.5: Average treatment effect in permuted subsamples using the standard binary and augmented binary methods for the log-odds treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	0.011 (1.171)	0.005 (0.742)	0.001 (0.665)	0.049 (0.641)	-0.002 (0.673)	-0.002 (0.646)
40	0.012 (0.755)	0.006 (0.645)	-0.005 (0.559)	0.000 (0.537)	0.014 (0.583)	0.002 (0.555)
50	0.000 (0.608)	0.004 (0.575)	0.007 (0.487)	-0.003 (0.475)	-0.003 (0.512)	0.012 (0.492)
60	-0.005 (0.557)	0.004 (0.527)	-0.003 (0.449)	0.000 (0.437)	-0.004 (0.461)	-0.005 (0.452)
70	-0.004 (0.507)	0.001 (0.411)	0.004 (0.413)	0.000 (0.487)	-0.004 (0.424)	-0.007 (0.417)
80	0.005 (0.467)	-0.009 (0.457)	0.005 (0.390)	-0.002 (0.354)	0.000 (0.401)	-0.001 (0.386)

Results shown on logarithmic scale

GLS generalised least squares, GEE generalised estimating equations

Table C.6: Average treatment effect in permuted subsamples using the standard binary and augmented binary methods for the difference in response probabilities treatment effect

	Standard binary		Augmented binary (GLS)		Augmented binary (GEE)	
N	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
30	-0.003 (0.182)	-0.002 (0.166)	0.001 (0.151)	0.000 (0.144)	-0.002 (0.154)	0.000 (0.145)
40	0.001 (0.156)	0.000 (0.148)	0.001 (0.130)	-0.002 (0.123)	-0.004 (0.136)	0.000 (0.130)
50	-0.001 (0.142)	-0.002 (0.134)	0.002 (0.117)	0.000 (0.112)	0.001 (0.121)	0.000 (0.118)
60	0.000 (0.129)	-0.002 (0.121)	0.002 (0.106)	-0.003 (0.104)	-0.002 (0.113)	0.001 (0.108)
70	-0.004 (0.120)	0.000 (0.114)	0.001 (0.101)	0.001 (0.098)	0.001 (0.104)	-0.001 (0.101)
80	-0.002 (0.112)	0.000 (0.108)	0.001 (0.094)	0.000 (0.092)	0.000 (0.097)	-0.001 (0.094)

Results shown on probability scale

GLS generalised least squares, GEE generalised estimating equations

C.3 Supplementary Results: ACR50, ACR70

Table C.7: Type I error of the log-odds ACR50 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Type I error			Small sample adjusted type I error		
N	Binary	AugBin (GLS)	AugBin (GEE)	Binary	AugBin (GLS, PML)	AugBin (GEE adj, PML)
30	0.005 (0.001)	0.059 (0.003)	0.069 (0.004)	0.005 (0.001)	0.054 (0.003)	0.035 (0.003)
40	0.011 (0.001)	0.058 (0.003)	0.068 (0.004)	0.010 (0.001)	0.048 (0.003)	0.040 (0.003)
50	0.019 (0.002)	0.053 (0.003)	0.063 (0.003)	0.014 (0.002)	0.056 (0.003)	0.044 (0.002)
60	0.025 (0.002)	0.055 (0.003)	0.060 (0.003)	0.020 (0.002)	0.052 (0.003)	0.041 (0.002)
70	0.034 (0.003)	0.057 (0.003)	0.059 (0.003)	0.025 (0.002)	0.050 (0.003)	0.044 (0.003)
80	0.034 (0.003)	0.056 (0.003)	0.059 (0.003)	0.029 (0.002)	0.053 (0.003)	0.050 (0.003)

Table C.8: Power of the log-odds ACR50 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Power			Small sample adjusted power		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.022 (0.002)	0.196 (0.006)	0.232 (0.006)	0.034 (0.003)	0.208 (0.006)	0.150 (0.005)
40	0.061 (0.003)	0.263 (0.006)	0.298 (0.006)	0.073 (0.004)	0.265 (0.006)	0.216 (0.006)
50	0.133 (0.005)	0.324 (0.007)	0.359 (0.007)	0.142 (0.005)	0.321 (0.007)	0.279 (0.006)
60	0.193 (0.006)	0.381 (0.007)	0.407 (0.007)	0.207 (0.006)	0.378 (0.007)	0.339 (0.007)
70	0.269 (0.006)	0.436 (0.007)	0.459 (0.007)	0.277 (0.006)	0.433 (0.007)	0.400 (0.007)
80	0.357 (0.007)	0.493 (0.007)	0.512 (0.007)	0.348 (0.007)	0.490 (0.007)	0.460 (0.007)

Table C.9: Type I error of the ACR50 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Type I error			Small sample adjusted type I error		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.091 (0.004)	0.054 (0.003)	0.065 (0.003)	0.039 (0.003)	0.051 (0.003)	0.031 (0.002)
40	0.065 (0.003)	0.059 (0.003)	0.066 (0.003)	0.045 (0.003)	0.054 (0.003)	0.038 (0.003)
50	0.067 (0.004)	0.055 (0.003)	0.061 (0.003)	0.042 (0.003)	0.054 (0.003)	0.035 (0.003)
60	0.055 (0.003)	0.061 (0.003)	0.062 (0.003)	0.047 (0.003)	0.052 (0.003)	0.046 (0.003)
70	0.054 (0.003)	0.051 (0.003)	0.060 (0.003)	0.051 (0.003)	0.053 (0.003)	0.044 (0.003)
80	0.064 (0.003)	0.049 (0.003)	0.061 (0.003)	0.045 (0.003)	0.058 (0.003)	0.044 (0.003)

Table C.10: Power of the ACR50 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Power			Small sample adjusted power		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.243 (0.006)	0.202 (0.006)	0.208 (0.006)	0.157 (0.005)	0.211 (0.006)	0.133 (0.005)
40	0.273 (0.006)	0.270 (0.006)	0.284 (0.006)	0.224 (0.006)	0.272 (0.006)	0.199 (0.005)
50	0.318 (0.007)	0.334 (0.007)	0.344 (0.007)	0.274 (0.007)	0.329 (0.007)	0.264 (0.007)
60	0.363 (0.007)	0.395 (0.007)	0.394 (0.007)	0.314 (0.007)	0.385 (0.007)	0.324 (0.007)
70	0.415 (0.007)	0.441 (0.007)	0.444 (0.007)	0.370 (0.007)	0.436 (0.007)	0.387 (0.007)
80	0.474 (0.007)	0.500 (0.007)	0.498 (0.007)	0.432 (0.007)	0.496 (0.007)	0.446 (0.007)

Table C.11: Type I error of the log-odds ACR70 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Type I error			Small sample adjusted type I error		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.000 (0.000)	0.063 (0.003)	0.074 (0.004)	0.000 (0.000)	0.056 (0.003)	0.036 (0.003)
40	0.000 (0.000)	0.061 (0.003)	0.068 (0.004)	0.000 (0.000)	0.054(0.003)	0.041 (0.003)
50	0.000 (0.000)	0.052 (0.003)	0.071 (0.004)	0.000 (0.000)	0.058 (0.003)	0.045 (0.003)
60	0.000 (0.000)	0.056 (0.003)	0.064 (0.003)	0.000 (0.000)	0.058 (0.003)	0.038 (0.003)
70	0.000 (0.000)	0.048 (0.003)	0.058 (0.003)	0.002 (0.001)	0.058 (0.003)	0.046 (0.003)
80	0.002 (0.001)	0.055 (0.003)	0.055 (0.003)	0.002 (0.001)	0.053 (0.003)	0.046 (0.003)

Table C.12: Power of the log-odds ACR70 response in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Power			Small sample adjusted power		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.000 (0.000)	0.191 (0.006)	0.229 (0.006)	0.000 (0.000)	0.205 (0.006)	0.142 (0.005)
40	0.000 (0.000)	0.247 (0.006)	0.284 (0.006)	0.000 (0.000)	0.253 (0.006)	0.199 (0.006)
50	0.001 (0.001)	0.307 (0.007)	0.340 (0.007)	0.004 (0.001)	0.308 (0.007)	0.258 (0.006)
60	0.003 (0.001)	0.352 (0.007)	0.378 (0.007)	0.007 (0.001)	0.350 (0.007)	0.312 (0.007)
70	0.007 (0.001)	0.411 (0.007)	0.433 (0.007)	0.021 (0.002)	0.411 (0.007)	0.370 (0.007)
80	0.015 (0.001)	0.460 (0.007)	0.478 (0.007)	0.036 (0.003)	0.460 (0.007)	0.424 (0.007)

Table C.13: Type I error of the ACR70 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Type I error			Small sample adjusted type I error		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.056 (0.003)	0.029 (0.002)	0.031 (0.002)	0.006 (0.001)	0.030 (0.002)	0.008 (0.001)
40	0.040 (0.003)	0.032 (0.002)	0.039 (0.003)	0.013 (0.002)	0.034 (0.003)	0.010 (0.001)
50	0.032 (0.003)	0.031 (0.002)	0.037 (0.003)	0.020 (0.002)	0.034 (0.003)	0.015 (0.002)
60	0.041 (0.003)	0.034 (0.003)	0.041 (0.003)	0.022 (0.002)	0.036 (0.003)	0.022 (0.002)
70	0.049 (0.003)	0.040 (0.003)	0.040 (0.003)	0.020 (0.002)	0.045 (0.003)	0.025 (0.002)
80	0.048 (0.003)	0.045 (0.003)	0.039 (0.003)	0.026 (0.002)	0.040 (0.003)	0.031 (0.002)

Table C.14: Power of the ACR70 difference in response probabilities in standard binary and augmented binary methods in 5000 sub-samples where GLS is generalised least squares, GEE is generalised estimating equations, PML is penalised maximum likelihood and GEE adj is the GEE small sample adjustment

	Power			Small sample adjusted power		
N	Binary	Aug Bin (GLS)	Aug Bin (GEE)	Binary	Aug Bin (GLS, PML)	Aug Bin (GEE adj, PML)
30	0.108 (0.004)	0.118 (0.005)	0.110 (0.004)	0.029 (0.002)	0.120 (0.005)	0.034 (0.003)
40	0.112 (0.004)	0.173 (0.005)	0.165 (0.005)	0.065 (0.003)	0.174 (0.005)	0.077 (0.004)
50	0.136 (0.004)	0.243 (0.006)	0.228 (0.006)	0.095 (0.004)	0.240 (0.006)	0.140 (0.005)
60	0.185 (0.005)	0.304 (0.007)	0.281 (0.006)	0.128 (0.005)	0.301 (0.006)	0.206 (0.006)
70	0.250 (0.006)	0.366 (0.007)	0.345 (0.007)	0.156 (0.005)	0.363 (0.007)	0.269 (0.006)
80	0.288 (0.006)	0.417 (0.007)	0.399 (0.007)	0.184 (0.005)	0.415 (0.007)	0.331 (0.007)

C.4 Supplementary Results: Simulated Example

Model

$$Y_{ij} = \alpha + \beta_1 T_i I\{j=1\} + \beta_2 T_i I\{j=2\} + \gamma y_{i0} + \delta_j + \varepsilon_{ij}$$
$$(\varepsilon_{i1}, \varepsilon_{i2} | T_i, y_{i0}) \sim N\left((0,0), \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \tag{C.13}$$

$$logit\left(P(F_{i1}=1|T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \alpha_{F1} + \beta_{F1} T_i + \gamma_{F1} y_{i0} \tag{C.14}$$

$$logit\left(P(F_{i2}=1|F_{i1}=0, T_i, y_{i0}, Y_{i1}, Y_{i2})\right) = \alpha_{F2} + \beta_{F2} T_i + \gamma_{F2} Y_{i1} \tag{C.15}$$

Results

We investigate the power and type I error rate for the small sample adjusted measures for the difference in response probability estimator, as shown below.

$$\tilde{\delta}_1 = \frac{\sum_{i=1}^{n} \tilde{p}_{i1} - \sum_{i=1}^{n} \tilde{p}_{i0}}{n} \tag{C.16}$$

where $\tilde{p}_{i1}$ and $\tilde{p}_{i0}$ are the fitted probabilities of response for patient i on the experimental treatment and the control treatment respectively.

Table C.15: Power and average confidence interval width in ACR20 response in the small sample adjusted standard binary and augmented binary methods in 5000 simulations

Total sample	$\tilde{\delta}_1$ (S.D.)		Power		Average CI width		Sample size
size	Bin	Aug bin	Bin	Aug bin	Bin	Aug bin	reduction (%)
30	0.128 (0.167)	0.130 (0.121)	0.145	0.172	0.630	0.496	38.0
40	0.132 (0.145)	0.133 (0.106)	0.179	0.226	0.550	0.431	38.6
50	0.138 (0.129)	0.135 (0.097)	0.213	0.278	0.493	0.386	38.7
60	0.137 (0.120)	0.136 (0.088)	0.240	0.329	0.452	0.353	39.0
70	0.135 (0.113)	0.136 (0.083)	0.269	0.367	0.419	0.328	38.7
80	0.138 (0.103)	0.138 (0.077)	0.293	0.425	0.392	0.306	39.1

$\alpha = -15, \beta_1 = 2.5, \beta_2 = 2, \gamma = 4.1, \delta_1 = 6, \delta_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6, \alpha_{F1} = -3.8, \beta_{F1} = -0.1, \gamma_{F1} = 0.4, \alpha_{F2} = -0.8, \beta_{F2} = -0.08, \gamma_{F2} = -0.008, \tilde{\delta}_1 \approx 0.134$

Table C.16: Type I error rate and average confidence interval width in ACR20 response in the small sample adjusted standard binary and augmented binary methods in 5000 simulations

Total sample	$\tilde{\delta}_1$ (S.D.)		Type I error		Average CI width	
size	Bin	Aug bin	Bin	Aug bin	Bin	Aug bin
30	0.002 (0.157)	0.001 (0.102)	0.068	0.047	0.596	0.426
40	-0.001 (0.143)	0.001 (0.092)	0.080	0.047	0.517	0.370
50	0.000 (0.128)	-0.002 (0.081)	0.081	0.044	0.465	0.332
60	-0.001 (0.118)	0.000 (0.075)	0.079	0.043	0.425	0.303
70	-0.001 (0.107)	0.000 (0.070)	0.073	0.043	0.394	0.282
80	0.000 (0.104)	0.000 (0.065)	0.081	0.049	0.369	0.263

$\alpha = -15, \beta_1 = 0, \beta_2 = 0, \gamma = 4.1, \delta_1 = 6, \delta_2 = 12, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.6, \alpha_{F1} = -3.8, \beta_{F1} = 0, \gamma_{F1} = 0.4, \alpha_{F2} = -0.8, \beta_{F2} = 0, \gamma_{F2} = -0.008, \tilde{\delta}_1 = 0$

Appendix D

Small Sample Adjusted Methods: R Code

D.1 Augmented Binary: GLS

```
library(gtools)
library(geepack)
library(nlme)
library(R2Cuba)
library(boot)
library(mvtnorm)
library(MASS)
library(brglm)
library(gee)


###AUGMENTED BINARY METHOD - GENERALISED LEAST SQUARES (GLS)
##UNADJUSTED

integrand<-function(acrn,meantreated,meanuntreated,Sigma,failure1,failure2,baseline)
{
  n=length(baseline)

  f1treat=inv.logit(cbind(rep(1,n),baseline,rep(1,n))%*%failure1$coefficient)
  f1untreat=inv.logit(cbind(rep(1,n),baseline,rep(0,n))%*%failure1$coefficient)

  f2treat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(1,n))%*%failure2$coefficient)
  f2untreat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(0,n))%*%failure2$coefficient)

  pdftreat=dmvnorm(cbind(-meantreated[,1]+acrn[1],-meantreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(Sigma
       [1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))
  pdfuntreat=dmvnorm(cbind(-meanuntreated[,1]+acrn[1],-meanuntreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(
       Sigma[1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))

  return(c(mean((1-f1treat)*(1-f2treat)*pdftreat),mean((1-f1untreat)*(1-f2untreat)*pdfuntreat)))
}


probofsuccess=function(continuous,baseline,failure1,failure2,dich)
{

  n=length(baseline)

  meantreated=cbind(cbind(rep(1,n),rep(1,n),rep(0,n),baseline,rep(1,n))%*%continuous$coefficient,
                    cbind(rep(1,n),rep(0,n),rep(1,n),baseline,rep(2,n))%*%continuous$coefficient)

  meanuntreated=cbind(cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(1,n))%*%continuous$coefficient,
```

```
                    cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(2,n))%*%continuous$coefficient)


  #find lower and upper points for integration:
  maxmean1=max(c(meantreated[,1],meanuntreated[,1]))
  maxmean2=max(c(meantreated[,2],meanuntreated[,2]))
  minmean1=min(c(meantreated[,1],meanuntreated[,1]))
  minmean2=min(c(meantreated[,2],meanuntreated[,2]))


  #integrate

  a=cuhre(2,2,integrand=integrand,meantreated=meantreated,meanuntreated=meanuntreated,Sigma=getVarCov(continuous),
       failure1=failure1,failure2=failure2,baseline=baseline,lower=c(qnorm(1e-08,minmean1,sqrt(getVarCov(
       continuous)[1,1])),qnorm(1e-08,minmean2,sqrt(getVarCov(continuous)[2,2]))),
          upper=c(qnorm(1-1e-08,maxmean1,sqrt(getVarCov(continuous)[1,1])), dich),flags=list(verbose=0,final=1,
               pseudo.random=0,mersenne.seed=NULL))


  #return(a$value[1]-a$value[2]) #ABSOLUTE VALUE
  return(log(a$value[1]/(1-a$value[1]))-log(a$value[2]/(1-a$value[2]))) #LOGODDS
}

getVarCov.gls <-
  function(obj, ...)
  {

    S <- corMatrix(obj$modelStruct$corStruct)
    #get which individuals have a full correlation matrix:

    temp1=sapply(S,function(x)return(sum(dim(x))))


    vw  <-  1/varWeights(obj$modelStruct$varStruct)

    cor=S[[min(which(temp1==4))]]

    varianceweights=c(min(vw),max(vw))

    vars=(obj$sigma*varianceweights)^2

    result  <-  t(cor * sqrt(vars))*sqrt(vars)
    result
  }


get.partials<-function(continuous,baseline,failure1,failure2,dich)
{

  fit1=probofsuccess(continuous,baseline,failure1,failure2,dich)
  augbin.partials=as.vector(rep(0,11))

  #continuous model

  for(i in 1:5)
  {

    valueupdate1=continuous
    valueupdate1$coefficient[i]=valueupdate1$coefficient[i]+0.000001

    updateprob=probofsuccess(valueupdate1,baseline,failure1,failure2,dich)

    augbin.partials[i]=(updateprob-fit1)/0.000001

  }

  #failure model1

  for(i in 1:3)
  {

    valueupdate2=failure1
    valueupdate2$coefficient[i]=valueupdate2$coefficient[i]+0.000001

    updateprob=probofsuccess(continuous,baseline,valueupdate2,failure2,dich)
```

```
113
114     augbin.partials[i+5]=(updateprob-fit1)/0.000001
115   }
116
117   #failure model2
118
119   for(i in 1:3)
120   {
121
122     valueupdate3=failure2
123     valueupdate3$coefficient[i]=valueupdate3$coefficient[i]+0.000001
124
125     updateprob=probofsuccess(continuous,baseline,failure1,valueupdate3,dich)
126
127     augbin.partials[i+8]=(updateprob-fit1)/0.000001
128   }
129
130   return(c(augbin.partials,fit1))
131 }
132
133
134
135 ##################################################
136 #APPLY IN RHEUMATOID ARTHRITIS EXAMPLE FROM PAPER:#
137 ##################################################
138
139 dichotomisationthreshold=20 #change this depending on example
140 data=#OSKIRA-1 TRIAL
141
142   acrn.12<-data$acrn12 #CONTINUOUS MEASURE AT TIME POINT ONE
143 acrn.24<-data$acrn24 #CONTINUOUS MEASURE AT TIME POINT TWO
144 rescuemedicationupto12<-data$rescuemedicationupto12 #BINARY MEASURE AT TIME POINT ONE
145 rescuemedicationupto24<-data$rescuemedicationupto24 #BINARY MEASURE AT TIME POINT TWO
146 baselinediseaseactivity<-data$baselinediseaseactivity #BASELINE CONTINUOUS SCORE
147 arm<-data$arm #BINARY TREATMENT INDICATOR
148 patientid<-data$patientid #PATIENT NUMBER
149
150
151 #CHANGE TO RELATIVE ACRN AND BOXCOX TRANSFORM
152 acrn.12=1-acrn.12/100
153 acrn.24=1-acrn.24/100
154
155 lm1=lm(acrn.12~1)
156 lm2=lm(acrn.24~1)
157
158 temp=boxcox(lm1,plotit=F,lambda=seq(0,2,length=100))
159 lambda.12=temp$x[which.max(temp$y)]
160 temp=boxcox(lm2,plotit=F,lambda=seq(0,2,length=100))
161 lambda.24=temp$x[which.max(temp$y)]
162 lambda=mean(c(lambda.12,lambda.24))
163
164 acrn.12=boxcoxtransform(acrn.12,mean(c(lambda.12,lambda.24)))
165 acrn.24=boxcoxtransform(acrn.24,mean(c(lambda.12,lambda.24)))
166
167 y=c(acrn.12,acrn.24)
168
169 #DATA FRAME FOR CONTINUOUS MODEL
170 id=c(1:(length(y)/2),1:(length(y)/2))
171 y=y[order(id)]
172 X=data.frame(intercept=rep(1,length(y)),trt1=c(armnew,rep(0,length(y)/2)),trt2=c(rep(0,length(y)/2),armnew),
      baselinediseaseactivity=c(baselinediseaseactivity,baselinediseaseactivity),time=c(rep(1,length(y)/2),rep(2,
      length(y)/2)))
173 X=X[order(id),]
174 id=sort(id)
175
176 #CONTINUOUS MODEL
177 continuousmodel=gls(y~trt1+trt2+baselinediseaseactivity+time,correlation=corSymm(form=~1|id),weights=varIdent(form
      =~1|time),na.action=na.omit,data=X)
178
179 #BINARY MODEL 1
180 failuremodel1=glm(rescuemedicationupto12~baselinediseaseactivity+armnew,family="binomial")
181
182 #BINARY MODEL 2
183
184 rescuemedicationupto24=rescuemedicationupto24[rescuemedicationupto12==0]
```

```
interim=(baselinediseaseactivity*acrn.12)[rescuemedicationupto12==0]
armnew=armnew[rescuemedicationupto12==0]

failuremodel2=glm(rescuemedicationupto24~interim+armnew,family="binomial")


partials=get.partials(continuousmodel,baselinediseaseactivity[!is.na(baselinediseaseactivity)],failuremodel1,
     failuremodel2,((1-dichotomisationthreshold/100)^lambda-1)/lambda)
mean=partials[12]
partials=partials[1:11]
covariance=matrix(0,11,11)

covariance[1:5,1:5]=continuousmodel$varBeta
covariance[6:8,6:8]=summary(failuremodel1)$cov.unscaled
covariance[9:11,9:11]=summary(failuremodel2)$cov.unscaled

variance=t(partials)%*%covariance%*%partials

CI.augbin=c(mean-1.96*sqrt(variance),mean,mean+1.96*sqrt(variance))




##ADJUSTED

#For adjustments change failure model 1 and failure model 2 to:

failuremodel1=brglm(rescuemedicationupto12~baselinediseaseactivity+armnew,family="binomial")
failuremodel2=brglm(rescuemedicationupto24~interim+armnew,family="binomial")
```

