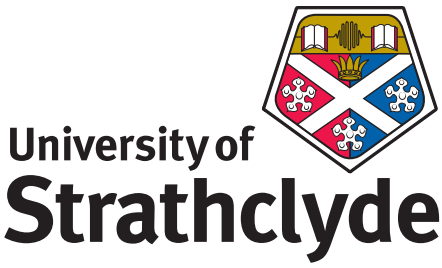


Detection of Safety Signals in Randomised Controlled Trials

Raymond Carragher

Department of Mathematics and Statistics

University of Strathclyde

Glasgow

2017

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

I would like to thank to my supervisors, Professor Chris Robertson, University of Strathclyde, and Doctor Ian Bradbury, Frontier Science (Scotland) Ltd., for their support and guidance. I would also like to thank Thérèse O'Donnell for her support.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (award reference 1521741) and Frontier Science (Scotland) Ltd.

Abstract

The occurrence, severity, and duration of patient adverse events are routinely recorded during randomised clinical trials. This data is used by a trial's Data Monitoring Committee to make decisions regarding the safety of a treatment and may lead to the alteration or discontinuation of a trial if real safety issues are detected. There are many different types of adverse event and the statistical analysis of this data, particularly with regard to hypothesis testing, must take into account potential multiple comparison issues. Unadjusted hypothesis tests may lead to large numbers of false positive results, but simple adjustments are generally too conservative. In addition, the anticipated effect sizes of adverse events in clinical trials are generally small and consequently the power to detect such effects is low.

A number of recent classical and Bayesian methods, which use groupings of adverse events, have been proposed to address this problem. We illustrate and compare a number of these approaches, and investigate if their use of a common underlying model, which involves groupings of adverse events by body-system or System Organ Class, is useful in detecting adverse events associated with treatments. For data where this type of grouped approach is appropriate, the methods considered are shown to correctly flag more adverse event effects than standard approaches, while maintaining control of the overall error rate.

While controlling for multiple types of adverse event, these proposed methods do not take into account event timings or patient exposure time, and are more suited to end of trial analysis. In order to address the desire for the early detection of safety issues in clinical trials a number of Bayesian methods are introduced to analyse the accumulation of adverse events as the trial progresses, taking into account event timing, patient time in study, and body-system. These methods are suitable for use at interim trial safety analyses. The models which performed best were those that had a common body-system dependence over the duration of the trial.

Contents

List of Tables **xiii**

List of Figures **xxi**

1 Safety in the Context of Clinical Trials **1**

1.1 Introduction 1
1.2 Overview 3
1.3 Data Monitoring Committees 5
1.4 ICH Guidelines 6
1.4.1 Long-Term Treatment 6
1.4.2 Pharmacovigilance 7
1.4.3 Reporting and Statistical Analysis 8
1.4.4 Statistical Principles for Clinical Trials 8
1.5 Clinical Trial Outcomes 9
1.5.1 Primary and Secondary Clinical Trial Outcomes 9
1.5.2 Safety Outcomes in Clinical Trials 10
1.6 Recording Data in Clinical Trials 11
1.6.1 Medical Dictionaries 12
1.6.2 Recording and Classifying Adverse Events 13
1.7 Safety Analyses in Clinical Trials 15
1.8 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer 17
1.8.1 HER2 Positive Breast Cancer 17
1.8.2 Trial Overview 17
1.8.2.1 Primary and Secondary Outcomes 18
1.8.2.2 Data Monitoring Committee Role 18
1.8.2.3 Interim and Final Analyses 19
1.8.2.4 Statistical Methods 19
1.8.2.5 Study Populations and Randomisation 19

1.8.2.6 Safety 20
1.8.3 Trial Conduct 21
1.8.3.1 Data Collection and Recording 21
1.8.4 Interim Analysis and Trial Changes 23
1.8.5 Efficacy Results 24
1.8.6 Safety Analysis 24
1.8.7 Interim Analysis Safety Report 25
1.8.8 Final Safety Report 27
1.9 Grouping of Adverse Events and the Body-System Approach to Safety Analysis 30
1.10 Research Questions 36

2 Error Controlling Procedures in Safety Analysis 38
2.1 Introduction 38
2.2 ICH Guidelines 39
2.3 Error Rate Controlling Procedures 41
2.3.1 The Familywise Error Rate 42
2.3.1.1 Controlling the FWER 42
2.3.2 False Discovery Rate 43
2.3.2.1 Control of the FDR by the Benjamini-Hochberg Procedure 44
2.3.2.2 Extension to Positive Regression Dependent Test Statistics 44
2.3.2.3 Adjusted BH p-values 45
2.3.2.4 Further Extensions of False Discovery Rate Control 45
2.3.2.5 Increasing the Power through Grouped False Discovery Rate Methods 47
2.4 Double False Discovery Rate 50
2.4.1 Notation 50
2.4.2 Original Double FDR Approach 50
2.4.3 Double FDR Approach Procedure (2012) 52
2.5 Group Benjamini-Hochberg 54
2.5.1 Notation 54
2.5.2 Group Benjamini-Hochberg Procedure 54
2.6 Comparison Between DFDR and GBH 56
2.7 Discussion 58