D.2 Augmented Binary: GEE

```
library(geesmv)
library(gee)
library(gtools)
library(geepack)
library(gee)
library(nlme)
library(R2Cuba)
library(boot)
library(mvtnorm)
library(MASS)
library(brglm)



##AUGMENTED BINARY - GENERALISED ESTIMATING EQUATIONS (GEE)

##UNADJUSTED

integrand<-function(acrn,meantreated,meanuntreated,Sigma,failure1,failure2,baseline)
{
  n=length(baseline)

  f1treat=inv.logit(cbind(rep(1,n),baseline,rep(1,n))%*%failure1$coefficient)
  f1untreat=inv.logit(cbind(rep(1,n),baseline,rep(0,n))%*%failure1$coefficient)

  f2treat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(1,n))%*%failure2$coefficient)
  f2untreat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(0,n))%*%failure2$coefficient)

  pdftreat=dmvnorm(cbind(-meantreated[,1]+acrn[1],-meantreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(Sigma
       [1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))
  pdfuntreat=dmvnorm(cbind(-meanuntreated[,1]+acrn[1],-meanuntreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(
       Sigma[1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))

  return(c(mean((1-f1treat)*(1-f2treat)*pdftreat),mean((1-f1untreat)*(1-f2untreat)*pdfuntreat)))
}

```

```
35
36
37
38 probofsuccess=function(continuous,baseline,failure1,failure2,dich)
39 {
40
41   n=length(baseline)
42
43   meantreated=cbind(cbind(rep(1,n),rep(1,n),rep(0,n),baseline,rep(1,n))%*%continuous$coefficient,
44                     cbind(rep(1,n),rep(0,n),rep(1,n),baseline,rep(2,n))%*%continuous$coefficient)
45
46   meanuntreated=cbind(cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(1,n))%*%continuous$coefficient,
47                       cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(2,n))%*%continuous$coefficient)
48
49
50   #find lower and upper points for integration:
51   maxmean1=max(c(meantreated[,1],meanuntreated[,1]))
52   maxmean2=max(c(meantreated[,2],meanuntreated[,2]))
53   minmean1=min(c(meantreated[,1],meanuntreated[,1]))
54   minmean2=min(c(meantreated[,2],meanuntreated[,2]))
55
56
57   #integrate
58
59   a=cuhre(2,2,integrand=integrand,meantreated=meantreated,meanuntreated=meanuntreated,Sigma=Sigma1,failure1=
         failure1,failure2=failure2,baseline=baseline,lower=c(qnorm(1e-08,minmean1,sqrt(Sigma1[1,1])),qnorm(1e-08,
         minmean2,sqrt(Sigma1[2,2]))),
60          upper=c(qnorm(1-1e-08,maxmean1,sqrt(Sigma1[1,1])), dich),flags=list(verbose=0,final=1,pseudo.random=0,
                 mersenne.seed=NULL))
61
62
63   #return(a$value[1]-a$value[2])
64   return(log(a$value[1]/(1-a$value[1]))-log(a$value[2]/(1-a$value[2])))
65 }
66
67
68
69
70 get.partials<-function(continuous,baseline,failure1,failure2,dich)
71 {
72
73   fit1=probofsuccess(continuous,baseline,failure1,failure2,dich)
74   augbin.partials=as.vector(rep(0,11))
75
76   #continuous model
77
78   for(i in 1:5)
79   {
80
81     valueupdate1=continuous
82     valueupdate1$coefficient[i]=valueupdate1$coefficient[i]+0.000001
83
84     updateprob=probofsuccess(valueupdate1,baseline,failure1,failure2,dich)
85
86     augbin.partials[i]=(updateprob-fit1)/0.000001
87
88   }
89
90   #failure model1
91
92   for(i in 1:3)
93   {
94
95     valueupdate2=failure1
96     valueupdate2$coefficient[i]=valueupdate2$coefficient[i]+0.000001
97
98     updateprob=probofsuccess(continuous,baseline,valueupdate2,failure2,dich)
99
100    augbin.partials[i+5]=(updateprob-fit1)/0.000001
101  }
102
103  #failure model2
104
105  for(i in 1:3)
106  {
```

```
107
108     valueupdate3=failure2
109     valueupdate3$coefficient[i]=valueupdate3$coefficient[i]+0.000001
110
111     updateprob=probofsuccess(continuous,baseline,failure1,valueupdate3,dich)
112
113     augbin.partials[i+8]=(updateprob-fit1)/0.000001
114   }
115
116   return(c(augbin.partials,fit1))
117 }
118
119
120
121
122
123 ##################################################
124 #APPLY IN RHEUMATOID ARTHRITIS EXAMPLE FROM PAPER:#
125 ##################################################
126
127
128 dichotomisationthreshold=20 #change this depending on example
129 data=#OSKIRA-1 TRIAL
130
131 acrn.12<-data$acrn12 #CONTINUOUS MEASURE AT TIME POINT ONE
132 acrn.24<-data$acrn24 #CONTINUOUS MEASURE AT TIME POINT TWO
133 rescuemedicationupto12<-data$rescuemedicationupto12 #BINARY MEASURE AT TIME POINT ONE
134 rescuemedicationupto24<-data$rescuemedicationupto24 #BINARY MEASURE AT TIME POINT TWO
135 baselinediseaseactivity<-data$baselinediseaseactivity #BASELINE CONTINUOUS SCORE
136 arm<-data$arm #BINARY TREATMENT INDICATOR
137 patientid<-data$patientid #PATIENT NUMBER
138
139
140 #CHANGE TO RELATIVE ACRN AND BOXCOX TRANSFORM
141 acrn.12=1-acrn.12/100
142 acrn.24=1-acrn.24/100
143
144 lm1=lm(acrn.12~1)
145 lm2=lm(acrn.24~1)
146
147 temp=boxcox(lm1,plotit=F,lambda=seq(0,2,length=100))
148 lambda.12=temp$x[which.max(temp$y)]
149 temp=boxcox(lm2,plotit=F,lambda=seq(0,2,length=100))
150 lambda.24=temp$x[which.max(temp$y)]
151 lambda=mean(c(lambda.12,lambda.24))
152
153 acrn.12=boxcoxtransform(acrn.12,mean(c(lambda.12,lambda.24)))
154 acrn.24=boxcoxtransform(acrn.24,mean(c(lambda.12,lambda.24)))
155
156 y=c(acrn.12,acrn.24)
157
158 #DATA FRAME FOR CONTINUOUS MODEL
159
160 id=c(1:(length(y)/2),1:(length(y)/2))
161 y=y[order(id)]
162 X=data.frame(intercept=rep(1,length(y)),trt1=c(arm,rep(0,length(y)/2)),trt2=c(rep(0,length(y)/2),arm),
      baselinediseaseactivity=c(baselinediseaseactivity,baselinediseaseactivity),time=c(rep(1,length(y)/2),rep(2,
      length(y)/2)))
163 X=X[order(id),]
164 id=sort(id)
165
166 #CONTINUOUS MODEL
167 continuousmodel=gee(y~trt1+trt2+baselinediseaseactivity+time, data=X, id=id, corstr="exchangeable", na.action=na.
      omit)
168 Sigma1<-as.matrix(continuousmodel$scale*continuousmodel$working.correlation)
169
170 #BINARY MODEL 1
171 failuremodel1=glm(rescuemedicationupto12~baselinediseaseactivity+arm,family="binomial")
172
173 #BINARY MODEL 2
174
175 rescuemedicationupto24=rescuemedicationupto24[rescuemedicationupto12==0]
176 interim=(baselinediseaseactivity*acrn.12)[rescuemedicationupto12==0]
177 arm=arm[rescuemedicationupto12==0]
178
```

```
failuremodel2=glm(rescuemedicationupto24~interim+arm,family="binomial")



partials=get.partials(continuousmodel,baselinediseaseactivity[!is.na(baselinediseaseactivity)],failuremodel1,
     failuremodel2,((1-dichotomisationthreshold/100)^lambda-1)/lambda)

mean=partials[12]
partials=partials[1:11]
covariance=matrix(0,11,11)

covariance[1:5,1:5]=continuousmodel$robust.variance
covariance[6:8,6:8]=summary(failuremodel1)$cov.unscaled
covariance[9:11,9:11]=summary(failuremodel2)$cov.unscaled

variance=t(partials)%*%covariance%*%partials

CI.augbin=c(mean-1.96*sqrt(variance),mean,mean+1.96*sqrt(variance))









###ADJUSTED


GEE.mbn <- function (formula, id, data, corstr = "independence",
                     d = 2, r = 1)
{

  init <- model.frame(formula, data)
  init$num <- 1:length(init[, 1])
  m <- model.frame(formula, data)
  mt <- attr(m, "terms")
  data$response <- model.response(m, "numeric")
  mat <- as.data.frame(model.matrix(formula, m))
  gee.fit <- gee(formula, data = data, id = id, family = "gaussian",
                 corstr = corstr)
  beta_est <- gee.fit$coefficient
  alpha <- gee.fit$working.correlation[1, 2]
  scale <- summary(gee.fit)$scale
  len <- length(beta_est)
  len_vec <- len^2
  data$id <- gee.fit$id
  cluster <- cluster.size(data$id)
  ncluster <- max(cluster$n)
  size <- cluster$m
  mat$subj <- rep(unique(data$id), cluster$n)
  var <- switch(corstr, independence = cormax.ind(ncluster),
                exchangeable = cormax.exch(ncluster, alpha), `AR-M` = cormax.ar1(ncluster,
                                                                                 alpha), unstructured = summary(
                                                                                      gee.fit)$working.correlation
                                                                                      )


  cov.beta <- unstr <- matrix(0, nrow = len, ncol = len)
  step11 <- matrix(0, nrow = len, ncol = len)
  for (i in 1:size) {
    y <- as.matrix(data$response[data$id == unique(data$id)[i]])
    covariate <-  as.matrix(subset(mat[,1:5], data$id==i))
    ncluster = cluster$n[i]
    var1 = var[1:ncluster, 1:ncluster]
    Vi <- gee.fit$scale * var1
    xx <- t(covariate) %*% solve(Vi) %*% covariate
    step11 <- step11 + xx

  }
  k <- (sum(cluster$n) - 1)/(sum(cluster$n) - len) * size/(size -
                                                           1)
  delta <- ifelse(size > ((d + 1) * len), len/(size - len),
                  1/d)
  step00 <- matrix(0, nrow = len, ncol = len)
  for (i in 1:size) {
```

```
    y <- as.matrix(data$response[data$id == unique(data$id)[i]])
    ncluster = cluster$n[i]
    covariate <-  as.matrix(subset(mat[,1:5], data$id==i))
    var1 = var[1:ncluster, 1:ncluster]

    Vi <- gee.fit$scale * var1
    xy <- t(covariate) %*% solve(Vi) %*% (y - covariate %*%
                                          beta_est)
    step00 <- step00 + xy %*% t(xy)


  }
  tracexi<-max(diag(solve(step11) %*% step00))/len
  xi <- pmax(r, max(diag(solve(step11) %*% step00))/len)
  step12 <- matrix(0, nrow = len, ncol = len)
  step13 <- matrix(0, nrow = len_vec, ncol = 1)
  step14 <- matrix(0, nrow = len_vec, ncol = len_vec)
  p <- matrix(0, nrow = len_vec, ncol = size)
  for (i in 1:size) {
    y <- as.matrix(data$response[data$id == unique(data$id)[i]])
    covariate <-  as.matrix(subset(mat[,1:5], data$id==i))
    ncluster = cluster$n[i]
    var1 = var[1:ncluster, 1:ncluster]

    Vi <- gee.fit$scale * var1
    xy <- t(covariate) %*% solve(Vi) %*% (k * (y - covariate %*%
                                                 beta_est) %*% t(y - covariate %*% beta_est) +
                                          delta * xi * Vi) %*% solve(Vi) %*% covariate
    step12 <- step12 + xy
    step13 <- step13 + vec(xy)
    p[, i] <- vec(xy)
  }
  for (i in 1:size) {
    dif <- (p[, i] - step13/size) %*% t(p[, i] - step13/size)
    step14 <- step14 + dif
  }
  cov.beta <- solve(step11) %*% (step12) %*% solve(step11)
  cov.var <- size/(size - 1) * kronecker(solve(step11), solve(step11)) %*%
    step14 %*% kronecker(solve(step11), solve(step11))
  return(list(summary(gee.fit)$coefficients, summary(gee.fit)$scale, summary(gee.fit)$working.correlation,cov.beta
        = cov.beta))
}


integrand<-function(acrn,meantreated,meanuntreated,Sigma,failure1,failure2,baseline)
{
  n=length(baseline)

  f1treat=inv.logit(cbind(rep(1,n),baseline,rep(1,n))%*%failure1$coefficient)
  f1untreat=inv.logit(cbind(rep(1,n),baseline,rep(0,n))%*%failure1$coefficient)

  f2treat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(1,n))%*%failure2$coefficient)
  f2untreat=inv.logit(cbind(rep(1,n),acrn[1]*baseline,rep(0,n))%*%failure2$coefficient)

  pdftreat=dmvnorm(cbind(-meantreated[,1]+acrn[1],-meantreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(Sigma
       [1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))
  pdfuntreat=dmvnorm(cbind(-meanuntreated[,1]+acrn[1],-meanuntreated[,2]+acrn[2]), mean=c(0,0),sigma=matrix(c(
       Sigma[1,1],Sigma[1,2],Sigma[2,1],Sigma[2,2]),2,2))

  return(c(mean((1-f1treat)*(1-f2treat)*pdftreat),mean((1-f1untreat)*(1-f2untreat)*pdfuntreat)))
}




probofsuccess=function(continuous,baseline,failure1,failure2,dich)
{

  n=length(baseline)

  meantreated=cbind(cbind(rep(1,n),rep(1,n),rep(0,n),baseline,rep(1,n))%*%continuous[[1]][,1],
                    cbind(rep(1,n),rep(0,n),rep(1,n),baseline,rep(2,n))%*%continuous[[1]][,1])

  meanuntreated=cbind(cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(1,n))%*%continuous[[1]][,1],
                      cbind(rep(1,n),rep(0,n),rep(0,n),baseline,rep(2,n))%*%continuous[[1]][,1])
```

```
323
324
325   #find lower and upper points for integration:
326   maxmean1=max(c(meantreated[,1],meanuntreated[,1]))
327   maxmean2=max(c(meantreated[,2],meanuntreated[,2]))
328   minmean1=min(c(meantreated[,1],meanuntreated[,1]))
329   minmean2=min(c(meantreated[,2],meanuntreated[,2]))
330
331
332   #integrate
333
334   a=cuhre(2,2,integrand=integrand,meantreated=meantreated,meanuntreated=meanuntreated,Sigma=Sigma1,failure1=
          failure1,failure2=failure2,baseline=baseline,lower=c(qnorm(1e-08,minmean1,sqrt(Sigma1[1,1])),qnorm(1e-08,
          minmean2,sqrt(Sigma1[2,2]))),
335          upper=c(qnorm(1-1e-08,maxmean1,sqrt(Sigma1[1,1])), dich),flags=list(verbose=0,final=1,pseudo.random=0,
                mersenne.seed=NULL))
336
337
338   return(a$value[1]-a$value[2])
339   #return(log(a$value[1]/(1-a$value[1]))-log(a$value[2]/(1-a$value[2])))
340 }
341
342
343
344
345 get.partials<-function(continuous,baseline,failure1,failure2,dich)
346 {
347
348   fit1=probofsuccess(continuous,baseline,failure1,failure2,dich)
349   augbin.partials=as.vector(rep(0,11))
350
351
352   #split in to three separate models
353
354   #continuous model
355
356   for(i in 1:5)
357   {
358
359     valueupdate1=continuous
360     valueupdate1[[1]][i,1]=valueupdate1[[1]][i,1]+0.000001
361
362     updateprob=probofsuccess(valueupdate1,baseline,failure1,failure2,dich)
363
364     augbin.partials[i]=(updateprob-fit1)/0.000001
365
366   }
367
368   #failure model1
369
370   for(i in 1:3)
371   {
372
373     valueupdate2=failure1
374     valueupdate2$coefficient[i]=valueupdate2$coefficient[i]+0.000001
375
376     updateprob=probofsuccess(continuous,baseline,valueupdate2,failure2,dich)
377
378     augbin.partials[i+5]=(updateprob-fit1)/0.000001
379   }
380
381   #failure model2
382
383   for(i in 1:3)
384   {
385
386     valueupdate3=failure2
387     valueupdate3$coefficient[i]=valueupdate3$coefficient[i]+0.000001
388
389     updateprob=probofsuccess(continuous,baseline,failure1,valueupdate3,dich)
390
391     augbin.partials[i+8]=(updateprob-fit1)/0.000001
392   }
393
394   return(c(augbin.partials,fit1))
```

```
}


################################################
#APPLY IN RHEUMATOID ARTHRITIS EXAMPLE FROM PAPER:#
################################################

dichotomisationthreshold=20 #change this depending on example
data=#OSKIRA-1 TRIAL

  acrn.12<-data$acrn12 #CONTINUOUS MEASURE AT TIME POINT ONE
acrn.24<-data$acrn24 #CONTINUOUS MEASURE AT TIME POINT TWO
rescuemedicationupto12<-data$rescuemedicationupto12 #BINARY MEASURE AT TIME POINT ONE
rescuemedicationupto24<-data$rescuemedicationupto24 #BINARY MEASURE AT TIME POINT TWO
baselinediseaseactivity<-data$baselinediseaseactivity #BASELINE CONTINUOUS SCORE
arm<-data$arm #BINARY TREATMENT INDICATOR
patientid<-data$patientid #PATIENT NUMBER


#CHANGE TO RELATIVE ACRN AND BOXCOX TRANSFORM
acrn.12=1-acrn.12/100
acrn.24=1-acrn.24/100

lm1=lm(acrn.12~1)
lm2=lm(acrn.24~1)

temp=boxcox(lm1,plotit=F,lambda=seq(0,2,length=100))
lambda.12=temp$x[which.max(temp$y)]
temp=boxcox(lm2,plotit=F,lambda=seq(0,2,length=100))
lambda.24=temp$x[which.max(temp$y)]
lambda=mean(c(lambda.12,lambda.24))

acrn.12=boxcoxtransform(acrn.12,mean(c(lambda.12,lambda.24)))
acrn.24=boxcoxtransform(acrn.24,mean(c(lambda.12,lambda.24)))

y=c(acrn.12,acrn.24)

#DATA FRAME FOR CONTINUOUS MODEL
id=c(1:(length(y)/2),1:(length(y)/2))
y=y[order(id)]
X=data.frame(intercept=rep(1,length(y)),trt1=c(arm,rep(0,length(y)/2)),trt2=c(rep(0,length(y)/2),arm),
      baselinediseaseactivity=c(baselinediseaseactivity,baselinediseaseactivity),time=c(rep(1,length(y)/2),rep(2,
      length(y)/2)))
X=X[order(id),]
id=sort(id)
X$y<-y
X$id<-id
X<-na.omit(X)
Xnew<-transform(X, id=match(X$id, unique(X$id)))

#CONTINUOUS MODEL
continuousmodel=GEE.mbn(y~trt1+trt2+baselinediseaseactivity+time, data=Xnew, id=Xnew$id, corstr="exchangeable")
Sigma1<-as.matrix(continuousmodel[[2]]*continuousmodel[[3]])

#BINARY MODEL 1
failuremodel1=brglm(rescuemedicationupto12~baselinediseaseactivity+arm,family="binomial")

#BINARY MODEL 2

rescuemedicationupto24=rescuemedicationupto24[rescuemedicationupto12==0]
interim=(baselinediseaseactivity*acrn.12)[rescuemedicationupto12==0]
arm=arm[rescuemedicationupto12==0]
failuremodel2=brglm(rescuemedicationupto24~interim+arm,family="binomial")



partials=get.partials(continuousmodel,baselinediseaseactivity[!is.na(baselinediseaseactivity)],failuremodel1,
      failuremodel2,((1-dichotomisationthreshold/100)^lambda-1)/lambda)

mean=partials[12]
partials=partials[1:11]

covariance=matrix(0,11,11)

covariance[1:5,1:5]=continuousmodel$cov.beta
```

```
covariance[6:8,6:8]=summary(failuremodel1)$cov.unscaled
covariance[9:11,9:11]=summary(failuremodel2)$cov.unscaled

variance=t(partials)%*%covariance%*%partials

CI.augbin=c(mean-1.96*sqrt(variance),mean,mean+1.96*sqrt(variance))
```

D.3 Standard Binary

```
library(gtools)
library(geepack)
library(nlme)
library(R2Cuba)
library(boot)
library(mvtnorm)
library(MASS)
library(brglm)
library(gee)


##STANDARD BINARY METHOD
#UNADJUSTED


boxcoxtransform=function(y,lambda)
{
  return((y^lambda-1)/lambda)
}



differenceinprob.binary=function(glm1,t,x)
{
  #get fitted probs for each arm from model:

  fittedvalues.control=as.double(inv.logit(cbind(rep(1,length(t[t==0])),rep(0,length(t[t==0])),x[t==0])%*%glm1$
      coef))

  fittedvalues.exp=as.double(inv.logit(cbind(rep(1,length(t[t==1])),rep(1,length(t[t==1])),x[t==1])%*%glm1$coef))


  return(log(mean(fittedvalues.exp,na.rm=T)/(1-mean(fittedvalues.exp,na.rm=T)))-log(mean(fittedvalues.control,na.
      rm=T)/(1-mean(fittedvalues.control,na.rm=T))))
  #return(mean(fittedvalues.exp,na.rm=T)-mean(fittedvalues.control,na.rm=T))

}


partialderivatives.binary=function(glm1,t,x)
{


  value=differenceinprob.binary(glm1,t,x)

  partials=rep(0,3)

  tempglm1=glm1
  tempglm1$coef[1]=tempglm1$coef[1]+0.00001

  partials[1]=(differenceinprob.binary(tempglm1,t,x)-value)/0.00001

  tempglm1=glm1
  tempglm1$coef[2]=tempglm1$coef[2]+0.00001

  partials[2]=(differenceinprob.binary(tempglm1,t,x)-value)/0.00001

  tempglm1=glm1
  tempglm1$coef[3]=tempglm1$coef[3]+0.00001

```

```
  partials[3]=(differenceinprob.binary(tempglm1,t,x)-value)/0.00001

  return(c(value,partials))

}


#####################################################
#APPLY IN RHEUMATOID ARTHRITIS EXAMPLE FROM PAPER:#
#####################################################


#UNADJUSTED

dichotomisationthreshold=20 #change this depending on example
data=#OSKIRA-1 TRIAL

  acrn.12<-data$acrn12 #CONTINUOUS MEASURE AT TIME POINT ONE
acrn.24<-data$acrn24 #CONTINUOUS MEASURE AT TIME POINT TWO
rescuemedicationupto12<-data$rescuemedicationupto12 #BINARY MEASURE AT TIME POINT ONE
rescuemedicationupto24<-data$rescuemedicationupto24 #BINARY MEASURE AT TIME POINT TWO
baselinediseaseactivity<-data$baselinediseaseactivity #BASELINE CONTINUOUS SCORE
arm<-data$arm #BINARY TREATMENT INDICATOR
patientid<-data$patientid #PATIENT NUMBER

success.binary=ifelse(acrn.24>dichotomisationthreshold & rescuemedicationupto24==0,1,0)

glm1=glm(success.binary~armnew+baselinediseaseactivity,family="binomial") #CHANGE THIS FOR ADJUSTED METHOD

partials.binary=partialderivatives.binary(glm1,armnew,baselinediseaseactivity)
mean.binary=partials.binary[1]
partials.binary=partials.binary[-1]

covariance=summary(glm1)$cov.unscaled

var.binary=t(partials.binary)%*%covariance%*%partials.binary


CI.binary=c(mean.binary-1.96*sqrt(var.binary),mean.binary,mean.binary+1.96*sqrt(var.binary))


#ADJUSTED

#CHANGE GLM1 ABOVE TO:
glm1=glm(success.binary~armnew+baselinediseaseactivity,family="binomial")
```

Appendix E

Preprint: Complex Composite Structures

Employing latent variable models to improve efficiency in composite endpoint analysis

Martina McMenamin*[1], Jessica K. Barrett[1], Anna Berglind[2], James M.S. Wason[1,3]

[1] MRC Biostatistics Unit, School of Clinical Medicine, Cambridge Institute of Public Health Forvie Site, Robinson Way, Cambridge Biomedical Campus Cambridge, UK

[2] Global Medicines Development, Biometrics and Information Sciences, AstraZeneca, Gothenburg, Sweden

[3] Institute of Health and Society, Newcastle University, Newcastle, UK

Summary

Composite endpoints that combine multiple outcomes on different scales are common in clinical trials, particularly in chronic conditions. In many of these cases, patients will have to cross a predefined responder threshold in each of the outcomes to be classed as a responder overall. One instance of this occurs in systemic lupus erythematosus (SLE), where the responder endpoint combines two continuous, one ordinal and one binary measure. The overall binary responder endpoint is typically analysed using logistic regression, resulting in a substantial loss of information. We propose a latent variable model for the SLE endpoint, which assumes that the discrete outcomes are manifestations of latent continuous measures and can proceed to jointly model the components of the composite. We perform a simulation study and find the method to offer large efficiency gains over the standard analysis. We find that the magnitude of the precision gains

*To whom correspondence should be addressed martina.mcmenamin@mrc-bsu.cam.ac.uk

are highly dependent on which components are driving response. Bias is introduced when joint normality assumptions are not satisfied, which we correct for using a bootstrap procedure. The method is applied to the Phase IIb MUSE trial in patients with moderate to severe SLE. We show that it estimates the treatment effect 2.5 times more precisely, offering a 60% reduction in required sample size.

Key words: Latent variable models; Composite endpoints; Responder analysis; Systemic lupus erythematosus

1. Introduction

Composite endpoints combine multiple outcomes in order to determine the effectiveness or efficacy of a treatment for a given disease. They are typically recommended when a disease is complex or multi-system and meaningful improvement cannot be captured in a single outcome. Furthermore, the endpoint may be a combination of continuous and discrete outcomes which are collapsed in to a single binary responder index.

Table 1 shows examples of diseases that use composite endpoints combining multiple continuous and discrete components. Responders in fibromyalgia must respond in two continuous and one ordinal component however responders in trials for frailty or soft tissue infections must respond in a total of five continuous and discrete components. Generally, these composite responder endpoints will be treated as a single binary outcome and analysed using a logistic regression model, which we term the standard binary method. This solves problems with multiplicity however results in large losses in efficiency (Wason and Seaman (2013)). The aim of this paper is to propose a joint modelling framework within which we can model the components of the composite, retaining the information on the original scales of the outcomes, hence increasing efficiency. One likelihood based method for handling mixed data is the factorisation model. The objective is to factorise the

Table 1. Examples of diseases that use complex composite endpoints combining multiple discrete and continuous measures to determine effectiveness of a treatment including criteria for response and how each component is measured

Disease	Responder endpoint	Measured by
Fibromyalgia	• achieved a 30% improvement in pain	Electronic diary
	• 30% improvement in functional status	Subscale of Fibromyalgia Impact Questionnaire (FIQ)
	• improved, much improved, or very much improved	7-point Patient Global Impression of Change (PGIC) scale
Frailty	• BMI<18.5 kg/m2 OR >10% weight loss since last wave	weight and height
	• One positive answer to exhaustion questions	CES-D questionnaire
	• Low grip strength (M $<$ 31.12 kg, F $<$ 17.60 kg)	Eg. Jamar hand dynamometer
	• Gait speed (M $<$ 0.691 m/s, F $<$ 0.619 m/s)	Distance/time
	• Low activity (M $<$ 16.5 activity units, F $<$ 13.5 activity units)	Activity units derived using intensity vs. frequency
Necrotizing Soft Tissue Infections	• Alive until day 28	yes/no
	• Day 14 debridements $\leqslant$ 3	surface area
	• No amputation if debridement	yes/no
	• Day 14 mSOFA score $\leqslant$ 1	mSOFA score - composite additively combining scores in different systems
	• Reduction of at least 3 score points in mSOFA score	mSOFA score - composite additively combining scores in different systems
Systemic lupus erythematosus	• Change in SLEDAI $\leqslant$ -4	SLE Disease Activity Index
	• Change in PGA $<$ 0.3	Physicians Global Assessment
	• No Grade A or more than one Grade B in BILAG	British Isles Lupus Assessment Group
	• Reduction in oral corticosteroids	Notes

joint distribution and fit a univariate model to each component of the factorisation (de Leon and Carriere (2013)). This accounts for correlations between the outcomes by including one response as a covariate in the model for the other response. In the graphical modelling literature this has been termed the 'Conditional Gaussian Distribution' (Whittaker (1990); Lauritzen and Wermuth

(1989)). An advantage of these methods in relation to the composite endpoint problem is that we may account for correlations between measurements whilst making inference directly on the outcomes that we have measured, hence they fall within a broader class of 'direct methods'. Examples of applications of these ideas, which build on the work of Olkin and Tate (1961), include developmental toxicity studies by Fitzmaurice and Laird (1995) and in the longitudinal setting, the augmented binary method was developed for application to composite endpoints in clinical trials where the composite is formed of one continuous and one binary outcome (Wason and Seaman (2013); Wason and Jenkins (2016); McMenamin *and others* (2018)). One difficulty with these methods beyond the bivariate scenario is the range of possibilities for the factorisations, with no consensus on how this should be determined. In the case of the SLE responder endpoint with four components, this amounts to 24 possible factorisations, each of which may result in different conclusions (Verbeke *and others* (2014); de Leon and Carriere (2013)).

Another family of models used to model mixed outcome types which feature frequently in economics and finance are copulas. These are functions that join or couple multivariate distribution functions to their uniform one-dimensional marginal distribution functions as discussed by Nelsen (1999). Copulas offer a flexible framework in this setting, as the marginal distribution functions need not come from the same parametric family. While the construction of copulas is considered to be mathematically elegant and the flexibility with which we can model appealing, they are not without their shortcomings. Extensions beyond the bivariate setting are difficult and have failed to perform well in many applications (de Leon and Carriere (2013)). Other practical implications include poor out-of-sample predictions due to the wide variety of copulas available. These restrictions, along with difficulties in longitudinal settings with unbalanced data structures, have seen few applications of copulas for mixed outcome types in the medical statistics literature (Verbeke *and others* (2014)). Applications of copulas in mixed outcome settings include de Leon and Wu (2010) and Wu and de Leon. A.R. (2014).

Another likelihood based method that allows for more flexibility when modelling the correlations between outcomes falls within the framework of latent variable models (Skrondal and Rabe-Hesketh (2004)). The multiple outcomes are assumed to be physical manifestations of some underlying latent process. This is modelled by including the same latent variable in each of the models for the observed responses. The outcomes are then assumed to be independent conditional on this latent variable. This solves the problem of deciding the order of factorisations in previously discussed methods however this formulation results in the inclusion of some covariance parameters in the mean structure, leaving the model sensitive to misspecification of the correlation structure (Sammel and Ryan (2002)). One example of these models is introduced by Sammel *and others* (1997), where effects of covariates of interest are modelled through this shared latent variable. Although these models have the intuitive interpretation that each outcome is attempting to capture underlying disease activity, the correlation matrix is restricted to allow for the same correlation between each pair of outcomes, which is unlikely in practice. This structure is relaxed in work by Dunson (2000), where the effects of covariates are included in the model separately from the latent variable. The correlation structure can be further relaxed to allow for a different latent variable for each outcome, meaning that pairs of outcomes are not assumed to have the same correlation. However these models would require integrating out each of the latent variables in order to obtain the joint distribution of interest (McCulloch (2008)). Furthermore, they are relevant in applications with multiple time points however less so for a single time point, as is the case for the composite endpoint problem.

Latent variables have also been used in the setting of mixed continuous and discrete variables to a different end. Namely, the outcomes adopt a correlated Gaussian distribution by assuming that the discrete outcomes are coarsely measured manifestations of underlying continuous variables subject to some threshold specifications, as seen in Ashford and Sowden (1970) and Chib and Greenberg (1998). Specifying discrete variables in terms of a partitioning of the latent variable

space into non-overlapping intervals dates back to Pearson (1904) in relation to his generalised theory of alternative inheritance and has received much consideration in the literature since. Terminology surrounding these models is inconsistent but they are often referred to as multivariate probit models (Ashford and Sowden (1970)). This theory can also be found to underpin conditional grouped continuous models (Poon and Lee (1987)). By formulating the distribution in this way, we can correlate the error terms between models and work within the familiar paradigm of Gaussian distributions and maximum likelihood theory. The theory and application of these ideas for a mixture of continuous and binary outcomes has featured in the statistics literature, see for example work by Tate (1955), Cox and Wermuth (1992) and Catalano and Ryan (1992). Generalisations of these ideas, which appear less frequently in the literature, lead to methods for modelling continuous and ordinal variables, with applications in developmental toxicology (Catalano (1997); Samani and Ganjali (2008); Arminger and Kusters (1988); Regan and Catalano (2000); Faes *and others* (2002)). Despite the advantages, the multivariate probit model has not realised its full potential in the applied biostatistics literature. This was noted by Lessafre and Molenberghs (1991) and we believe it still to be the case today. The few applications that do appear tend to demonstrate bivariate scenarios or those that mix continuous and binary or continuous and ordinal, rather than all three. Furthermore, these models have not been considered specifically to address challenges with modelling composite endpoints. Other work has combined thresholding the response variables and introducing latent variables in the model, examples include work by Gueorguieva and Agresti (2001), de Leon and Carriere (2013), Gueorguieva and Sanacora (2003) and Gueorguieva and Sanacora (2006), however these ideas are most applicable in the longitudinal setting. We will therefore employ the latent outcome framework and investigate its use for the composite endpoint problem.

The paper proceeds as follows. In Section 2 we discuss systemic lupus erythematosus (SLE), the motivating example for the methods. In Section 3 we introduce the latent variable model and

discuss how we conduct estimation and inference. In Section 4 we compare the behaviour of the latent variable model with the augmented binary and standard binary methods, including the case when the key assumptions are not satisfied. In Section 5 we apply the methods to the Phase IIb MUSE trial in patients with moderate to severe SLE. Finally, in Section 6 we discuss our findings and make some recommendations.

2. Motivating example

In what follows we will focus specifically on systemic lupus erythematosus (SLE), however the methods introduced will be relevant to other diseases using endpoints with a similar structure. The SLE endpoint is shown in Figure 1. It combines a continuous PGA measure, a continuous SLEDAI measure, an ordinal BILAG measure and a binary corticosteroids measure, where patients must meet the response criteria in all components in order to be classed as a responder overall. Note that the SLEDAI and BILAG measures are themselves composite scores deriving from a combination of items, however this will not be considered in the analysis.