3 Modelling Methods for Safety Analysis **60**
3.1 Introduction . . . 60
3.2 Survival Analysis Models . . . 63
3.3 Recurrent Event Analysis . . . 65
3.3.1 The Non-Parametric Estimate of the Mean Cumulative Function . . . 68
3.3.1.1 Using the MCF to Analyse Adverse Event Data from Randomised Clinical Trials . . . 70
3.3.1.2 Further Discussion . . . 73
3.4 Random and Mixed Effects Models . . . 74
3.5 Longitudinal Analysis . . . 76
3.6 Bayesian Approaches to Clinical Trials . . . 77
3.6.1 Bayesian Models for Adverse Event Incidence . . . 78
3.6.2 Berry and Berry Model . . . 81
3.6.2.1 Model Notation . . . 81
3.6.2.2 Data Model . . . 81
3.6.2.3 First Level . . . 82
3.6.2.4 Second Level . . . 82
3.6.2.5 Third Level . . . 83
3.6.2.6 Modelling the Body-System . . . 83
3.6.2.7 Discussion . . . 85
3.6.3 Multivariate Bayesian Logistic Regression . . . 86
3.6.3.1 The Model . . . 88
3.6.3.2 Interpretation of Coefficients . . . 89
3.6.3.3 Model Fitting . . . 90
3.6.3.4 Discussion . . . 90
3.6.4 Further Discussion . . . 92
3.7 Discussion . . . 96
3.7.1 Bayesian Models . . . 96
3.7.2 Frequentist Models . . . 96
3.7.3 Further Extensions of the Above . . . 97

4 Adverse Event Detection in GSK EGF100151 **98**
4.1 Introduction . . . 98
4.2 End of Trial Safety Data . . . 99
4.3 Berry and Berry Model . . . 100
4.3.1 Model with Point-Mass (c212.BB) . . . 100

4.3.2 Model without Point-Mass (c212.1a) . . . 102
4.4 Error Controlling Procedures . . . 104
4.4.1 Double False Discovery Rate . . . 104
4.4.2 Group Benjamini-Hochberg . . . 108
4.5 Discussion . . . 109

5 Simulation Study **111**
5.1 Introduction . . . 111
5.2 Simulated Adverse Event Incidence Data Model . . . 113
5.3 Trial Layout and Body-Systems . . . 114
5.4 Simulation Definition . . . 116
5.4.1 Simulation Data Model Parameters . . . 116
5.4.2 Simulation Structure and Adverse Events with Raised Treatment Rates . . . 122
5.5 Results . . . 125
5.5.1 Individual Simulations and Model Parameter Estimation . . . 125
5.5.1.1 Equal Treatment and Control Event Rates (TDM1, SIM6) . . . 125
5.5.1.2 Low Increase in Treatment Event Rate (TDM15, SIM6) . . . 127
5.5.1.3 Medium Increase in Treatment Event Rate (TDM15, SIM84) . . . 131
5.5.1.4 High Increase in Treatment Event Rate (TDM15, SIM162) . . . 135
5.5.1.5 Assessment . . . 139
5.5.2 Overall Results . . . 140
5.5.2.1 Treatment Arm Effect Simulations . . . 140
5.5.2.2 Repeated Simulations . . . 145
5.6 Conclusions . . . 149

6 Methods for Interim Data Analysis **152**
6.1 Introduction . . . 152
6.2 Adverse Event Data . . . 152
6.3 Analysing Interim Data . . . 154
6.3.1 Censoring, Observation, Terminal Events, and Time Scales 155
6.3.2 Body-System . . . 156
6.3.3 Detecting Raised Adverse Event Rates . . . 156

6.4 Modelling Approach . . . 157
6.4.1 Patient Level Models . . . 157
6.4.2 Summary Models . . . 158
6.4.3 Incidence and Recurrent Event Analysis . . . 158
6.4.4 Sequential Analysis of Interim Data . . . 159
6.5 Models for Interim Adverse Event Analysis . . . 159
6.5.1 Poisson Process Models for Adverse Events . . . 161
6.5.2 Independence and Exchangeability . . . 162
6.5.3 Modelling Body-System and Longitudinal Relationships . . . 163
6.6 Poisson Bayesian Models: Three-Level Hierarchies . . . 165
6.6.1 BB_{30} Poisson Point-mass Model (level 0) . . . 166
6.6.2 BB_{31} Poisson Point-mass Model (level 1) . . . 167
6.6.3 BB_{32} Poisson Point-mass Model (level 2) . . . 168
6.6.4 Poisson Models Without Point-mass . . . 168
6.6.5 Summary . . . 169
6.7 Poisson Bayesian Models: Two-Level Hierarchy . . . 169
6.7.1 BB_{20} Poisson Point-mass Model (Level 0) . . . 170
6.7.2 Choice of Prior for $\pi_{b,h}$ . . . 171
6.7.3 BB_{21} Poisson Point-mass Model (level 1) . . . 171
6.7.4 Poisson Models Without Point-mass . . . 172
6.8 Flagging Adverse Events as Having Raised Treatment Rates . . . 172