The real data underpinning this motivation comes from the MUSE study (Furie *and others* (2017)). It was a Phase IIb, randomised, double-blind, placebo-controlled study investigating the efficacy and safety of anifrolumab in adults with moderate to severe SLE. Patients (n=305) were randomised to receive anifrolumab (300mg or 1000mg) or placebo, in addition to standard therapy every 4 weeks for 48 weeks. The primary end point was the percentage of patients achieving an SRI response at week 24 with sustained reduction of oral corticosteroids (<10mg/day and less than or equal to the dose at week 1 from week 12 through 24). The methods discussed will make inference at one time point, as this is the case in the trial, although they can be easily extended for the longitudinal case.

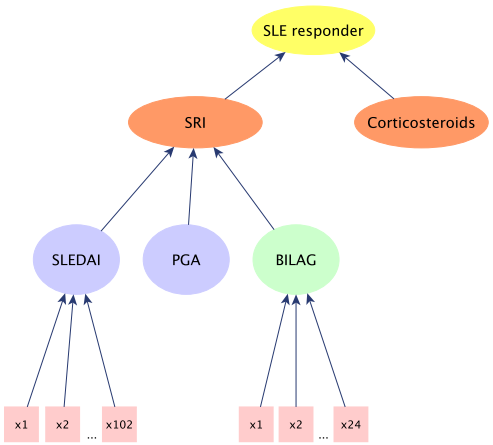


Fig. 1. Structure of the composite endpoint use in trials of systemic lupus erythematosus. The continuous SLEDAI, continuous PGA and ordinal BILAG measures are dichotomised and combined to form the binary SRI indicator which is then combined with the binary taper variable to form the overall binary SLE responder index

3. Methods

3.1 *Notation*

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}^*, Y_{i4}^*)$ represent the vector of observed and latent continuous measures for patient i. Y_{i1} and Y_{i2} are the observed continuous SLEDAI and PGA measures. Let Y_{i3} denote BILAG, the observed ordinal manifestation of Y_{i3}^* and Y_{i4} the observed binary taper variable for Y_{i4}^*. T_i represents the treatment indicator for patient i, y_{i10} and y_{i20} are the baseline measures for Y_{i1} and Y_{i2} respectively.

3.2 *Model*

The mean structure for the outcomes is shown in (3.1). The baseline measures y_{10} and y_{20} are included in the model for Y_1 and Y_2 respectively.

$$
\begin{aligned}
Y_{i1} &= \alpha_0 + \alpha_1 T_i + \alpha_2 y_{i10} + \varepsilon_{i1} \\
Y_{i2} &= \beta_0 + \beta_1 T_i + \beta_2 y_{i20} + \varepsilon_{i2} \\
Y^*_{i3} &= \gamma_1 T_i + \varepsilon^*_{i3} \\
Y^*_{i4} &= \psi_0 + \psi_1 T_i + \varepsilon^*_{i4}
\end{aligned}
\tag{3.1}
$$

The observed discrete variables are related to the latent continuous variables by partitioning the latent variable space, as shown in (3.2). The lower and upper thresholds for both discrete variables are set at $\tau_{03} = \tau_{04} = -\infty, \tau_{53} = \tau_{24} = \infty$. The intercept term for the ordinal variable in (3.1) is set at $\gamma_0 = 0$ so that the cut-points $\tau_{13}, \tau_{23}, \tau_{33}, \tau_{43}$ may be estimated. The intercept for the binary outcome ψ_0 may be estimated, as $\tau_{14} = 0$.

$$
Y_{i3} = \begin{cases}
\text{Grade E} & \text{if } \tau_{03} \leqslant Y^*_{i3} < \tau_{13}, \\
\text{Grade D} & \text{if } \tau_{13} \leqslant Y^*_{i3} < \tau_{23}, \\
\text{Grade C} & \text{if } \tau_{23} \leqslant Y^*_{i3} < \tau_{33}, \\
\text{Grade B} & \text{if } \tau_{33} \leqslant Y^*_{i3} < \tau_{43}, \\
\text{Non-responder} & \text{if } \tau_{43} \leqslant Y^*_{i3} < \tau_{53}
\end{cases}
\qquad
Y_{i4} = \begin{cases}
0, & \text{if } \tau_{04} \leqslant Y^*_{i4} < \tau_{14}, \\
1, & \text{if } \tau_{14} \leqslant Y^*_{i4} < \tau_{24}
\end{cases}
\tag{3.2}
$$

Following these assumptions, we can model the error terms in (3.1) as multivariate normal with zero mean and variance-covariance matrix Σ, as shown in (3.3). Note that the error variances for $\varepsilon^*_3, \varepsilon^*_4$ are $\sigma_3 = 1$ and $\sigma_4 = 1$. This does not represent a constraint on the model but rather a rescaling required for identifiability.

$$
(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon^*_{i3}, \varepsilon^*_{i4}) \sim N(0, \Sigma)
\qquad
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 & \rho_{14}\sigma_1 \\
\rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 & \rho_{24}\sigma_2 \\
\rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 & \rho_{34} \\
\rho_{14}\sigma_1 & \rho_{24}\sigma_2 & \rho_{34} & 1
\end{pmatrix}
\tag{3.3}
$$

Subsequently, we may factorise the joint likelihood contribution for patient i as shown below.

$$l(\boldsymbol{\theta};\mathbf{Y}) = f(Y_{i1}, Y_{i2};\boldsymbol{\theta}) f(Y_{i3}^*, Y_{i4}^* | Y_{i1}, Y_{i2};\boldsymbol{\theta}) \tag{3.4}$$

where $\boldsymbol{\theta}$ is a vector which contains all model parameters. The observed likelihood can then be expressed as in (3.5).

$$l(\boldsymbol{\theta};\mathbf{Y}) = \prod_{i=1}^{n} \prod_{w=1}^{5} \prod_{k=1}^{2} f(Y_{i1}, Y_{i2};\boldsymbol{\theta}) \left[pr\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = y_{i1}, Y_{i2} = y_{i2};\boldsymbol{\theta}\right)\right]^{I\{Y_{i3}=w, Y_{i4}=k\}} \tag{3.5}$$

The joint probability of patients having discrete measurements $Y_{i3} = w$ and $Y_{i4} = k$ must be multiplied over the five ordinal levels and two binary levels resulting in ten combinations of the probabilities in (3.6) to be calculated. We discuss the intuition for (3.6) in Appendix A.

$$\begin{aligned} &pr\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = Y_{i1}, Y_{i2} = Y_{i2};\boldsymbol{\theta}\right) = \\ &\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \\ &\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) \end{aligned} \tag{3.6}$$

where Φ_2 is the bivariate standard normal distribution function and $\mu_{3|1,2}, \mu_{4|1,2}$ and $\Sigma_{3,4|1,2}$ are derived using the rules of conditional multivariate normality, resulting in (3.7).

$$\begin{aligned} \mu_{3|1,2} =& \mu_3 + \tfrac{(\rho_{13} - \rho_{12}\rho_{23})}{\sigma_1\left(1-\rho_{12}^2\right)}\left(Y_{i1} - \mu_1\right) + \tfrac{(\rho_{23} - \rho_{12}\rho_{13})}{\sigma_2\left(1-\rho_{12}^2\right)}\left(Y_{i2} - \mu_2\right) \\ \mu_{4|1,2} =& \mu_4 + \tfrac{(\rho_{14} - \rho_{12}\rho_{24})}{\sigma_1\left(1-\rho_{12}^2\right)}\left(Y_{i1} - \mu_1\right) + \tfrac{(\rho_{24} - \rho_{12}\rho_{14})}{\sigma_2\left(1-\rho_{12}^2\right)}\left(Y_{i2} - \mu_2\right) \end{aligned} \tag{3.7}$$

$$\Sigma_{3,4|1,2} = \begin{pmatrix} 1 - \frac{\rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23} + \rho_{23}^2}{1-\rho_{12}^2} & \rho_{34} - \frac{\rho_{13}\rho_{14} - \rho_{12}\rho_{13}\rho_{24} - \rho_{12}\rho_{14}\rho_{23} + \rho_{23}\rho_{24}}{1-\rho_{12}^2} \\ \rho_{34} - \frac{\rho_{13}\rho_{14} - \rho_{12}\rho_{13}\rho_{24} - \rho_{12}\rho_{14}\rho_{23} + \rho_{23}\rho_{24}}{1-\rho_{12}^2} & 1 - \frac{\rho_{14}^2 - 2\rho_{12}\rho_{14}\rho_{24} + \rho_{24}^2}{1-\rho_{12}^2} \end{pmatrix}$$

3.3 *Estimation*

As the variance parameters (σ_1, σ_2) are required to be greater than 0, we introduce parameters (δ_1, δ_2) such that $\sigma_1 = exp(\delta_1)$ and $\sigma_2 = exp(\delta_2)$. This transformation ensures that the variance is above 0 whilst allowing the parameter we estimate to take any real value. We must also ensure that the correlation parameters $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$ are estimated within (-1,1) by introducing $(\delta_{12}, \delta_{13}, \delta_{14}, \delta_{23}, \delta_{24}, \delta_{34})$, where $\rho_{12} = 2expit(\delta_{12}) - 1, \rho_{13} = 2expit(\delta_{13}) - 1, \rho_{14} = 2expit(\delta_{14}) - 1, \rho_{23} = 2expit(\delta_{23}) - 1, \rho_{24} = 2expit(\delta_{24}) - 1, \rho_{34} = 2expit(\delta_{34}) - 1$.

We fit the model in R by coding the likelihood function, probability of response and using the delta method to obtain standard errors. The bivariate distribution functions in (3.6) are estimated using 'pmvnorm', using the method of Genz (1992). The likelihood maximisation is conducted using the 'nlminb' function in the 'optimx' package, which is the best performing method in terms of accuracy and convergence rate, however is the slowest. We use the 'Hessian' function in the 'numDeriv' package to obtain the Hessian matrix and invert this to get the covariance matrix of the model parameters. In a small number of cases the Hessian is not positive definite because of computational error, meaning that it cannot be inverted. This is rectified in these cases by using the 'near PD' function in the 'Matrix' package, which computes the nearest positive definite matrix.

3.4 *Inference*

We wish to make inference on the probability of response. Let S_i be an indicator for patient i denoting whether or not they achieved response defined by S_i=1 if $Y_{i1} \leqslant \theta_1, Y_{i2} \leqslant \theta_2, Y^*_{i3} \leqslant \theta_3, Y^*_{i4} \leqslant \theta_4$. Therefore,

$$P(S_i = 1 \mid T_i, y_{i10}, y_{i20}) = \int_{-\infty}^{\theta_1} \int_{-\infty}^{\theta_2} \int_{-\infty}^{\theta_3} \int_{-\infty}^{\theta_4} f_{\mathbf{Y}}(\mathbf{Y}; T_i, y_{i10}, y_{i20}) dy^*_{i4} dy^*_{i3} dy_{i2} dy_{i1} \tag{3.8}$$

We obtain the integrand in (3.8) by using the fitted values of the parameters in the conditional

mean and conditional covariance matrix in (3.7). Parameter estimates from these methods are maximum likelihood estimates and so we avail of asymptotic maximum likelihood theory. The integral in (3.8) is evaluated using the 'R2Cuba' package to obtain estimates for each patient, assuming they were treated $\tilde{p_{i1}}$ and not treated $\tilde{p_{i0}}$. The odds ratio treatment effect is then defined as shown in (3.9).

$$\tilde{\delta} = \frac{\left(\frac{\sum_{i=1}^{N} \tilde{p_{i1}}}{N - \sum_{i=1}^{N} \tilde{p_{i1}}}\right)}{\left(\frac{\sum_{i=1}^{N} \tilde{p_{i0}}}{N - \sum_{i=1}^{N} \tilde{p_{i0}}}\right)} \tag{3.9}$$

Note that we can easily define a risk difference or risk ratio using these quantities but in what follows we consider $\tilde{\delta}$ to be the effect of interest. The standard error estimates are obtained using the delta method. This requires the covariance matrix of the maximum likelihood estimates $\text{Cov}(\hat{\boldsymbol{\theta}})$ and $''\tilde{\boldsymbol{\delta}}$, the vector of partial derivatives of $\tilde{\delta}$ with respect to each of the parameter estimates. The variance of $\tilde{\delta}$ is obtained as shown in (3.10).

$$Var(\tilde{\delta}) = (''\tilde{\boldsymbol{\delta}})^T Cov(\hat{\boldsymbol{\theta}})(''\tilde{\boldsymbol{\delta}}) \tag{3.10}$$

Alternatively, the quantity in (3.8) can also be considered to be a multivariate Gaussian hidden truncation distribution, from which we can obtain a closed form solution, and proceed as detailed by Arnold (2009).

Another important consideration for the model is how to assess goodness-of-fit. We propose an extension to an existing method for application in this case, which is detailed in Appendix B in the supplementary material.

4. Simulation study

We are interested in comparing the performance of the latent variable, augmented binary and standard binary methods through simulation. The models for the augmented binary and standard

binary methods are included in Appendix C in the supplementary material.

4.1 *Data generating model*

Initially, we investigate the properties of the methods when the assumptions of the latent variable model are satisfied. The parameter values in the 'baseline' scenario are chosen to simulate a scenario where composite endpoints are typically recommended for use. Namely, that all four components drive response and items are correlated but not so highly that the composite becomes redundant. The parameter values have been informed by the MUSE trial dataset, in particular the correlation structure. The response probability in the control arm is 0.275 and in the treatment arm is 0.381, resulting in an odds ratio equal to 1.6, values typically observed in trials requiring response in all four components. The parameter values selected for the model in (3.1) are shown in Table 2. From this baseline case, we vary parameters to determine how the methods behave under various scenarios of interest. In particular, under varying treatment effect, varying responder threshold and varying drivers of response. The parameter values for these data generating models are included in Appendix D in the supplementary material.

Table 2. Parameter values for the data generating model in the baseline simulation scenario comparing the performance of the latent variable, augmented binary and standard binary methods for analysing a composite endpoint, where the values correspond to a treatment effect in all components and all components drive response

Purpose	Values
Total sample size	N=300
Intercept	$\alpha_0 = -4.9, \beta_0 = -1.2, \psi_0 = -0.2$
Treatment	$\alpha_1 = -0.28, \beta_1 = -0.35, \gamma_1 = -0.24, \psi_1 = -0.18$
Baseline value	$\alpha_2 = -0.5, \beta_2 = -0.5$
Variance	$\sigma_1 = \sigma_2 = 1$
Correlation	$\rho_{12} = 0.5, \rho_{13} = \rho_{24} = 0.35, \rho_{14} = 0.25, \rho_{23} = 0.4, \rho_{34} = 0.3$
Discrete cut-point	$\tau_{13} = -1, \tau_{23} = -0.1, \tau_{33} = 0.45, \tau_{43} = 1.3$
Responder threshold	$\theta_1 = -4, \theta_2 = -0.6, \theta_3 = 0.45, \theta_4 = 0$

4.2 *Results*

The methods are evaluated against a range of performance criteria, which are included with their Monte Carlo standard errors in Appendix E of the supplementary material. For further details see Morris *and others* (2017).

4.2.1 *Varying treatment effect* Figure 2 shows the bias of the methods as the treatment effect varies. The standard binary method is unbiased, as we would expect for a logistic regression in a large sample. The latent variable method is unbiased for smaller treatment effects but a small bias towards the null is introduced as the treatment effect increases. The augmented binary method is biased away from the null in this setting and the bias increases as the treatment effect increases. Given that this performance is worse than is suggested from previous applications of the augmented binary method in Wason and Seaman (2013) and Wason and Jenkins (2016), this would suggest that the treatment effect from the augmented binary method may be biased if the model is misspecified. The coverage of the methods is shown in Figure 3. The binary method has

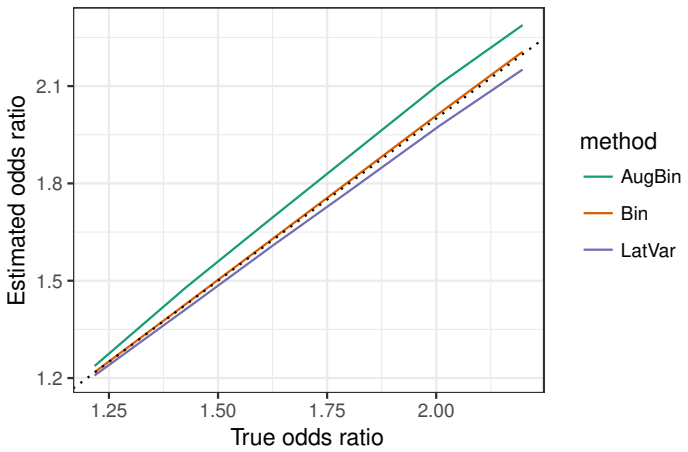


Fig. 2. Bias reported from the latent variable method, augmented binary method and standard binary method when n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

approximately nominal coverage. The latent variable method has nominal coverage for smaller treatment effects, however the coverage probability decreases as the treatment effect increases. The augmented binary method has coverage of approximately 0.91, which also decreases when the treatment effect increases. In order to diagnose this under-coverage in the joint modelling methods

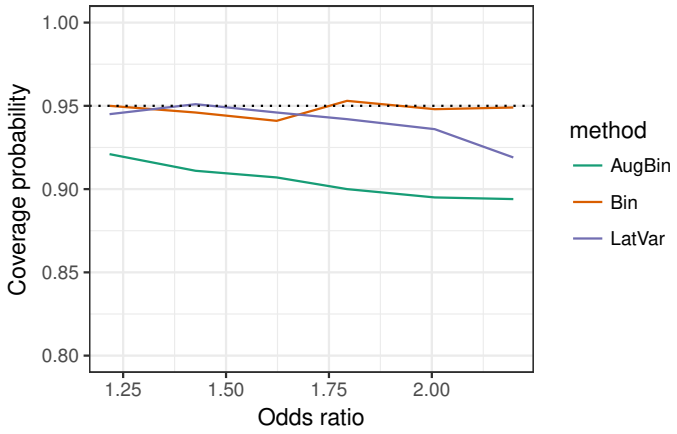


Fig. 3. Coverage probability reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

we can look at bias-corrected coverage, as recommended in Morris *and others* (2017). Figure 4 shows both the coverage and bias-corrected coverage for the three methods. The properties of the standard binary method remain unchanged. The bias-corrected coverage of the latent variable method is 0.95, which indicates that any under-coverage is due to the bias present. This is not true for the augmented binary method which shows small improvements in bias-corrected coverage, indicating that under-coverage is present in this method due to reasons other than bias. Again, this may be down to model misspecification. The power of the three methods is shown in Figure 5. The performance of the binary and augmented binary methods are as we would expect based on previous findings in Wason and Seaman (2013) and Wason and Jenkins (2016). The latent variable method offers much higher power. In this setting it has close to 100% power for odds

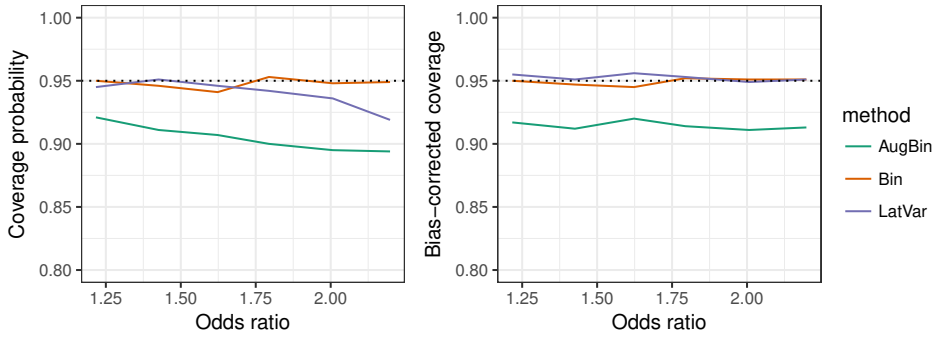


Fig. 4. Coverage probability (left) and bias-corrected coverage probability (right) reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

ratios larger than 1.6, an effect that is plausible to observe in a trial. These findings have indicated

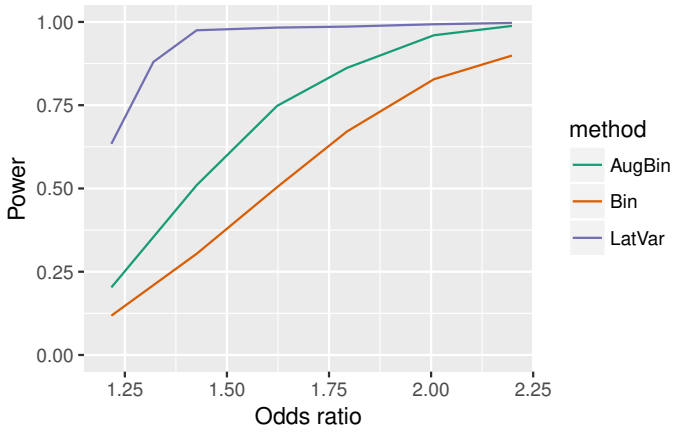


Fig. 5. Statistical power reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

that the standard binary method has the smallest bias and that the latent variable method has the smallest variance. The mean squared error (MSE) provides a combined measure of bias and

variance. Figure 6 shows the MSE of the three methods as the treatment effect varies. The MSE for the standard and augmented binary methods is approximately 6.5 times that of the latent variable method. However, this measure should be interpreted with care due to the fact that the MSE is more sensitive to the sample size than comparisons of bias or empirical SE alone (Morris *and others* (2017)).