7 Demonstration Interim Analyses 174
7.1 Introduction . . . 174
7.2 Trial Structure . . . 176
7.2.1 Body-Systems and Adverse Event Severity . . . 177
7.2.2 Intervals . . . 178
7.2.3 Trial Types . . . 178
7.3 Trial Simulation Parameters . . . 179
7.3.1 Background Trial Adverse Event Rates . . . 179
7.3.2 Increased Treatment Rates . . . 179
7.4 Adverse Events with Raised Treatment Rates . . . 180
7.5 Patient Recruitment . . . 183
7.6 Flagging Adverse Events as Having Raised Treatment Rates . . . 183
7.7 Demonstration Analysis Results . . . 184
7.7.1 Trial II(a) Incidence Event Analysis . . . 184
7.7.1.1 Cumulative Adverse Event Incidence Totals . . . 184

7.7.1.2 Interim and Final Adverse Event Incidence Counts 187
7.7.1.3 Model Analyses: Model 1a (No Point-mass) 189
7.7.1.4 Model Analyses: Model BB (Point-mass) 196
7.7.1.5 Assessing the Adverse Event Rates by Posterior Distribution 200
7.7.2 Trial II(a) Total Adverse Event Analysis 207
7.7.2.1 Cumulative Adverse Event Totals 207
7.7.2.2 Interim and Final Analyses Adverse Event Counts 210
7.7.2.3 Model Analyses: Model 1a (No Point-mass) 211
7.7.2.4 Model Analyses: Model BB (Point-mass) 216
7.7.2.5 All Events versus Event Incidence Analysis 219
7.7.2.6 Model Parameter Estimation 220
7.7.3 Overall Results 223
7.7.3.1 Incidence Event Analysis 223
7.7.3.2 Total Event Analysis 229
7.7.3.3 Model Comparison 235
7.8 Sensitivity Analysis 240
7.8.1 Changing the Flagging Threshold 240
7.8.1.1 Incidence Data (All Trials) 240
7.8.2 Missing Data 245
7.8.2.1 Missing Data Case 1 246
7.8.2.2 Missing Data Case 2 247
7.8.2.3 Missing Data Case 3 248
7.8.2.4 Missing Data Case 4 250
7.8.2.5 Discussion 251
7.8.3 Lower Background Trial Adverse Event Rates 252
7.8.3.1 Results 254
7.8.4 Mixed Adverse Event Background Rates 258
7.8.4.1 Results 259
7.9 Discussion 263
7.10 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer 266
7.11 Conclusions 268

8 Conclusions **270**

A Software Implementations and Model Fitting Algorithms 278
A.1 MCMC Approach to Bayesian Model Fitting . . . 279
A.2 MCMC Sampling Algorithms . . . 279
A.3 Default Simulation Parameter Values . . . 280
A.4 Initial Values for MCMC Chains . . . 282
A.4.1 Top-Level Parameters . . . 282
A.4.1.1 End of Trial Methods . . . 282
A.4.1.2 Interim Analysis Methods . . . 283
A.4.2 Hyperparameters . . . 284
A.5 Convergence Diagnostics and Summary Statistics . . . 285
A.6 Inverse Gamma Distribution . . . 285
A.7 Direct Error Controlling Methods . . . 285
A.8 BUGs Models . . . 286
A.8.1 c212.BB . . . 286
A.8.2 c212.1a . . . 287

B Joint Distributions and Complete Conditional Distributions 288
B.1 Distributions . . . 288
B.2 General Results . . . 288
B.3 Model c212.1a . . . 292
B.3.1 Complete Conditional Distributions . . . 294
B.4 Model c212.BB . . . 297
B.4.1 Complete Conditional Distributions . . . 298
B.5 Model $1a_{20}$ . . . 301
B.5.1 Complete Conditional Distributions . . . 302
B.6 Model $1a_{21}$ . . . 304
B.6.1 Complete Conditional Distributions . . . 305
B.7 Model BB_{20} . . . 307
B.7.1 Complete Conditional Distributions . . . 308
B.8 Model BB_{21} . . . 310
B.8.1 Complete Conditional Distributions . . . 310
B.9 Model $1a_{30}$ . . . 314
B.9.1 Complete Conditional Distributions . . . 315
B.10 Model $1a_{31}$ . . . 319
B.10.1 Complete Conditional Distributions . . . 320
B.11 Model $1a_{32}$ . . . 323
B.11.1 Complete Conditional Distributions . . . 324

B.12 Model BB_{30} . . . 327
B.12.1 Complete Conditional Distributions . . . 329
B.13 Model BB_{31} . . . 333
B.13.1 Complete Conditionals Distributions . . . 335
B.14 Model BB_{32} . . . 339
B.14.1 Complete Conditional Distributions . . . 341

C Model Tuning and Monitoring Convergence 346
C.1 Approach to Determining Approximate Convergence . . . 346
C.2 Tuning Model Fitting Parameters . . . 348
C.3 Simulation Study . . . 349
C.3.1 Individual Simulations . . . 349
C.3.2 All Simulations . . . 351
C.4 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer . . . 351
C.5 Demonstration Interim Analyses . . . 353
C.6 Sensitivity Analysis . . . 354
C.6.1 Low Background Event Rate . . . 354
C.6.2 Mixed Background Event Rates . . . 356
C.7 Summary . . . 356

D Grouped FDR Controlling Methods 357
D.1 Controlling Error Rates for DFDR . . . 357
D.1.1 Large Body-System Properties for the DFDR Under Independence Assumptions . . . 357
D.2 Comparison of DFDR and GBH using Simulated Data . . . 359
D.2.1 Simulation Definition . . . 359
D.2.1.1 Adverse Events with Raised Treatment Rates . . . 360
D.2.2 Results Summary . . . 361
D.2.3 Conclusions . . . 364

E Trial Adverse Event Simulation 365
E.1 Simulating Patient Recruitment . . . 365
E.2 Simulating Adverse Event Data . . . 365

F Table of Methods 367

References 386

List of Tables

1.1 NCI CTCAE: Adverse event severities. 14
1.2 NCI CTCAE: Adverse event severity grade definitions for *Hyperkalemia*. 15
1.3 Trial EGF100151: Clinical study reports. 18
1.4 Trial EGF100151: Population randomisation. 20
1.5 Trial EGF100151: Adverse events of special interest. 21
1.6 Trial EGF100151: Incidence of the most common adverse events (all grades), 15 November 2005 25
1.7 Trial EGF100151: Serious adverse events experienced by more than 1% of subjects, final clinical study report. 27
1.8 Trial EGF100151: Most common adverse events (all grades), final clinical study report. 28
1.9 Trial EGF100151: The six most common adverse events by grade on the *lapatinib and capecitabine* arm, final clinical study report. . . 29
1.10 Trial EGF100151: The six most common adverse events, and rash, by grade, on the *capecitabine* arm, final clinical study report. 29
1.11 Trial EGF100151: System organ classes and adverse event totals. . . 31
1.12 Trial EGF100151: Adverse events significant at the 5% level, final clinical study report. 32
1.13 Trial EGF100151: Number of subjects who experienced at least one adverse event in a system organ class. 35

2.1 Null and Alternative Hypotheses: actual status and test outcomes. . 42
2.2 Comparison of DFDR and GBH procedures for controlling the FDR. 57

4.1 Trial EGF100151: Methods applied to safety data. 99
4.2 Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry point-mass model (c212.BB), `R` implementation. . 101

4.3 Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry point-mass model (c212.BB), `OpenBUGS` implementation. . . . 102
4.4 Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry model without point-mass (c212.1a), `R` implementation. . . . 103
4.5 Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry model without point-mass (c212.1a), `OpenBUGS` implementation. . . . 104
4.6 Trial EGF100151: adverse events flagged as significant by the Double False Discover Rate. . . . 105
4.7 Trial EGF100151: Adverse events flagged at 10% significance level by the Double False Discover Rate (low counts removed). . . . 107
4.8 Trial EGF100151: Adverse events flagged as significant by the Group Benjamini-Hochberg procedure. . . . 108
4.9 Trial EGF100151: Adverse events flagged as significant by the Group Benjamini-Hochberg procedure, with re-weighted p-values. . . . 108
4.10 Trial EGF100151: Adverse events flagged as significant at the 5% level by the Group Benjamini-Hochberg procedure (low counts removed). . . . 109
4.11 Trial EGF100151: Adverse events flagged as significant at the 10% level by the Group Benjamini-Hochberg procedure (low counts removed). . . . 109
4.12 Trial EGF100151: Adverse events of interest flagged by method. . . . 110

5.1 Methods used in the Simulation Study. . . . 112
5.2 Simulation body-systems and numbers of adverse events. . . . 115
5.3 Trial Sizes and numbers of participants used in simulation study. . . . 115
5.4 Simulation study: overall mean and treatment arm parameter values. 116
5.5 Simulations with trial arm and body-system effects only. . . . 117
5.6 Simulations with trial arm and adverse event effects only. . . . 118
5.7 Simulations with trial arm, body-system, and adverse event effects. 119
5.8 Simulations with trial arm and body-system only effects: fixed parameter values. . . . 120
5.9 Simulations with trial arm and adverse event only effects: fixed parameter values. . . . 120