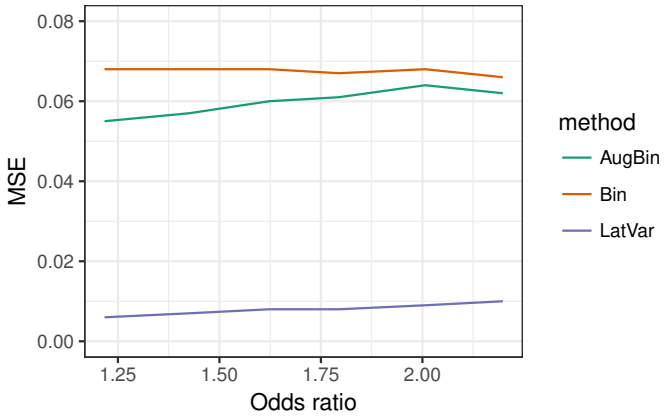


Fig. 6. Mean Squared Error (MSE) reported from the latent variable method, augmented binary method and standard binary method for n_{sim}=5000, total sample size N=300 for true log-odds treatment effect between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

4.2.2 *Varying* θ_1 To understand more about the precision performance of the augmented binary method in particular, we vary the responder threshold θ_1 to change the proportion of responders in that outcome. Figure 7 shows the density of the Y_1 variable and the relative precision of the methods, as the responder threshold varies. The precision gains from the augmented binary method diminish as the threshold increases. This is intuitive, as improvements in efficiency fall as the continuous component becomes less responsible for driving response. It is interesting to note that all precision gains are lost for any thresholds above -4. Therefore, even when 20% of patients are non-responders, all efficiency gains are lost. The percentage of responders needed to improve

efficiency using the augmented binary method will of course depend on the correlation structure employed. Due to the additional information in the other components, the latent variable method is still five times as precise as the other methods.

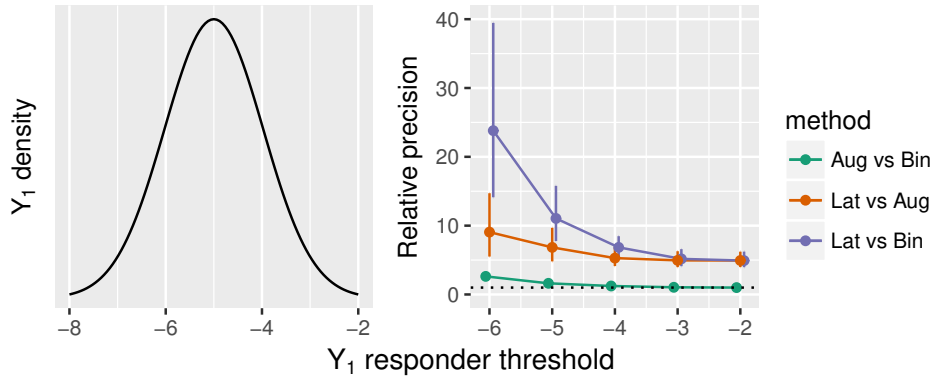


Fig. 7. Density of continuous Y_1 variable (left) and estimated relative precision of augmented binary versus standard binary method, latent variable versus augmented binary method and latent variable versus standard binary method as the Y_1 responder threshold θ_1 varies between $\theta_1 = -6$ and $\theta_1 = -2$ (right) for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

4.2.3 *Components contributing to response* An important consideration when investigating performance is how the precision changes when different combinations of outcomes are responsible for driving response. Figure 8 shows boxplots of the relative precision for the methods for four different response combinations, namely when response is driven by $(Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_4)$ and (Y_4), where Y_1 and Y_2 are observed as continuous variables, Y_3 is ordinal and Y_4 is binary.

When all four components contribute to response, the latent variable method outperforms the other methods, offering large precision gains. The variability in the magnitude of these gains is large, with the median result showing that the latent variable method reports the treatment effect 8 times more precisely than the binary method and 6 times more precisely than the augmented binary method. If response is driven by (Y_1, Y_2, Y_3) then the relative median gains for the latent

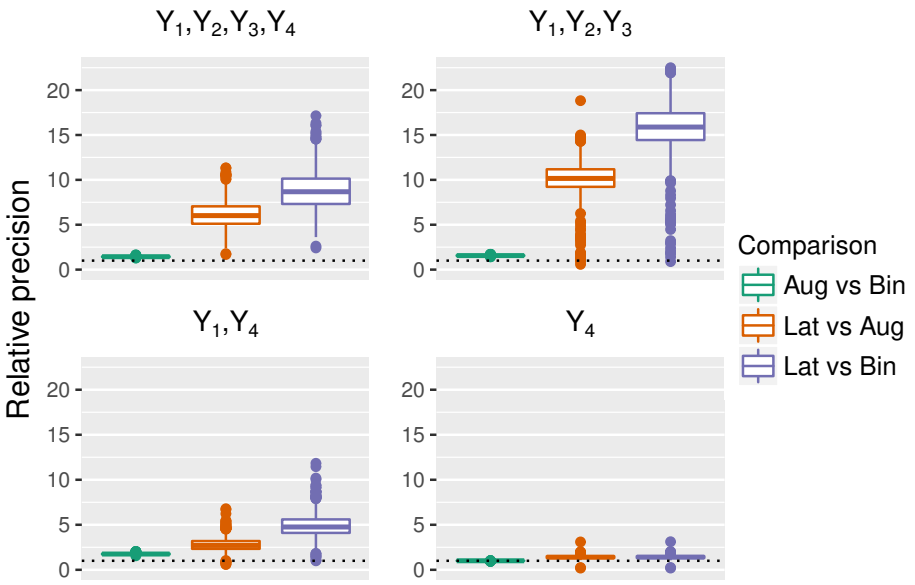


Fig. 8. Estimated relative precision gains from augmented binary versus standard binary method, latent variable versus augmented binary method and latent variable versus standard binary method when different combinations of components driving response. Response driven by Y_1, Y_2, Y_3, Y_4 (top left), Y_1, Y_2, Y_3 (top right), Y_1, Y_4 (bottom left) and Y_4 (bottom right) where Y_1, Y_2 are continuous, Y_3 is ordinal, Y_4 is binary for n_{sim}=5000 and total sample size N=300. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components

variable method are larger, however note that in less that 2% of cases the treatment effect is reported equally or less precisely than from both of the other methods. The findings are similar when response is driven by (Y_1, Y_4), however the median gains are much smaller. The treatment effect is reported 5 times more precisely from the latent variable method than the binary method in this setting. Note that as the augmented binary method models the relevant components it still performs well and again better than the latent variable method in a very small number of cases. When binary Y_4 determines response, the augmented binary method offers no improvement in precision whereas the latent variable method is approximately 1.5 times more precise. It is clear from the results that the magnitude of the precision gains from the latent variable method is

highly dependent on the structure of the data.

4.3 *Sensitivity analysis*

The key assumptions in this model are that of joint normality of the four components and that the discrete variables are realisations of latent continuous variables. Although it is not possible to test these assumptions in real data, we can investigate how robust the latent variable method is to deviations from these conditions. We can do this by drawing from the multivariate skew-normal distribution with different degrees of skew in each of the components. The first scenario investigated considers when all four components are skewed. Scenarios 2-3 consider different magnitudes of skew in the latent continuous components only. This tests the robustness of the method to the assumption that the observed discrete variables manifest from a true normal continuous variable. Scenario 4 is the null case for scenario 3. The results are shown in Appendix F of the supplementary material.

In summary, scenarios 1-3 have increased bias resulting in under-coverage as the bias-corrected coverage is close to nominal for all scenarios. The coverage of the latent variable method is nominal in the null case. This is consistent with our previous findings however the magnitude of the bias is much smaller when the assumptions are satisfied. The latent variable method still offers large power gains over the other methods. The MSE is smallest for the latent variable method across all scenarios investigated, indicating that the large reduction in variance is useful despite the introduction of bias. The latent variable method estimates the probability of response in the control arm well however underestimates the probability of response in the treatment arm. The magnitude of this underestimation is unaffected by the degree of skew or whether the skew is present in the observed continuous components. The relative precision of the methods are consistent with our previous findings indicating that the violation of joint normality only affects the bias and not the variance. The augmented binary and standard binary methods behave

similarly to when the joint normality assumptions are satisfied, which is expected given that the assumptions of those models are violated in both contexts.

5. Case study

5.1 *Data structure*

Due to data sharing policy, we conduct the analysis for a subset of the patients, N=278 rather than N=305 reported in the paper, so the results will differ from the original paper. Furthermore, only the anifrolumab 300mg arm (n=95) and the placebo arm (n=87) will be used to illustrate the methods.

The simulation results have suggested that the structure of the data is important for how the methods will perform, in particular the magnitude of the precision gains depends highly on which components drive response. Table 3 shows the criteria for response in each component and the rates of response in each by treatment arm. This suggests that the components responsible for responder discrimination are the continuous SLEDAI measurement and the binary taper measure. We can further explore the structure of the data by visualising the 4-D endpoint. Figure 9 shows a plot of the four components in the SLE index. The two panels show taper responders and non-responders, the levels in BILAG are denoted using colours where any coloured data points representing Grade B - Grade E are responders. The response thresholds for the continuous measurements are included, where a patient must be below the threshold to be considered a responder. We can conclude that response is entirely driven by SLEDAI and the taper variable, as there are no PGA non-responders not already accounted for by SLEDAI and no purple data points in the responder quadrant.

Table 3. Observed response rates in each of the SLE responder index components in the anifrolmab 300mg arm and placebo arm of the Phase IIb MUSE trial. SLE index is comprised of a continuous SLEDAI outcome, continuous PGA outcome, ordinal BILAG outcome and binary taper outcome where response in each component is achieved when the patient meets the criteria shown

Components	Response criteria	Treatment arm	
		Anifrolumab 300mg	Placebo
SLEDAI	Change in SLEDAI $\leqslant$ -4	58/89	41/76
PGA	Change in PGA $<$ 0.3	87/89	75/76
BILAG	No Grade A or more than one Grade B	86/89	72/76
Taper	Sustained reduction in oral corticosteroids	53/95	37/87
SLE responder index	Responder in all four components	34/95	18/87

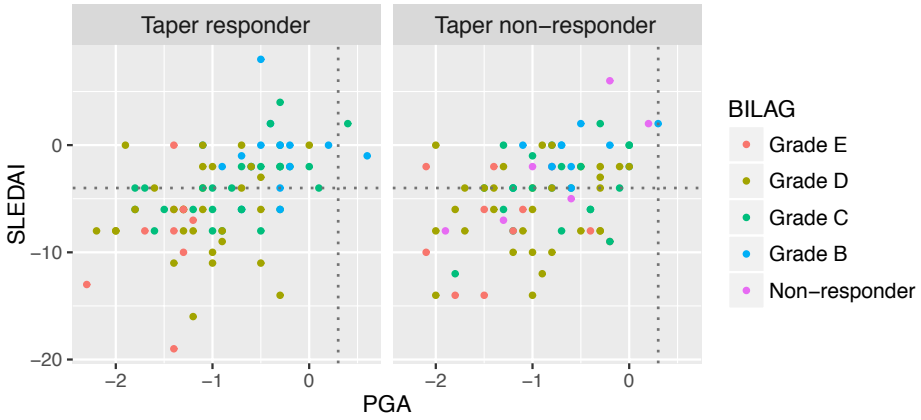


Fig. 9. Observed response rates in each of the SLE responder index components in the Phase IIb MUSE trial. SLEDAI is plotted on the y-axis and PGA on the x-axis, along with their corresponding dichotomisation thresholds. Levels of BILAG are represented by different colours and taper responders and non-responders are split across two panels.

5.2 *Results*

The probability of response in the placebo arm is estimated as 0.199 by the latent variable method, 0.211 by the augmented binary method and 0.224 by the standard binary method.

A much larger discrepancy between the methods is shown in the treatment arm, where the probability of response is estimated at 0.311, 0.324 and 0.382 in the latent variable, augmented binary and standard binary methods respectively.

The log-odds treatment effect point estimates and confidence intervals for the MUSE trial are shown in Table 4. Both joint modelling methods estimate the treatment effect more precisely. Although there may be bias present in the point estimates for the joint modelling methods, the confidence intervals entirely overlap with that of the binary method. All three methods indicate that anifrolumab 300mg performs better than placebo, as in the original findings. The latent variable model fits the data well according to the modified Pearson residuals, see Appendix G.

The simulation results indicated that the latent variable method may report the treatment effect with bias and have problems with bias related under-coverage when the treatment effect is large and when the assumption of joint normality is not satisfied. As the problems with the performance are bias related, we suggest implementing a bootstrap procedure to correct for this. In this scenario N=182 and n_{boot}=1000, therefore the procedure is as follows:

1. Sample with replacement N=182 patients from the MUSE trial
2. Compute the treatment effect using the latent variable, augmented binary and standard binary methods
3. Repeat step 1 and 2 n_{boot}=1000 times
4. Obtain an estimate of the bias using the difference between the treatment effect in the MUSE trial and the mean of the bootstrap treatment effects

A 95% bootstrap confidence interval for the treatment effect estimate can be obtained by ordering the 1000 bootstrap estimates of the treatment effect and taking the 25^{th} and 975^{th} estimate. The point estimates and 95% confidence intervals from the MUSE trial and from the bootstrap re-sampling are shown in Table 4. The log-odds point estimate from the latent variable

Table 4. Log-odds treatment effect estimates and 95% confidence intervals from the latent variable method, augmented binary method and standard binary method in the Phase IIb MUSE trial and the bootstrap sample when N=182 and $n_{boot} = 1000$

Method	Log-odds treatment effect	
	MUSE trial estimate	Bootstrap estimate
Latent Variable	0.641 (0.217, 1.072)	0.682 (0.275, 1.137)
Augmented binary	0.580 (0.139, 1.021)	0.608 (0.096, 1.111)
Binary	0.763 (0.078, 1.449)	0.809 (0.112, 1.561)

method has been shifted away from the null by approximately 0.04. This is the magnitude of bias that the simulation results suggested for this treatment effect. The width of the confidence interval hasn't changed much from the original estimate in the bootstrap sample, indicating that the variance is well estimated in the trial dataset. The point estimate for the binary method has also been shifted substantially, despite the simulations showing this method to be unbiased. This is likely due to the large imprecision in the treatment effect reported by the binary method.

In terms of estimated precision, it is interesting to determine where the trial data set lies in the distribution of datasets generated in the simulation study. The latent variable method reports the treatment effect 2.5 times more precisely than the standard binary method in this setting, whilst the augmented binary method is 2.4 times more precise. We would have expected this similar performance as the augmented binary method models the SLEDAI and taper variables - the only components driving response. This increase in precision from the latent variable method compared with the binary method amounts to a 60% reduction in required sample size.

6. Discussion

In this paper we addressed the issue of substantial losses of information when modelling complex composite endpoints. By employing concepts of partitioning latent variable outcome spaces we

could model the observed structure of the composite endpoint, which resulted in large gains in efficiency. Sensitivity analyses showed that a bias is introduced when the assumptions of joint normality were not satisfied, however similar reductions in variance were observed. When applying the methods to the MUSE trial, we implemented a bootstrap procedure to correct for the presumed bias, as joint normality could not be assessed. The treatment effect was reported 2.5 times more precisely than that reported from the standard binary method.

Bias correction appears to perform well in the real data, where the crucial assumptions cannot be tested. The point estimate is shifted by a magnitude that would have been expected from the simulation results and the estimate of the variance is similar to that obtained in the single trial dataset. Furthermore the bootstrap confidence interval for the treatment effect is contained within that for the binary method, which offers further reassurance for application. However, more work could be done to investigate different structures and scenarios to ensure that the bias correction is always performing as expected. Ideally, we would investigate this further across a large number of datasets however this is too computationally intensive. To perform this on one replicate, where $n_{boot} = 1000$ using 200 cores on a high performance computer (HPC) currently takes 7 hours. Exploring this further through bootstrapping or employing alternative multivariate distributions is an area for future research.

The precision gains offered by the latent variable method offer justification for the additional complexity. However, the magnitude of these gains are highly dependent on the components that drive response. The baseline case in the simulations was chosen to reflect when a composite endpoint is recommended for use, i.e. when all four components determine response rates. In this scenario, the precision gains achieved resulted in the latent variable method reporting the effect 2.5 to 17.5 times more precisely than the standard binary method. However, in practice in SLE trials, this has not been found to be the case. A review of two phase 3 trials (N= 2262) using the SRI-5 index found the SRI-5 response rate at week 52 for all patients was 32.8% (Kalunian *and*

others (2018)). Non-response due to a lack of SLEDAI improvement, concomitant medication non-compliance or dropout was 31, 16.5 and 19.1%, respectively. Non-response due to deterioration in BILAG or Physicians Global Assessment after SLEDAI improvement, concomitant medication compliance and trial completion was 0.5%. This is in agreement with our findings from the MUSE trial data, which suggests that the precision gains in the baseline case are optimistic. The simulation results show that when one continuous and one binary component drive response, the latent variable method may be anywhere between 1 and 12 times as precise as the binary method and up to 7 times as precise as the augmented binary method. In a very small number of cases ($<2\%$) there are no efficiency gains from using the latent variable method in this scenario. However the potential gains available in 98% of cases ensures that implementing the latent variable method is still very much a worthwhile endeavour, for all stakeholders in a clinical trial.