5.10 Simulations with trial arm, body-system, and adverse event effects: fixed parameter values. 121
5.11 Simulation Study: Parameter ranges. 122
5.12 Treatment Arm Effect Simulations: Total numbers of adverse events. 123
5.13 Repeated Simulations (no treatment arm effects): Total numbers of adverse events. 124
5.14 Individual simulation cases. 126
5.15 TDM1, SIM6: Type-I error rates (number of events incorrectly declared significant) by trial size. 126
5.16 TDM15, SIM6: Large trial results. 128
5.17 TDM15, SIM6: Medium trial results. 129
5.18 TDM15, SIM6: Small trial results. 129
5.19 TDM15, SIM84: Large trial results. 132
5.20 TDM15, SIM84: Medium trial results. 133
5.21 TDM15, SIM84: Small trial results. 133
5.22 TDM15, SIM162: Large trial results. 136
5.23 TDM15, SIM162: Medium trial results. 137
5.24 TDM15, SIM162: Small trial results. 137
5.25 Treatment Arm Effect Simulations: Large trial results. 141
5.26 Treatment Arm Effect Simulations: Medium trial results. 142
5.27 Treatment Arm Effect Simulations: Small trial results. 142
5.28 Treatment Arm Effect simulations: Combined results. 143
5.29 Repeated Simulations: Large trial results. 145
5.30 Repeated Simulations: Medium trial results. 146
5.31 Repeated Simulations: Small trial results. 146
5.32 Repeated Simulations: Combined results. 147

6.1 Hierarchical methods for interim analyses. 165

7.1 Demonstration Analysis: Common trial details. 176
7.2 Demonstration Analysis: Planned safety reviews. 176
7.3 Demonstration Analysis: Simulation body-systems and numbers of adverse events. 177
7.4 Demonstration Analysis: Adverse event severity probabilities. 178
7.5 Demonstration Analysis: Trial types. 178
7.6 Demonstration Analysis: Background trial adverse event rate. . . . 179
7.7 Demonstration Analysis: Body-systems and intervals with increased treatment rates. 180

7.8 Demonstration Analysis: Adverse event and interval totals per trial. 181
7.9 Demonstration Analysis: Adverse event and interval totals, all trials combined. 182
7.10 Demonstration Analysis: Simulated trial patient enrolment totals. . 183
7.11 Demonstration Analysis: Trial II(a) total adverse event incidence by trial arm at each interim safety analysis. 188
7.12 Demonstration Analysis: Trial II(a) total adverse event incidence for Bdy-sys_3 by trial arm at each interim safety analysis. 188
7.13 Demonstration Analysis: Trial II(a) model 1a severity 1+ adverse event incidence analysis results for the first 3 safety analyses. 189
7.14 Demonstration Analysis: Trial II(a) model 1a severity 1+ adverse event incidence analysis results for the final 4 safety analyses. . . . 190
7.15 Demonstration Analysis: Trial II(a) model 1a severity 3+ adverse event incidence analysis results for the first 6 safety analyses. 191
7.16 Demonstration Analysis: Trial II(a) model 1a severity 3+ adverse event incidence analysis results for the final safety analysis. 192
7.17 Demonstration Analysis: Trial II(a) model BB severity 1+ adverse event incidence analysis results for the first 3 safety analyses. 196
7.18 Demonstration Analysis: Trial II(a) model BB severity 1+ adverse event incidence analysis results for the final 4 safety analyses. . . . 197
7.19 Demonstration Analysis: Trial II(a) model BB severity 3+ adverse event incidence analysis results for the first 6 safety analyses. 198
7.20 Demonstration Analysis: Trial II(a) model BB severity 3+ adverse event incidence analysis results for the final safety analysis. 199
7.21 Demonstration Analysis: Trial II(a) total adverse events by trial arm at each interim safety analysis. 210
7.22 Demonstration Analysis: Trial II(a) total adverse events by trial arm at each interim safety analysis for Bdy-sys_3. 210
7.23 Demonstration Analysis: Trial II(a) model 1a severity 1+ total adverse event analysis results for the first 5 safety analyses. 211
7.24 Demonstration Analysis: Trial II(a) model 1a severity 1+ total adverse event analysis results for the final 2 safety analyses. 212
7.25 Demonstration Analysis: Trial II(a) model 1a severity 3+ total adverse event analysis results for the first 3 safety analyses. 212
7.26 Demonstration Analysis: Trial II(a) model 1a severity 3+ total adverse event analysis results for the final 4 safety analyses. 213

7.27 Demonstration Analysis: Trial II(a) model BB severity 1+ total adverse event analysis results for the first 5 safety analyses. 216
7.28 Demonstration Analysis: Trial II(a) model BB severity 1+ total adverse event analysis results for the final 2 safety analyses. 217
7.29 Demonstration Analysis: Trial II(a) model BB severity 3+ total adverse event analysis results for the first 3 safety analyses. 217
7.30 Demonstration Analysis: Trial II(a) model BB severity 3+ total adverse event analysis results for the final 4 safety analyses. 218
7.31 Demonstration Analysis: All trials model 1a severity 1+ adverse event incidence results for the first 2 safety analyses. 223
7.32 Demonstration Analysis: All trials model 1a severity 1+ adverse event incidence results for the final 5 safety analyses. 224
7.33 Demonstration Analysis: All trials model 1a severity 3+ adverse event incidence results for the first 6 safety analyses. 225
7.34 Demonstration Analysis: All trials model 1a severity 3+ adverse event incidence results for the final safety analysis. 226
7.35 Demonstration Analysis: All trials model BB severity 1+ adverse event incidence results for the first 2 safety analyses. 226
7.36 Demonstration Analysis: All trials model BB severity 1+ adverse event incidence results for the final 5 safety analyses. 227
7.37 Demonstration Analysis: All trials model BB severity 3+ adverse event incidence results for the first 6 safety analyses. 228
7.38 Demonstration Analysis: All trials model BB severity 3+ adverse event incidence results for the final safety analysis. 229
7.39 Demonstration Analysis: All trials model 1a severity 1+ total adverse event results for the first 2 safety analyses. 229
7.40 Demonstration Analysis: All trials model 1a severity 1+ total adverse event results for the final 5 safety analyses. 230
7.41 Demonstration Analysis: All trials model 1a severity 3+ total adverse event results for the first 6 safety analyses. 231
7.42 Demonstration Analysis: All trials model 1a severity 3+ total adverse event results for the final safety analysis. 232
7.43 Demonstration Analysis: All trials model BB severity 1+ total adverse event results for the first 4 safety analyses. 232
7.44 Demonstration Analysis: All trials model BB severity 1+ total adverse event results for the final 3 safety analyses. 233

7.45 Demonstration Analysis: All trials model BB severity 3+ total adverse event results for the first 2 safety analyses. 233
7.46 Demonstration Analysis: All trials model BB severity 3+ total adverse event results for the final 5 safety analyses. 234
7.47 Sensitivity Analysis: Changed threshold model $1a_{21}$ results (all trials), severity 1+ events. 240
7.48 Sensitivity Analysis: Changed threshold model $1a_{21}$ results (all trials), severity 3+ events. 241
7.49 Sensitivity Analysis: Changed threshold model BB_{21} results (all trials), severity 1+ events. 241
7.50 Sensitivity Analysis: Changed threshold model BB_{21} results (all trials), severity 3+ events. 241
7.51 Sensitivity Analysis: Missing data scenarios. 245
7.52 Sensitivity Analysis: Missing data case 1, model $1a_{21}$ results (all trials), severity 1+. 246
7.53 Sensitivity Analysis: Missing data case 1, model $1a_{21}$ results (all trials), severity 3+. 246
7.54 Sensitivity Analysis: Missing data case 1, model BB_{21} results (all trials), severity 1+. 246
7.55 Sensitivity Analysis: Missing data case 1, model BB_{21} results (all trials), severity 3+. 247
7.56 Sensitivity Analysis: Missing data case 2, model $1a_{21}$ results (all trials), severity 1+. 247
7.57 Sensitivity Analysis: Missing data case 2, model $1a_{21}$ results (all trials), severity 3+. 247
7.58 Sensitivity Analysis: Missing data case 2, model BB_{21} results (all trials), severity 1+. 248
7.59 Sensitivity Analysis: Missing data case 2, model BB_{21} results (all trials), severity 3+. 248
7.60 Sensitivity Analysis: Missing data case 3, model $1a_{21}$ results (all trials), severity 1+. 248
7.61 Sensitivity Analysis: Missing data case 3, model $1a_{21}$ results (all trials), severity 3+. 249
7.62 Sensitivity Analysis: Missing data case 3, model BB_{21} results (all trials), severity 1+. 249
7.63 Sensitivity Analysis: Missing data case 3, model BB_{21} results (all trials), severity 3+. 249