In addition to SLE, we have identified other disease areas that have a similar complex composite structure, meaning the potential to improve efficiency extends well beyond SLE. However, it must be acknowledged that the exact structure of the endpoint may offer different magnitudes of bias and precision, and may require longer computational time. Furthermore, in conditions where longitudinal data is required to sufficiently capture disease activity, trials may include multiple follow-up times and the method will need to be extended to include latent variables in the mean structure to account for this. In terms of scalability to more complex endpoints, the computational time depends on many things, in particular the number of outcomes, the outcome scale and the number of levels in the ordinal variable. In our case, we find the number of ordinal levels to be the most influential factor in computational time. This is due to the fact that 5 levels in the ordinal variable leads to 10 probability calculations in (3.6), however 3 levels would require the computation of 6 joint probabilities. Consequently, the run time will be substantially increased if there are multiple ordinal levels and decreased if the discrete variables are binary. If the computational time for a particular endpoint is deemed to be too large, then we may reduce

the complexity of the endpoint by collapsing the least informative components in to a single binary variable. It must be acknowledged that as we have coded the likelihood, with no package available to do this, the likelihood and probability of response code will have to be tailored specifically to each endpoint. The potential gains in efficiency justify this additional complexity. We have shown that the latent variable method is a powerful tool in composite endpoint analysis and should be considered as a primary analysis method in a trial using these endpoints. In order for implementation in the general case, where the composite contains any number of continuous and discrete outcomes and to ensure the uptake of the method in clinical trials, we will need to develop a software package. Furthermore, if patients and investigators are to benefit from the efficiency gains, we will need a method to calculate the required sample size in a given trial. We are currently working on addressing these issues to aid in the application of the method.

Acknowledgments

Disclosure: The MUSE trial data set was received under a data sharing contract with AstraZeneca.

References

Arminger, G. and Kusters, U. (1988). *Latent trait and latent class models*, Chapter Latent trait models with indicators of mixed measurement level. Plenum, pp. 51–73.

Arnold, B.C. (2009). Flexible univariate and multivariate models based on hidden truncation. *Journal of Statistical Planning and Inference* **139**, 3741–3749.

Ashford, J.R. and Sowden, R.R. (1970). Multivariate probit analysis. *Biometrics* **26**, 535–46.

Azzalini, A. and A., Dalla Valle. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

Catalano, P.J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine* **16**, 883–900.

Catalano, P.J. and Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87**(419), 651–658.

Chib, Siddhartha and Greenberg, Edward. (1998). Analysis of multivariate probit models. *Biometrika* **85**(2), 347–361.

Cox, D.R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**(3), 441–61.

de Leon, A.R. and Carriere, K.C. (editors). (2013). *Analysis of Mixed Data Methods and Applications*. Chapman and Hall/CRC.

de Leon, A.R and Wu, B. (2010). Copula based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* **30**, 175–185.

Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 355–366.

Faes, C., Geys, H., Aerts, M., Catalano, P.J. and Molenberghs, G. (2002). Modelling combined continuous and ordinal outcomes from developmental toxicity studies. In: Stasinopoulos, M. and Touloumi, G. (editors), *In Proceedings of the 17th International Workshop on Statistical Modelling*. Chania, Crete.

Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 845–852.

Furie, R., Khamashta, M., Merrill, J.T., Werth, V.P., Kalunian, K., Brohawn, P., Illei, G. G., Drappa, J., Wang, L., Yoo, S. *and others*. (2017). Anifrolumab, an anti

interferon alpha receptor monoclonal antibody, in moderate-to-severe systemic lupus erythematosus. *Arthritis and Rheumatology* **69**(2), 376–386.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–150.

Gueorguieva, R.V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**(455), 1102–1112.

Gueorguieva, R.V. and Sanacora, G. (2003). A latent variable model for joint analysis of repeatedly measured ordinal and continuous outcomes. In: Verbeke, G., Molenberghs, G., Aerts, M. and Fiews, S. (editors), *In Proceedings of the 18th International Workshop on Statistical Modelling*. Katholieke Universiteit Leuven: Leuven. pp. 171–176.

Gueorguieva, R.V. and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*.

Kalunian, Kenneth C, Urowitz, Murray B, Isenberg, David, Merrill, Joan T, Petri, Michelle, Furie, Richard A, Morgan-Cox, Mary-Ann, Taha, Rebecca, Watts, Steven, Silk, Maria *and others*. (2018). Clinical trial parameters that influence outcomes in lupus trials that use the systemic lupus erythematosus responder index. *Rheumatology* **57**(1), 125–133.

Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics* **17**, 31–54.

Lessafre, E. and Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine* **10**(9), 1391–1403.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research* **17**, 53–73.

McMenamin, M., A., Berglind and Wason, J.M.S. (2018). Improving the analysis of composite endpoints in rare disease trials. *Orphanet Journal of Rare Diseases* **13**, 81.

Morris, T.P., White, I.R. and Crowther, M.J. (2017). Using simulation studies to evaluate statistical methods. *arXiv:1712.03198*.

Nelsen, R.B. (1999). *An Introduction to Copulas: Definitions and Basic Properties*. New York, NY: Springer New York.

Olkin, I. and Tate, R.F. (1961). Multivariate correction models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448–465.

Pearson, K. (1904). Mathematical contributions to the theory of evolution. xii. on a generalised theory of alternative inheritance, with special reference to mendel's laws. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **203**, 53–86.

Poon, W.Y. and Lee, S.Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* **52**(3), 409–430.

Regan, M.M. and Catalano, P.J. (2000). Regression models and risk estimation for mixed discrete and continuous outcomes in developmental toxicology. *Risk Analysis* **20**, 363–376.

Samani, E.B. and Ganjali, M. (2008). A multivariate latent variable model for mixed continuous and ordinal responses. *World Applied Sciences Journal* **3**(2), 294–299.

Sammel, M.D. and Ryan, L.M. (2002). Effects of covariance misspecification in a latent variable model for multiple outcomes. *Statistica Sinica* **12**, 1207–1222.

Sammel, M.D., Ryan, L.M. and Leger, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc. B* **59**(3), 667–678.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC.

Tate, R.F. (1955). The theory of correlation between two continuous variables when one is dichotomised. *Biometrika* **42**(1-2), 205–216.

Verbeke, G., Fieuws, S. and Molenberghs, G. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research* **23**(1), 42–59.

Wason, J. and Jenkins, M. (2016). Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. *Rheumatology* **55**(10), 1796–1802.

Wason, J. and Seaman, S. R. (2013). Using continuous data on tumour measurements to improve inference in phase ii cancer studies. *Statistics in Medicine* **32**(26), 4639–4650.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley and Sons.

Wu, B. and de Leon. A.R. (2014). Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *JABES* **19**(1), 39–56.

Supplementary material

Appendix A

The joint probability below expresses, for patient i with $Y_1 = y_{i1}$ and $Y_2 = y_{i2}$, the probability that they will have a Y_3 score w and a Y_4 score k.

$$pr\left(Y_{i3} = w, Y_{i4} = k | Y_{i1} = Y_{i1}, y_{i2} = y_{i2}; \boldsymbol{\theta}\right) =$$

$$\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) -$$

$$\Phi_2\left(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi_2\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) \quad \text{(A.1)}$$

The intuition for the joint probability can be seen below in Figure 10, specifically for the SLE endpoint, where $w = 5$ and $k = 2$.

The blue box indicates the region where $w = 3$ and $k = 2$. As $\tau_{03} = \tau_{04} = -\infty$ and $\tau_{53} = \tau_{24} = \infty$, the corresponding probability is shown in (A.2).

$$pr\left(Y_{i3} = 3, Y_{i4} = 2 | Y_{i1} = y_{i1}, Y_{i2} = y_{i2}; \boldsymbol{\theta}\right) =$$

$$\Phi_2\left(\tau_{33} - \mu_{3|1,2}, \infty - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi_2\left(\tau_{23} - \mu_{3|1,2}, \infty - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) -$$

$$\Phi_2\left(\tau_{33} - \mu_{3|1,2}, \tau_{(14} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi_2\left(\tau_{23} - \mu_{3|1,2}, \tau_{14} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) \quad \text{(A.2)}$$

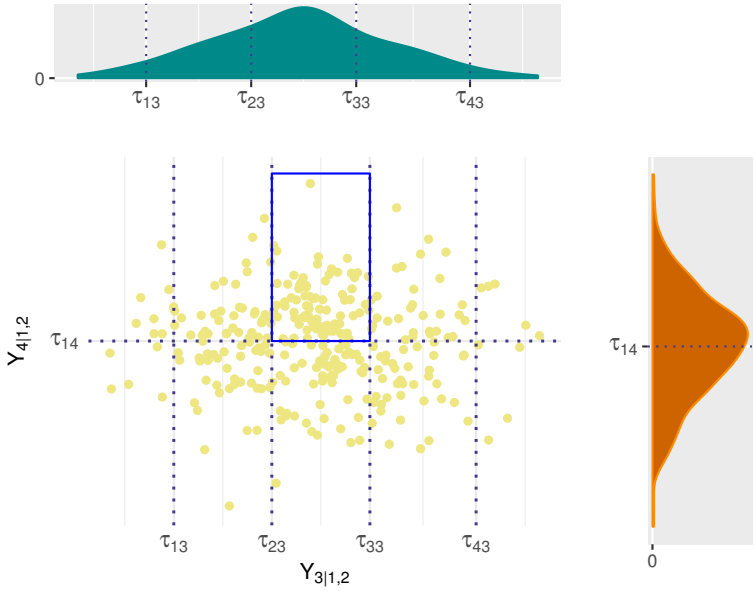


Fig. 10. The figure shows the conditional outcome $Y_{3|1,2}$ on the x-axis and $Y_{4|1,2}$ on the y-axis with their corresponding underlying continuous densities and partitioning thresholds. The area where $w = 3$ and $k = 2$ is highlighted for illustration

Appendix B

One suggestion in the literature for assessing goodness-of-fit in latent variable models is introduced by Samani and Ganjali (2008) for the case when there is one continuous and one ordinal variable. This may be extended to allow for two continuous, one ordinal and one binary outcome for application in SLE, as shown below.

As before, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ be the vector of observed responses for patient i. Then, partitioning the observed and latent continuous measures, we let $\mathbf{Y_{cts}} = (Y_1, Y_2)$ and $\mathbf{Y_{dis}} = (Y_3, Y_4)$. Then, $\hat{\Sigma}_{11} = \hat{Var}(\mathbf{Y_{cts}}), \hat{\Sigma}_{22} = \hat{Var}(\mathbf{Y_{dis}}), \hat{\Sigma}_{12} = \hat{\Sigma}_{21} = \hat{Cov}(\mathbf{Y_{cts}}, \mathbf{Y_{dis}})$.

The modified Pearson residuals, taking in to account the correlation between responses are shown below.

$$r_i^p = \hat{\Sigma}^{-\frac{1}{2}}(Y_i - \hat{\mu}_i) \tag{B.1}$$

where,

$$\boldsymbol{\hat{\mu}_i} = (\hat{E}(Y_{i1}, Y_{i2} \mid X_{i1}, X_{i2}), \hat{E}(Y_{i3}, Y_{i4} \mid X_{i3}, X_{i4}))' \tag{B.2}$$

and

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix} \tag{B.3}$$

A Cholesky decomposition may be used to obtain $\hat{\Sigma}^{-\frac{1}{2}}$ in (B.1). The covariance between the vector of observed continuous and observed discrete responses is shown below.

$$\begin{aligned}
\Sigma_{12} =& E(\mathbf{Y_{cts}}\mathbf{Y_{dis}}) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\
=& E(\mathbf{Y_{cts}}E(\mathbf{Y_{dis}} \mid \mathbf{Y_{cts}})) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\
=& E(Y_1 Y_2 E(Y_3, Y_4 \mid Y_1, Y_2)) - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}}) \\
=& \int_{y_1}\int_{y_2} y_1 y_2 \sum_{y_3}\sum_{y_4} y_3 y_4 P\left(Y_3 = w, Y_4 = k | Y_1 = y_1, Y_2 = y_2\right) f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 \\
& - E(\mathbf{Y_{cts}})E(\mathbf{Y_{dis}})
\end{aligned}$$

Where,

$$P\left(Y_3 = w, Y_4 = k | Y_1 = y_1, Y_2 = y_2\right) =$$

$$\Phi\left(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) - \Phi\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) -$$

$$\Phi\left(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right) + \Phi\left(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}\right)$$

$$E(\mathbf{Y_{cts}}) = \int_{y_1} \int_{y_2} y_1 y_2 f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2$$

$$E(\mathbf{Y_{dis}}) = \sum_{y_3} \sum_{y_4} y_3 y_4 P\left(Y_3 = w, Y_4 = k\right)$$

and

$$P(Y_3 = w, Y_4 = k) = \Phi(\tau_{w3} - \mu_3, \tau_{k4} - \mu_4; \rho_{3,4}) - \Phi(\tau_{(w-1)3} - \mu_3, \tau_{k4} - \mu_4; \rho_{3,4}) -$$

$$\Phi(\tau_{w3} - \mu_3, \tau_{(k-1)4} - \mu_4; \rho_{3,4}) + \Phi(\tau_{(w-1)3} - \mu_3, \tau_{(k-1)4} - \mu_4; \rho_{3,4})$$

The Pearson residual is based on the Pearson goodness-of-fit statistics

$$\chi^2_p = \sum_{i=1}^{n} \chi^2_p(Y_i, \hat{\mu}_i) \tag{B.4}$$

with ith component

$$\chi^2_p(Y_i, \hat{\mu}_i) = (Y_i - \hat{\mu}_i)' \hat{\Sigma}^{-1} (Y_i - \hat{\mu}_i) \tag{B.5}$$

The distribution of the residuals should follow a chi-squared distribution with p degrees of freedom. Comparing the residuals to the chi-squared value allows us to identify observations which the model does not fit well. If there are many observations unexplained by the model then it could indicate a poor choice of model. This may be due to the covariance structure $\hat{\Sigma}$ and its assumed distribution. The model may be refitted with various covariance structures and to obtain a model which is found to satisfactorily explain the observed data. If this is not achieved then joint normality of the error terms may be an unreasonable assumption indicating that the latent variable model may not be appropriate. It is possible to fit latent variable models which assume a different multivariate distribution for the error terms, however this is not considered here.

Appendix C

Augmented binary method

The augmented binary model is shown below. The baseline measures for Y_{i1} and Y_{i2} are included for comparison, as they are accounted for in the mean structure of the latent variable method. As one time point is modelled we can use a linear model for Y_{i1} as shown in (C.1). Note that Y_{i1} or Y_{i2} may be chosen as the continuous measure to retain and should always be determined by which is the most informative.

$$Y_{i1} = \delta_0 + \delta_1 Ti + \delta_2 y_{i10} + \delta_3 y_{i20} + \varepsilon_i \tag{C.1}$$

where $\varepsilon_i | T_i, y_{i10}, y_{i20} \sim N(0, \sigma)$. In this case, the failure time binary indicator will contain information from the remaining three components. F_i is set to equal 0 if $Y_{i2} \leqslant \theta_2, Y_{i3}$ is Grade B-E and $Y_{i4} = 0$, otherwise the patient is labelled a non-responder in these components and $F_{i1} = 1$. F_i is modelled using the logistic regression model in (C.2).

$$logit(Pr(F_i = 1 | T_i, y_{i10}, y_{i20}) = \alpha_F + \beta_F T_i + \gamma_F y_{i10} + \psi_F y_{i20} \tag{C.2}$$

Maximum likelihood estimates for the parameters are obtained from fitting models (C.1) and (C.2). The probability of response is shown in (C.3).

$$P(Y_{i1} \leqslant \theta_1, F_{i1} = 0 | T_i, y_{i10}, y_{i20}) = \int_{-\infty}^{\theta_1} P(F_{i1} = 0 | T_i, y_{i10}, y_{i20}) f_{Y_1}(y_{i1}; T_i, y_{i10}, y_{i20}) dy_{i1} \tag{C.3}$$

As in the latent variable method, (C.3) is used to obtain probability of response estimates for each patient, assuming they were treated $\tilde{p_{i1}}$ and not treated $\tilde{p_{i1}}$, which are used to define an odds ratio, risk ratio or risk difference.

Standard binary method

The standard binary method is a logistic regression on the overall responder index, as shown in (C.4).

$$logit(Pr(S_i = 1|T_i, y_{i10}) = \alpha + \beta T_i + \gamma y_{i10} + \psi y_{i20} \tag{C.4}$$

The odds ratio and standard error estimates can be obtained directly.

Appendix D

Table 5. Performance measures and Monte Carlo standard errors used to assess the behaviour of the latent variable, augmented binary and binary methods in a simulation study for the systemic lupus erythematosus composite endpoint

Performance measure	Estimate	MCSE
Bias	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \hat{\theta}_j - \theta$	$\sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{j=1}^{n_{sim}} (\hat{\theta}_j - \bar{\theta})^2}$
Coverage	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(\hat{\theta}_{low,j} \leqslant \theta \leqslant \hat{\theta}_{upp,j})$	$\sqrt{\frac{\hat{cov.}(1-\hat{cov.})}{n_{sim}}}$
Bias-corrected coverage	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(\hat{\theta}_{low,j} \leqslant \bar{\theta} \leqslant \hat{\theta}_{upp,j})$	$\sqrt{\frac{\hat{BEcov.}(1-\hat{BEcov.})}{n_{sim}}}$
Power	$\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} 1(p_j < \alpha)$	$\sqrt{\frac{\hat{Power}(1-\hat{Power})}{n_{sim}}}$
MSE	$\sum_{j=1}^{n_{sim}} (\hat{\theta}_j - \theta)^2$	$\sqrt{\frac{\sum_{j=1}^{n_{sim}} [(\hat{\theta}_j-\theta)^2 - \hat{MSE}]^2}{n_{sim}(n_{sim}-1)}}$
Empirical SE	$\sqrt{\frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\hat{EmpSE}}{\sqrt{2(n_{sim}-1)}}$
Model SE	$\sqrt{\frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} \hat{Var}(\hat{\theta}_j)}$	$\sqrt{\frac{\hat{Var}[\hat{Var}(\hat{\theta})]}{4n_{sim}\hat{ModSE}^2}}$ †
Relative precision A vs. B	$\frac{\hat{Var}(\hat{\theta}_j)_B}{\hat{Var}(\hat{\theta}_j)_A}$	-

$\hat{\theta}_j$: estimated log-odds treatment effect in simulated data j
$\bar{\theta}$: mean log-odds treatment effect over n_{sim} datasets
$\hat{\theta}_{low,j}, \hat{\theta}_{upp,j}$ lower and upper limit of confidence interval for iteration j
† $\hat{Var}[\hat{Var}(\hat{\theta})] = \frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} \{\hat{Var}(\hat{\theta}_i) - \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \hat{Var}(\hat{\theta}_j)\}^2$

Appendix E

Table 6. Parameter values for the simulated scenarios which investigate the effect of varying responder threshold θ_1, changing the components driving response and differing treatment effects on the performance of the latent variable, augmented binary and standard binary methods for the systemic lupus erythematosus composite endpoint

Scenario	Parameters	Investigates
$\theta_1 = -2$	$\theta_1 = -2$	100% of patients respond in Y_1
$\theta_1 = -3$	$\theta_1 = -3$	96% of patients respond in Y_1
$\theta_1 = -4$	$\theta_1 = -4$	82% of patients respond in Y_1
$\theta_1 = -5$	$\theta_1 = -5$	52% of patients respond in Y_1
$\theta_1 = -6$	$\theta_1 = -6$	20% patients respond in Y_1
Y_1, Y_4	$\theta_1 = -5, \theta_2 = 2, \theta_3 = 2$	Continuous and binary variable driving response
Y_4	$\theta_1 = -2, \theta_2 = 2, \theta_3 = 2$	Binary variable driving response
Y_1, Y_2, Y_3	$\theta_4 = 2$	Two continuous and ordinal drive response
Treat case 1	$\alpha_0 = -4.9, \alpha_1 = -0.09, \beta_0 = -1.2, \beta_1 = -0.11, \gamma_1 = -0.145, \psi_0 = -0.2, \psi_1 = -0.07$	Odds ratio = 1.217
Treat case 2	$\alpha_0 = -4.9, \alpha_1 = -0.20, \beta_0 = -1.2, \beta_1 = -0.25, \gamma_1 = -0.2, \psi_0 = -0.2, \psi_1 = -0.12$	Odds ratio = 1.426
Treat case 3	$\alpha_0 = -4.9, \alpha_1 = -0.30, \beta_0 = -1.2, \beta_1 = -0.50, \gamma_1 = -0.3, \psi_0 = -0.2, \psi_1 = -0.22$	Odds ratio = 1.794
Treat case 4	$\alpha_0 = -4.9, \alpha_1 = -0.32, \beta_0 = -1.2, \beta_1 = -0.65, \gamma_1 = -0.39, \psi_0 = -0.2, \psi_1 = -0.27$	Odds ratio = 2.007
Treat case 5	$\alpha_0 = -4.9, \alpha_1 = -0.33, \beta_0 = -1.2, \beta_1 = -0.72, \gamma_1 = -0.45, \psi_0 = -0.2, \psi_1 = -0.33$	Odds ratio = 2.198

Appendix F

Multivariate skew-normal distribution

To test the robustness of the latent variable method to deviations from joint normality of the components, we can generate the data so that the components are drawn from a multivariate skew-normal. The multivariate skew-normal is an extension of the univariate skew-normal distribution introduced by Azzalini and A. (1996). They define it as follows. A random vector $\mathbf{Y}=(Y_1, ...,_k)^T$ has k-variate skew-normal distribution, if its density function is

$$f_k(\mathbf{y}) = 2\phi_k(\mathbf{y};\boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\mathbf{y}), \mathbf{y} \in \mathbf{R}^k \tag{F.1}$$

where $\phi_k(\mathbf{y};\boldsymbol{\Omega})$ is the probability density function of the k-variate normal distribution with standardised marginals and correlation matrix $\boldsymbol{\Omega}$. The shape parameter $\boldsymbol{\alpha}$ determines the skewness, where $\boldsymbol{\alpha} = \mathbf{0}$ reduces the density in (F.1) to the $\mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$ density.