7.64 Sensitivity Analysis: Missing data case 4, model $1a_{21}$ results (all trials), severity 1+ . . . 250
7.65 Sensitivity Analysis: Missing data case 4, model $1a_{21}$ results (all trials), severity 3+. . . . 250
7.66 Sensitivity Analysis: Missing data case 4, model BB_{21} results (all trials), severity 1+. . . . 250
7.67 Sensitivity Analysis: Missing data case 4, model BB_{21} results (all trials), severity 3+. . . . 251
7.68 Sensitivity Analysis: Low background rate adverse event rate. . . . 253
7.69 Sensitivity Analysis: Low background rate trial patient enrolment totals. . . . 253
7.70 Sensitivity Analysis: Low rate events, body-systems and intervals with increased treatment rates. . . . 253
7.71 Sensitivity Analysis: Low rate severity 1+ adverse event incidence early trial results. . . . 255
7.72 Sensitivity Analysis: Low rate severity 1+ adverse event incidence end of trial results. . . . 256
7.73 Sensitivity Analysis: Low rate severity 3+ adverse event incidence early trial results. . . . 257
7.74 Sensitivity Analysis: Low rate severity 3+ adverse event incidence end of trial results. . . . 258
7.75 Sensitivity Analysis: Mixed event rates, adverse events with higher background rates. . . . 258
7.76 Sensitivity Analysis: Mixed event rates, patient enrolment totals. . . . 259
7.77 Sensitivity Analysis: Mixed event rates, body-systems and intervals with increased treatment rates. . . . 259
7.78 Sensitivity Analysis: Mixed rate severity 1+ adverse event incidence early trial results. . . . 260
7.79 Sensitivity Analysis: Mixed rate severity 1+ adverse event incidence end of trial results. . . . 261
7.80 Sensitivity Analysis: Mixed rate severity 3+ adverse event incidence early trial results. . . . 262
7.81 Sensitivity Analysis: Mixed rate severity 3+ adverse event incidence end of trial results. . . . 263
7.82 Demonstration Analysis: Trial II(a) model 1a level 1 parameter estimates. . . . 266

7.83 Trial EGF100151: Top 10 adverse events by posterior probability for $1a_{21}$. 267
7.84 Trial EGF100151: Top 10 adverse events by posterior probability for BB_{21}. 268

A.1 Methods implemented in the package `c212`. 278
A.2 Model parameters with non-standard distributions. 280
A.3 Model MCMC defaults. 280
A.4 Global MCMC parameters for end of trial models. 281
A.5 Global MCMC parameters for the interim analysis models. 281
A.6 End of trial methods: top-level initial value generation for MCMC simulation. 283
A.7 Interim analysis methods: top-level parameter initial value generation for MCMC simulation. 283
A.8 End of trial methods: hyper-parameter initial value generation for MCMC simulation. 284
A.9 Interim analysis methods: hyper-parameter initial value generation for MCMC simulation. 284

B.1 Probability distributions. 288

D.1 DFDR representative p-value as body-system size increases for body-systems containing only true null hypotheses. 359
D.2 Large body-system simulation parameter values. 360
D.3 Adverse events totals - all simulations 361
D.4 Overall results. 362
D.5 LBS1 Results. 362
D.6 LBS2 Results. 363
D.7 LBS3 Results. 363
D.8 LBS4 Results. 363
D.9 DFDR power as body-system size increases. 364

E.1 Interim analysis trial simulation parameters. 365

F.1 Table of methods. 371

List of Figures

1.1 WHO-ART hierarchy for *Heart rate and rhythm disorders.* 13
1.2 Trial EGF100151: Non-serious adverse event recording form. 22
1.3 Trial EGF100151: Serious adverse event recording form. 23
1.4 EGF10015: Proportion of subjects who experienced an adverse event by system organ class. 33

3.1 Adverse events grouped by body-system. 79
3.2 Directed acyclic graph for the Berry and Berry Model. 84

4.1 GSK EGF100151 - End of Trial Data. 100

5.1 TDM15, SIM6: Large trial parameter estimates. 130
5.2 TDM15, SIM6: Medium trial parameter estimates. 130
5.3 TDM15, SIM6: Small trial parameter estimates. 131
5.4 TDM15, SIM84: Large trial parameter estimates. 134
5.5 TDM15, SIM84: Medium trial parameter estimates. 134
5.6 TDM15, SIM84: Small trial parameter estimates. 135
5.7 TDM15 - SIM162: Large trial parameter estimates. 138
5.8 TDM15 - SIM162: Medium trial parameter estimates. 138
5.9 TDM15 - SIM162: Small trial parameter estimates. 139
5.10 Treatment Arm Effect Simulations: Correctly flagged adverse events and Type-I error rates. 144
5.11 Repeated Simulations: Correctly flagged adverse events and Type-I error rates 148