Scenarios of interest are shown in Table 7. The first scenario considers when all four components are skewed. Scenarios 2-3 consider different magnitudes of skew in the latent continuous components only. This tests the robustness of the method to the assumption that the observed discrete variables manifest from continuous variables. Scenario 4 is the null case for scenario 3.

Table 7. Simulation scenarios considered to investigate deviations from joint normality for the components of the systemic lupus erythematosus composite endpoint based on the multivariate skew-normal distribution where $\boldsymbol{\alpha}$ determines the magnitude of the skew in each component

Scenario	$\boldsymbol{\alpha}$	Purpose
skew1	(0.1, 0.1, 0.1, 0.1)	Skew in all four components
skew2	(0, 0, 0.1, 0.1)	Skew in discrete components only
skew3	(0, 0, 0.05, 0.05)	Smaller skew in discrete components only
skew4	(0, 0, 0.05, 0.05)	Smaller skew in discrete components only in the null case

Results

The bias, coverage, bias-corrected coverage and power are shown in Table 8 for all four scenarios. In scenarios 1-3, the non-normality introduces bias which results under-coverage. The bias-corrected coverage is close to nominal for all scenarios however the coverage of the latent variable method is nominal in the null case. This is consistent with our findings when the joint normality assumption is satisfied in that bias is introduced in the estimation of the treatment arm, however the magnitude of this bias is much smaller when the assumptions are satisfied. The augmented binary and standard binary methods behave similarly to when the joint normality assumptions are satisfied, which is expected given that the assumptions of those models are violated in both contexts. The latent variable method still offers large power gains over the other methods.

Table 9 shows the MSE, empirical SE and model SE of the three methods. The latent variable method performs best consistently across these performance measures. The augmented binary and standard binary methods have an MSE across all scenarios of approximately 0.06 whilst the MSE of the latent variable method is between 0.01 and 0.04. This indicates that the large reduction in variance is useful despite the introduction of bias. We acknowledge however that this may not hold across all sample sizes (Morris *and others* (2017)).

Table 10 shows the probability of response in each arm for each of the methods. The findings are consistent with when the assumptions are satisfied. Namely, the latent variable method estimates the probability of response in the control arm well however underestimates the probability of response in the treatment arm. The magnitude of this underestimation is unaffected by the degree of skew or whether the skew is present in the observed continuous components.

The odds ratio treatment effect estimate from each method is shown in Table 11. The latent variable method is biased towards the null, the augmented binary method is biased away from the null. The binary method slightly underestimates the treatment effect in this setting however

Table 8. Operating characteristics of the latent variable, augmented binary and binary methods when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
Bias	skew1	-0.173 (0.012)	0.041 (0.252)	-0.015 (0.258)
	skew2	-0.103 (0.008)	0.036 (0.251)	-0.020 (0.255)
	skew3	-0.068 (0.008)	0.038 (0.244)	-0.016 (0.245)
	skew4	-0.033 (0.008)	0.007 (0.254)	0.001 (0.255)
Coverage	skew1	0.556 (0.018)	0.933 (0.009)	0.939 (0.009)
	skew2	0.811 (0.013)	0.928 (0.008)	0.941 (0.008)
	skew3	0.884 (0.010)	0.934 (0.008)	0.950 (0.007)
	skew4	0.933 (0.009)	0.923 (0.009)	0.950 (0.008)
Bias-corrected coverage	skew1	0.962 (0.007)	0.929 (0.009)	0.943 (0.008)
	skew2	0.936 (0.008)	0.930 (0.008)	0.943 (0.007)
	skew3	0.940 (0.008)	0.929 (0.008)	0.954 (0.007)
	skew4	0.948 (0.008)	0.926 (0.009)	0.950 (0.008)
Power	skew1	0.897 (0.011)	0.646 (0.017)	0.487 (0.018)
	skew2	0.959 (0.006)	0.637 (0.015)	0.471 (0.016)
	skew3	0.982 (0.004)	0.641 (0.015)	0.495 (0.016)
	skew4	-	-	-

all are close to true for the null case.

The median relative precision of the methods are shown in Table 12, with the 10th centile and 90th centile values. These are consistent with our previous findings indicating that the violation of joint normality only affects the bias and not the variance.

Table 9. Operating characteristics (Monte Carlo standard errors in parentheses) of the latent variable, augmented binary and binary methods when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

Performance measure	Scenario	Method		
		Latent Variable	Augmented Binary	Binary
MSE	skew1	0.039 (0.001)	0.063 (0.003)	0.066 (0.003)
	skew2	0.021 (0.001)	0.063 (0.003)	0.065 (0.003)
	skew3	0.014 (0.001)	0.060 (0.003)	0.060 (0.003)
	skew4	0.010 (0.001)	0.064 (0.004)	0.065 (0.003)
EmpSE	skew1	0.097 (0.003)	0.248 (0.006)	0.257 (0.007)
	skew2	0.102 (0.002)	0.249 (0.006)	0.254 (0.006)
	skew3	0.099 (0.002)	0.241 (0.005)	0.245 (0.006)
	skew4	0.094 (0.002)	0.254 (0.006)	0.255 (0.006)
ModSE	skew1	0.010 (0.006)	0.052 (0.001)	0.064 (0.001)
	skew2	0.010 (0.003)	0.050 (0.001)	0.060 (0.001)
	skew3	0.010 (0.015)	0.048 (0.001)	0.059 (0.001)
	skew4	0.009 (0.004)	0.051 (0.001)	0.063 (0.001)

Table 10. Estimated probability of response in the treatment and placebo arms from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	$Pr(resp \mid T = 0)$				$Pr(resp \mid T = 1)$			
Scenario	True	Lat Var	Aug Bin	Bin	True	Lat Var	Aug Bin	Bin
skew1	0.259	0.263	0.221	0.258	0.365	0.330	0.326	0.359
skew2	0.290	0.287	0.253	0.290	0.398	0.370	0.361	0.392
skew3	0.309	0.302	0.271	0.308	0.418	0.394	0.382	0.413
skew4	0.309	0.299	0.269	0.307	0.309	0.292	0.270	0.307

Table 11. Estimated odds ratio treatment effect from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Treatment effect			
Scenario	True	Lat Var	Aug Bin	Bin
skew1	1.640	1.379 (1.140, 1.668)	1.708 (1.093, 2.668)	1.616 (0.985, 2.651)
skew2	1.617	1.459 (1.203, 1.770)	1.676 (1.083, 2.594)	1.586 (0.980, 2.565)
skew3	1.611	1.505 (1.243, 1.822)	1.674 (1.089, 2.572)	1.585 (0.987, 2.548)
skew4	1.000	0.967 (0.807, 1.160)	1.007 (0.647, 1.566)	1.001 (0.613, 1.634)

Table 12. Estimated relative precision from the latent variable model (Lat Var), augmented binary method (Aug Bin) and standard binary method (Bin) when the components of the systemic lupus erythematosus endpoint are drawn from a multivariate skew-normal, N=300 and $n_{sim} = 1000$

	Treatment effect		
Scenario	Lat Var vs Bin	Lat Var vs Aug Bin	Aug Bin vs Bin
skew1	6.903 [5.336, 8.972]	5.579 [4.376, 7.313]	1.231 [1.189, 1.275]
skew2	6.263 [5.013, 7.917]	5.177 [4.096, 6.518]	1.213 [1.178, 1.252]
skew3	6.326 [5.016, 7.995]	5.192 [4.098, 6.548]	1.219 [1.184, 1.257]
skew4	7.384 [5.729, 9.343]	5.985 [4.655, 7.629]	1.231 [1.192, 1.273]

Appendix G

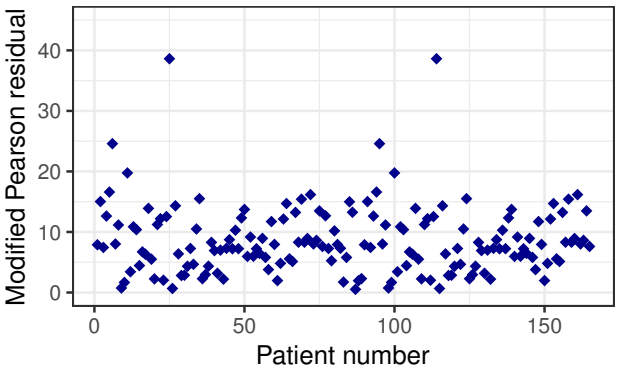


Fig. 11. Plot of the modified Pearson residuals from the latent variable model for each patient in the MUSE trial. The residuals highlight that two patients observations are poorly explained by the model but that the model is a good fit for the remaining patients.

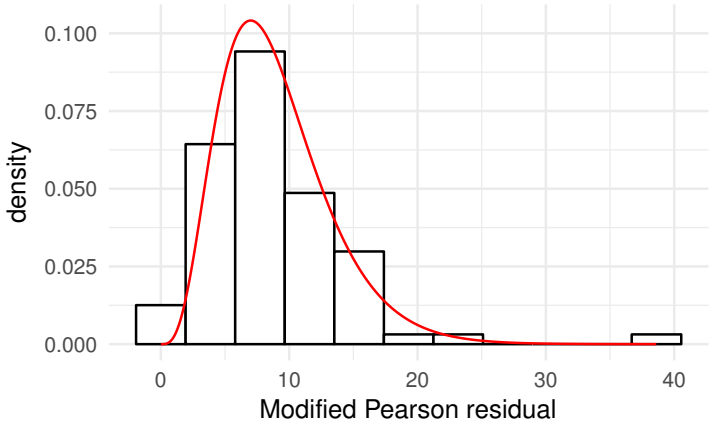


Fig. 12. Histogram of the modified Pearson residuals from the latent variable model in the MUSE trial dataset with the corresponding χ^2 density. The modified Pearson residuals should follow the distribution of the χ^2 density shown if the model fits well.

Appendix F

Latent Variable Method: R Code

```
library(MASS)
library(stats)
library(mvtnorm)
library(nlme)
library(boot)
#library(car)
library(matrixcalc)
#library(profvis)
library(numDeriv)
library(R2Cuba)
library(optimx)
library(brglm)
library(Matrix)

###LATENT VARIABLE CODE FOR FOUR DIMENSIONAL COMPOSITE ENDPOINT
###COMPONENTS: TWO CONTINUOUS, ONE ORDINAL, ONE BINARY

##STARTING VALUES

X<-c() #Add based on data
dat<-data.frame(id,treat,Z1,Z2,Z3ord,Z4ord,Z10,Z20) ## ORDER OF DATA INPUT

##LIKELIHOOD FUNCTION

f<-function(X,dat)
{

  dat<-dat[!is.na(dat[,3]),]

  #parameters
  alpha0 <- X[1]
  alpha1 <- X[2]
  beta0 <- X[3]
  beta1 <- X[4]
  gamma1 <- X[5]
  psi0 <- X[6]
  psi1 <- X[7]

  #covariance parameters
  sig1 <- exp(X[8])
  sig2 <- exp(X[9])
  rho12 <- 2*inv.logit(X[10])-1
  rho13 <- 2*inv.logit(X[11])-1
  rho14 <- 2*inv.logit(X[12])-1
  rho23 <- 2*inv.logit(X[13])-1
  rho24 <- 2*inv.logit(X[14])-1
  rho34 <- 2*inv.logit(X[15])-1

  #ordinal cutoffs
  tau13 <- X[16]
  tau23 <- X[17]
```

```
  tau33 <- X[18]
  tau43 <- X[19]

  #addition baseline parameters
  alpha2 <- X[20]
  beta2 <- X[21]


  #Known cutoffs
  tau03 <- -Inf
  tau53 <- +Inf
  tau04 <- -Inf
  tau14<- 0
  tau24<- +Inf

  #model means
  muz1<-alpha0+alpha1*dat[,2]+alpha2*dat[,7]
  muz2<-beta0+beta1*dat[,2]+beta2*dat[,8]
  muz3<-gamma1*dat[,2]#+gamma2*dat[,9]
  muz4<-psi0+psi1*dat[,2]

  #conditional means
  muz3cond <- muz3+((rho13-rho12*rho23)*(dat[,3]-muz1)/(sig1*(1-(rho12)^2)))+((rho23-rho13*rho12)*(dat[,4]-muz2)/(
       sig2*(1-(rho12)^2)))
  muz4cond <- muz4+((rho14-rho12*rho24)*(dat[,3]-muz1)/(sig1*(1-(rho12)^2)))+((rho24-rho14*rho12)*(dat[,4]-muz2)/(
       sig2*(1-(rho12)^2)))

  #condational covariance
  matcond11 <- 1-(((rho13)^2-2*rho12*rho13*rho23+(rho23)^2)/(1-(rho12)^2))
  matcond12 <- rho34-((rho13*rho14-rho13*rho12*rho24-rho12*rho14*rho23+rho23*rho24)/(1-(rho12)^2))
  matcond22 <- 1-(((rho14)^2-2*rho12*rho14*rho24+(rho24)^2)/(1-(rho12)^2))
  Sigcond <- matrix(c(matcond11, matcond12, matcond12, matcond22), nrow=2, ncol=2)
  Sigcond2 <-(Sigcond*t(Sigcond))^0.5

  #continuous bivariate covariance

  matbiv11 <- (sig1)^2
  matbiv12 <- rho12*sig1*sig2
  matbiv22 <- (sig2)^2
  Sigbiv   <- matrix(c(matbiv11, matbiv12, matbiv12, matbiv22), nrow=2, ncol=2)
  Sigbiv <- (Sigbiv*t(Sigbiv))^0.5

  #upperlimits
  mu11<-matrix(c(tau13-muz3cond, tau14-muz4cond), ncol=2)
  mu01<-matrix(c(tau03-muz3cond, tau14-muz4cond), ncol=2)
  mu10<-matrix(c(tau13-muz3cond, tau04-muz4cond), ncol=2)
  mu00<-matrix(c(tau03-muz3cond, tau04-muz4cond), ncol=2)
  mu21<-matrix(c(tau23-muz3cond, tau14-muz4cond), ncol=2)
  mu20<-matrix(c(tau23-muz3cond, tau04-muz4cond), ncol=2)
  mu31<-matrix(c(tau33-muz3cond, tau14-muz4cond), ncol=2)
  mu30<-matrix(c(tau33-muz3cond, tau04-muz4cond), ncol=2)
  mu41<-matrix(c(tau43-muz3cond, tau14-muz4cond), ncol=2)
  mu40<-matrix(c(tau43-muz3cond, tau04-muz4cond), ncol=2)
  mu51<-matrix(c(tau53-muz3cond, tau14-muz4cond), ncol=2)
  mu50<-matrix(c(tau53-muz3cond, tau04-muz4cond), ncol=2)
  mu12<-matrix(c(tau13-muz3cond, tau24-muz4cond), ncol=2)
  mu02<-matrix(c(tau03-muz3cond, tau24-muz4cond), ncol=2)
  mu22<-matrix(c(tau23-muz3cond, tau24-muz4cond), ncol=2)
  mu32<-matrix(c(tau33-muz3cond, tau24-muz4cond), ncol=2)
  mu42<-matrix(c(tau43-muz3cond, tau24-muz4cond), ncol=2)
  mu52<-matrix(c(tau53-muz3cond, tau24-muz4cond), ncol=2)

  pr11<-apply(mu11,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr01<-apply(mu01,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr10<-apply(mu10,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr00<-apply(mu00,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr21<-apply(mu21,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr20<-apply(mu20,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr31<-apply(mu31,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr30<-apply(mu30,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr41<-apply(mu41,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr40<-apply(mu40,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr51<-apply(mu51,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr50<-apply(mu50,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
  pr12<-apply(mu12,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
```

```
pr02<-apply(mu02,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
pr22<-apply(mu22,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
pr32<-apply(mu32,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
pr42<-apply(mu42,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
pr52<-apply(mu52,1,function(x){return(pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=c(0,0),sigma = Sigcond2))})
prz12<-dmvnorm(cbind(dat[,3],dat[,4]), c(mean(muz1), mean(muz2)), Sigbiv)

##Likelihood function

#components of likelihood, w=1,..5 (ordinal); k=1,2 (binary)
l1<-log(pr11-pr01-pr10+pr00)+log(prz12)#w=1,k=1
l2<-log(pr21-pr11-pr20+pr10)+log(prz12)#w=2, k=1
l3<-log(pr31-pr21-pr30+pr20)+log(prz12)#w=3, k=1
l4<-log(pr41-pr31-pr40+pr30)+log(prz12)#w=4, k=1
l5<-log(pr51-pr41-pr50+pr40)+log(prz12)#w=5, k=1
l6<-log(pr12-pr02-pr11+pr01)+log(prz12)#w=1, k=2
l7<-log(pr22-pr12-pr21+pr11)+log(prz12)#w=2, k=2
l8<-log(pr32-pr22-pr31+pr21)+log(prz12)#w=3, k=2
l9<-log(pr42-pr32-pr41+pr31)+log(prz12)#w=4, k=2
l10<-log(pr52-pr42-pr51+pr41)+log(prz12)#w=5, k=2

data0 <- cbind(dat[,5],dat[,6],l1)#1,1
data1 <- cbind(dat[,5],dat[,6],l2)#2,1
data2 <- cbind(dat[,5],dat[,6],l3)#3,1
data3 <- cbind(dat[,5],dat[,6],l4)#4,1
data4 <- cbind(dat[,5],dat[,6],l5)#5,1
data5 <- cbind(dat[,5],dat[,6],l6)#1,2
data6 <- cbind(dat[,5],dat[,6],l7)#2,2
data7 <- cbind(dat[,5],dat[,6],l8)#3,2
data8 <- cbind(dat[,5],dat[,6],l9)#4,2
data9 <- cbind(dat[,5],dat[,6],l10)#5,2

#1,1
data0[data0[,1]==1,3]<-0 #2,1==0
data0[data0[,1]==2,3]<-0 #3,1==0
data0[data0[,1]==3,3]<-0 #4,1==0
data0[data0[,1]==4,3]<-0 #5,1==0
data0[data0[,2]==1,3]<-0 #,1==0

#2,1
data1[data1[,1]==0,3]<-0 #1,1==0
data1[data1[,1]==2,3]<-0 #3,1==0
data1[data1[,1]==3,3]<-0 #4,1==0
data1[data1[,1]==4,3]<-0 #5,1==0
data1[data1[,2]==1,3]<-0 #1,2==0

#3,1
data2[data2[,1]==0,3]<-0 #1,1==0
data2[data2[,1]==1,3]<-0 #2,1==0
data2[data2[,1]==3,3]<-0 #4,1==0
data2[data2[,1]==4,3]<-0 #5,1==0
data2[data2[,2]==1,3]<-0 #1,2==0

#4,1
data3[data3[,1]==0,3]<-0 #1,1==0
data3[data3[,1]==1,3]<-0 #2,1==0
data3[data3[,1]==2,3]<-0 #3,1==0
data3[data3[,1]==4,3]<-0 #5,1==0
data3[data3[,2]==1,3]<-0 #1,2==0

#5,1
data4[data4[,1]==0,3]<-0 #1,1==0
data4[data4[,1]==1,3]<-0 #2,1==0
data4[data4[,1]==2,3]<-0 #3,1==0
data4[data4[,1]==3,3]<-0 #4,1==0
data4[data4[,2]==1,3]<-0 #1,2==0

#1,2
data5[data5[,2]==0,3]<-0 #1,1==0
data5[data5[,1]==1,3]<-0 #2,1==0
data5[data5[,1]==2,3]<-0 #3,1==0
data5[data5[,1]==3,3]<-0 #4,1==0
data5[data5[,1]==4,3]<-0 #5,1==0

#2,2
```

```
  data6[data6[,2]==0,3]<-0 #1,1==0
  data6[data6[,1]==0,3]<-0 #2,1==0
  data6[data6[,1]==2,3]<-0 #3,1==0
  data6[data6[,1]==3,3]<-0 #4,1==0
  data6[data6[,1]==4,3]<-0 #5,1==0

  #3,2
  data7[data7[,2]==0,3]<-0 #1,1==0
  data7[data7[,1]==0,3]<-0 #2,1==0
  data7[data7[,1]==1,3]<-0 #3,1==0
  data7[data7[,1]==3,3]<-0 #4,1==0
  data7[data7[,1]==4,3]<-0 #5,1==0

  #4,2
  data8[data8[,2]==0,3]<-0 #1,2==0
  data8[data8[,1]==0,3]<-0 #1,1==0
  data8[data8[,1]==1,3]<-0 #2,1==0
  data8[data8[,1]==2,3]<-0 #3,1==0
  data8[data8[,1]==4,3]<-0 #5,1==0

  #5,2
  data9[data9[,2]==0,3]<-0 #1,1==0
  data9[data9[,1]==0,3]<-0 #2,1==0
  data9[data9[,1]==1,3]<-0 #3,1==0
  data9[data9[,1]==2,3]<-0 #4,1==0
  data9[data9[,1]==3,3]<-0 #5,1==0


  t0 <- sum(data0[,3])
  t1 <- sum(data1[,3])
  t2 <- sum(data2[,3])
  t3 <- sum(data3[,3])
  t4 <- sum(data4[,3])
  t5 <- sum(data5[,3])
  t6 <- sum(data6[,3])
  t7 <- sum(data7[,3])
  t8 <- sum(data8[,3])
  t9 <- sum(data9[,3])

  #-log(likelihood)
  Tfinal<-sum(t0)+sum(t1)+sum(t2)+sum(t3)+sum(t4)+sum(t5)+sum(t6)+sum(t7)+sum(t8)+sum(t9)

  return(-Tfinal)
}

lowerlim <- c(-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf,-Inf)
upperlim <- c(+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf,+Inf)


##PROBABILITY OF RESPONSE

##INTEGRAND
integrand<-function(Zint,meantreat,meanuntreat,mle)
{

  sigmahat=matrix(nrow=4,ncol=4)
  sigmahat[1,1]=(exp(mle[8]))^2
  sigmahat[2,1]=(2*inv.logit(mle[10])-1)*(exp(mle[8]))*exp(mle[9])
  sigmahat[3,1]=(2*inv.logit(mle[11])-1)*(exp(mle[8]))
  sigmahat[4,1]=(2*inv.logit(mle[12])-1)*(exp(mle[9]))
  sigmahat[1,2]=sigmahat[2,1]
  sigmahat[2,2]=(exp(mle[9]))^2
  sigmahat[3,2]=(2*inv.logit(mle[13])-1)*(exp(mle[9]))
  sigmahat[4,2]=(2*inv.logit(mle[14])-1)*(exp(mle[9]))
  sigmahat[1,3]=sigmahat[3,1]
  sigmahat[2,3]=sigmahat[3,2]
  sigmahat[3,3]=1
  sigmahat[4,3]=2*inv.logit(mle[15])-1
  sigmahat[1,4]=sigmahat[4,1]
  sigmahat[2,4]=sigmahat[4,2]
  sigmahat[3,4]=sigmahat[4,3]
  sigmahat[4,4]=1

  xtreat<-cbind(-meantreat[,1]+Zint[1], -meantreat[,2]+Zint[2], -meantreat[,3]+Zint[3], -meantreat[,4]+Zint[4])
```

```
  xuntreat<-cbind(-meanuntreat[,1]+Zint[1],-meanuntreat[,2]+Zint[2],-meanuntreat[,3]+Zint[3],-meanuntreat[,4]+Zint
      [4])

  pdftreat=dmvnorm(xtreat, mean=c(0,0,0,0),sigma=sigmahat)
  pdfuntreat=dmvnorm(xuntreat, mean=c(0,0,0,0),sigma=sigmahat)

  return(c(mean(pdftreat),mean(pdfuntreat)))
}


####PROBABILITY OF SUCCESS

probofsuccess<-function(mle,n,dat,eta)
{
  n=n

  meantreat=cbind(cbind(rep(1,n),rep(1,n),dat[,7])%*%c(mle[1:2],mle[20]),cbind(rep(1,n),rep(1,n),dat[,8])%*%c(mle
      [3:4],mle[21]),rep(1,n)*mle[5],
                  cbind(rep(1,n),rep(1,n))%*%mle[6:7])
  meanuntreat=cbind(cbind(rep(1,n),rep(0,n),dat[,7])%*%c(mle[1:2],mle[20]),cbind(rep(1,n),rep(0,n),dat[,8])%*%c(
      mle[3:4],mle[21]),rep(0,n)*mle[5],
                    cbind(rep(1,n),rep(0,n))%*%mle[6:7])

  #lower and upper bounds
  minmean1=min(c(meantreat[,1],meanuntreat[,1]))
  minmean2=min(c(meantreat[,2],meanuntreat[,2]))
  minmean3=min(c(meantreat[,3],meanuntreat[,3]))
  minmean4=min(c(meantreat[,4],meanuntreat[,4]))

  maxmean1=max(c(meantreat[,1],meanuntreat[,1]))
  maxmean2=max(c(meantreat[,2],meanuntreat[,2]))
  maxmean3=max(c(meantreat[,3],meanuntreat[,3]))
  maxmean4=max(c(meantreat[,4],meanuntreat[,4]))

  lower=c(qnorm(1e-15,minmean1,exp(mle[8])),qnorm(1e-15,minmean2,exp(mle[9])),qnorm(1e-15,minmean3,1),qnorm(1e-15,
      minmean4,1))
    upper=c(eta[1],eta[2],eta[3],eta[4])

  a=cuhre(4,2,integrand=integrand,meantreat=meantreat,meanuntreat=meanuntreat,
          mle=mle,lower=lower,upper=upper,flags=list(verbose=0,final=1,pseudo.random=0,mersenne.seed=NULL))

  #return(c(a$value[1]-a$value[2],a$value[1],a$value[2])) ##RISK DIFFERENCE
  return(c((log(a$value[1]/(1-a$value[1]))-log(a$value[2]/(1-a$value[2]))),a$value[1],a$value[2])) ##LOG-ODDS
  #return(log(a$value[1]/a$value[2])) ##LOG RISK RATIO
}


#### PARTIAL DERIVATIVES

partials<-function(mle,n,dat,eta)
{
  p=length(mle)
  fit1<-probofsuccess(mle,n,dat,eta)
  fit<-fit1[1]
  partials.augbin<-as.vector(rep(0,p))

  for(i in 1:p){
    valueupdate=mle
    valueupdate[i]=valueupdate[i]+0.000001
    updateprob=probofsuccess(valueupdate,n,dat,eta)[1]
    partials.augbin[i]=(updateprob-fit)/0.000001
  }

  return(c(partials.augbin,fit1))
}



### AUGMENTED BINARY METHOD
##BOX-COX TRANSFORM FOR CONTINUOUS COMPONENT

boxcoxtransform=function(y,lambda)
{
  return((y^lambda-1)/lambda)
}
```

```
345
346
347 ###INTEGRAND
348
349 integrand.aug<-function(acrn,meantreated,meanuntreated,Sigma,failure1,baseline1,baseline2)
350 {
351   n=length(baseline1)
352
353   f1treat=inv.logit(cbind(rep(1,n),baseline1,baseline2,rep(1,n))%*%failure1$coefficient)
354   f1untreat=inv.logit(cbind(rep(1,n),baseline1,baseline2,rep(0,n))%*%failure1$coefficient)
355
356   pdftreat=dnorm(-meantreated[,1]+acrn[1], mean=0,sd=Sigma)
357   pdfuntreat=dnorm(-meanuntreated[,1]+acrn[1], mean=0,sd=Sigma)
358
359   return(c(mean((1-f1treat)*pdftreat),mean((1-f1untreat)*pdfuntreat)))
360
361 }
362
363 ##PROBABILITY OF SUCCESS
364
365 probofsuccess.aug=function(continuous,baseline1,baseline2,failure1,dich)
366 {
367
368   n=length(baseline1)
369
370   meantreated=cbind(rep(1,n),rep(1,n),baseline1,baseline2)%*%continuous$coefficient
371
372   meanuntreated=cbind(rep(1,n),rep(0,n),baseline1,baseline2)%*%continuous$coefficient
373
374
375   #find lower and upper points for integration:
376   maxmean1=max(c(meantreated[,1],meanuntreated[,1]))
377   minmean1=min(c(meantreated[,1],meanuntreated[,1]))
378
379
380   #integrate
381
382   a=cuhre(1,2,integrand=integrand.aug,meantreated=meantreated,meanuntreated=meanuntreated,Sigma=summary(continuous
        )$sigma,failure1=failure1,baseline1=baseline1,baseline2=baseline2,
383           lower=qnorm(1e-08,minmean1,summary(continuous)$sigma),upper=dich,flags=list(verbose=0,final=1,pseudo.
                random=0,mersenne.seed=NULL))
384
385   #return(c(a$value[1]-a$value[2],a$value[1],a$value[2])) ### RISK DIFFERENCE
386   return(c((log(a$value[1]/(1-a$value[1]))-log(a$value[2]/(1-a$value[2]))),a$value[1],a$value[2])) ## LOG-ODDS
387
388 }
389
390
391 ###PARTIAL DERIVATIVES
392
393 get.partials<-function(continuous,baseline1,baseline2,failure1,dich)
394 {
395
396   fit=probofsuccess.aug(continuous,baseline1,baseline2,failure1,dich)
397   fit1=fit[1]
398   augbin.partials=as.vector(rep(0,8))
399
400
401   #split in to three separate models
402
403   #continuous model
404
405   for(i in 1:4)
406   {
407
408     valueupdate1=continuous
409     valueupdate1$coefficient[i]=valueupdate1$coefficient[i]+0.000001
410
411     updateprob=probofsuccess.aug(valueupdate1,baseline1,baseline2,failure1,dich)[1]
412
413     augbin.partials[i]=(updateprob-fit1)/0.000001
414
415   }
416
417   #failure model1
```

```
418
419   for(i in 1:4)
420   {
421
422     valueupdate2=failure1
423     valueupdate2$coefficient[i]=valueupdate2$coefficient[i]+0.000001
424
425     updateprob=probofsuccess.aug(continuous,baseline1,baseline2,valueupdate2,dich)[1]
426
427     augbin.partials[i+4]=(updateprob-fit1)/0.000001
428   }
429
430   return(c(augbin.partials,fit))
431 }
432
433
434 #### STANDARD BINARY METHOD
435
436 differenceinprob.binary=function(glm1,t,x1,x2)
437 {
438   #get fitted probs for each arm from model:
439
440   fittedvalues.control=as.double(inv.logit(cbind(rep(1,length(t[t==0])),rep(0,length(t[t==0])),x1[t==0],x2[t==0])%
        *%glm1$coef))
441
442   fittedvalues.exp=as.double(inv.logit(cbind(rep(1,length(t[t==1])),rep(1,length(t[t==1])),x1[t==1],x2[t==1])%*%
        glm1$coef))
443
444
445   return(c(log(mean(fittedvalues.exp,na.rm=T)/(1-mean(fittedvalues.exp,na.rm=T)))-log(mean(fittedvalues.control,na
        .rm=T)/(1-mean(fittedvalues.control,na.rm=T))),mean(fittedvalues.exp,na.rm=T),mean(fittedvalues.control,na.
        rm=T))) ###LOG-ODDS
446
447   #return(c(mean(fittedvalues.exp,na.rm=T)-mean(fittedvalues.control,na.rm=T), mean(fittedvalues.exp,na.rm=T),
        mean(fittedvalues.control,na.rm=T))) ### RISK DIFFERENCE
448   #return(log(mean(fittedvalues.exp,na.rm=T)/mean(fittedvalues.control,na.rm=T))) ### LOG RISK RATIO
449 }
450
451
452 ## PARTIAL DERIVATIVES
453
454 partialderivatives.binary=function(glm1,t,x1,x2)
455 {
456
457   value1=differenceinprob.binary(glm1,t,x1,x2)
458   value=value1[1]
459
460   partials=rep(0,4)
461
462   tempglm1=glm1
463   tempglm1$coef[1]=tempglm1$coef[1]+0.00001
464
465   partials[1]=(differenceinprob.binary(tempglm1,t,x1,x2)[1]-value)/0.00001
466
467   tempglm1=glm1
468   tempglm1$coef[2]=tempglm1$coef[2]+0.00001
469
470   partials[2]=(differenceinprob.binary(tempglm1,t,x1,x2)[1]-value)/0.00001
471
472   tempglm1=glm1
473   tempglm1$coef[3]=tempglm1$coef[3]+0.00001
474
475   partials[3]=(differenceinprob.binary(tempglm1,t,x1,x2)[1]-value)/0.00001
476
477   tempglm1=glm1
478   tempglm1$coef[4]=tempglm1$coef[4]+0.00001
479
480   partials[4]=(differenceinprob.binary(tempglm1,t,x1,x2)[1]-value)/0.00001
481
482   return(c(value,partials,value1[2],value1[3]))
483
484
485 }
486
487 ##### CALCULATE PROBABILITY OF SUCCESS USING LATENT VARIABLE METHOD
```

```
488
489 n=dim(dat)[1]
490
491 eta=c() ##SET DICHOTOMISATION THRESHOLDS BASED ON DATA
492 mlefit=optimx(X,f,lower=lowerlim,upper=upperlim,method="nlminb",dat=dat,eta=eta,control=list(rel.tol=1e-12))
493 mle<-coef(mlefit[1,])
494 hess<-attr(mlefit,"details")["nlminb",]$nhatend
495 mlecov=ginv(hess)
496 mlecov<-nearPD(mlecov)$mat
497 se<-sqrt(diag(mlecov))
498 part<-partials(mle,n,dat,eta)
499 mean<-part[22]
500 parts<-part[1:21]
501 variance=t(parts)%*%mlecov%*%parts
502 variance=variance[1,1]
503
504 CI<-c(mean-1.96*sqrt(variance),mean,mean+1.96*sqrt(variance),part[23],part[24])
505
506
507
508 ##AUGMENTED BINARY
509 dichotomisationthreshold=eta[1] ###SET BASED ON DATA
510 cont<-dat$Z1
511 dat$myresp<-ifelse(dat$Z2<=(eta[2]) & dat$Z3ord!=3 & dat$Z3ord!=4 & dat$Z4ord==0,0,1) ### SET BINARY RESPONSE
      CRITERIA BASED ON DATA
512 failure<-dat$myresp
513 baselinediseaseactivity<-dat$Z10
514 baseline2<-dat$Z20
515 arm<-dat$treat
516 patientid<-dat$id
517
518 #fit continuous model
519 continuousmodel=gls(Z1~treat+Z10+Z20,data=dat)
520
521 #first model - all patients:
522 failuremodel1=glm(myresp~Z10+Z20+treat,family="binomial",data=dat)
523
524 partial.aug=get.partials(continuousmodel,baselinediseaseactivity,baseline2,failuremodel1,
      dichotomisationthreshold)
525
526 mean.aug=partial.aug[9]
527 partials.aug=partial.aug[1:8]
528
529 covariance.aug=matrix(0,8,8)
530 covariance.aug[1:4,1:4]=continuousmodel$varBeta
531 covariance.aug[5:8,5:8]=summary(failuremodel1)$cov.unscaled
532 variance.aug=t(partials.aug)%*%covariance.aug%*%partials.aug
533
534 #confidence interval for difference in mean probability of success
535 CI.augbin=c(mean.aug-1.96*sqrt(variance.aug),mean.aug,mean.aug+1.96*sqrt(variance.aug),partial.aug[10],partial.
      aug[11])
536
537
538 ###STANDARD BINARY
539
540 dat$resp<-ifelse(dat$Z1<=(eta[1]) & dat$Z2<=(eta[2]) & dat$Z3ord!=3 & dat$Z3ord!=4 & dat$Z4ord==0, 1,0) ##SET
      BASED ON DATA
541 success.binary=dat$resp
542
543 glm1=glm(success.binary~treat+Z10+Z20,family="binomial")
544
545 partial.binary=partialderivatives.binary(glm1,treat,Z10,Z20)
546 mean.binary=partial.binary[1]
547 partials.binary=partial.binary[2:5]
548 covariance=summary(glm1)$cov.unscaled
549 var.binary=t(partials.binary)%*%covariance%*%partials.binary
550
551 CI.binary=c(mean.binary-1.96*sqrt(var.binary),mean.binary,mean.binary+1.96*sqrt(var.binary),partial.binary[6],
      partial.binary[7])
```