6.1 Sample raw trial data. 153
6.2 Sample summary trial data (all event severities). 161

7.1 Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to day 720 of the trial. Severity 1+ events on left, severity 3+ events on the right. 185
7.2 Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to day 1800 of the trial. Severity 1+ events on left, severity 3+ events on the right. 186
7.3 Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to the end of trial. Severity 1+ events on left, severity 3+ events on the right. 187
7.4 Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged. 193
7.5 Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Proportion of severity 3+ adverse events with raised treatment rates correctly flagged. 194
7.6 Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Type-I error rates severity 1+ events. 194
7.7 Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Type-I error rates severity 3+ events. 195
7.8 Demonstration Analysis: Trial II(a) model BB adverse event incidence analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged. 199
7.9 Demonstration Analysis: Trial II(a) model BB adverse event incidence analysis. Proportion of severity 3+ adverse events with raised treatment rates correctly flagged. 200
7.10 Model $1a_{31}$: Top 10 adverse events by posterior probability at the end of trial (day 2520). 202
7.11 Model BB_{21}: Top 10 adverse events by posterior probability at the end of trial (day 2520). 203
7.12 Model $1a_{31}$: Bdy-sys3 and Bdy-sys_1 end of trial rates. 205
7.13 Model BB_{21}: Bdy-sys3 and Bdy-sys_1 end of trial rates. 206
7.14 Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval up to day 360. Severity 1+ events on left, severity 3+ events on the right. 207
7.15 Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval up to day 1440. Severity 1+ events on left, severity 3+ events on the right. 208

7.16 Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval to end of trial. Severity 1+ events on left, severity 3+ events on the right. 209
7.17 Demonstration Analysis: Trial II(a) model 1a total adverse event analysis. Proportion of adverse events with raised treatment rates correctly flagged (severity 1+ on the top, severity 3+ on the bottom). 214
7.18 Demonstration Analysis: Trial II(a) model 1a total adverse events analysis. Type-I error rates (severity 1+ on the top, severity 3+ on the bottom) 215
7.19 Demonstration Analysis: Trial II(a) model BB total adverse event analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged (severity 1+ on the top, severity 3+ on the bottom). 219
7.20 Demonstration Analysis: Day 2520 underlying parameter estimates model $1a_{21}$ (Bdy-sys_3) (severity 1+ on top, severity 3+ on bottom). 221
7.21 Demonstration Analysis: Day 2520 underlying parameter estimates model BB_{21} (Bdy-sys_3) (severity 1+ on top, severity 3+ on bottom). 222
7.22 Demonstration Analysis: All trials adverse event incidence data severity 1+ events. Proportion correct and Type-I error rates. . . . 236
7.23 Demonstration Analysis: All trials adverse event incidence data severity 3+ events. Proportion correct and Type-I error rates. . . . 237
7.24 Demonstration Analysis: All trials total adverse event data severity 1+ events. Proportion correct and Type-I error rates. 238
7.25 Demonstration Analysis: All trials total adverse event data severity 3+ events. Proportion correct and Type-I error rates. 239
7.26 Sensitivity Analysis: Changed threshold results severity 1+ adverse event incidence data (all trials). 243
7.27 Sensitivity Analysis: Changed threshold results severity 3+ adverse event incidence data (all trials). 244
7.28 Sensitivity Analysis: End of trial detection totals for missing data by case (model 1a on top, model BB on bottom). 252

A.1 c212.BB BUGs model. 286
A.2 c212.1a BUGs model. 287

C.1 c212.1a: Traceplots and posterior distributions for θ. 350
C.2 c212.BB: Traceplots and posterior distributions for θ. 350

C.3 Traceplot and posterior distribution for *Weight increased* with default simulation parameters. 351
C.4 Traceplot and posterior distribution plot for *Weight increased* with updated simulation parameters. 352
C.5 Traceplot for *Weight increased* with 10 chains and increased iterations. 353
C.6 Traceplot for *Adv_131* in interval 1260 - 1440 with default simulation parameters. 354
C.7 Traceplot for *Adv_131* in interval 1260 - 1440 with tuned simulation parameters. 355
C.8 Traceplot for *Adv_131* in interval 1260 - 1440 with tuned simulation parameters and additional iterations. 355

Chapter 1

Safety in the Context of Clinical Trials

1.1 Introduction

Randomised clinical trials (RCTs), conducted under the supervision of a Data Monitoring Committee (DMC), are the standard method for establishing the efficacy and safety of new treatments [1]. The DMC, in conjunction with an Institutional Review Board (IRB) or Research Ethics Committee (REC), is responsible for ensuring that a trial is carried out to the highest possible ethical and scientific standards [1], [2]. The DMC will meet at several prearranged times over the course of the trial to discuss the progress of the trial and consider the impact of any issues which may have arisen since the last meeting [2]. If deemed necessary (for example if an unexpected incident has occurred) the DMC may also meet at other times during the trial. The DMC may make recommendations about the conduct of the trial based on the evidence available to it when it meets. This may include recommendations regarding the continuation or termination of the whole trial or individual trial arms. Reasons for termination before the scheduled end of the trial may include early demonstrations of the efficacy of the new treatment, such that it would be unethical to withhold the treatment from the control or comparative groups. Alternatively, concerns regarding possible safety issues may arise where the new treatment group is at a higher risk of a serious health issue. In this case it would again be unethical to continue the new treatment. In addition to demonstrated benefit or harm, it is also possible that continuing the trial may be considered futile. In this case the DMC may take the view that it is very unlikely that the trial will be able to show any efficacy for the new treatment over the remaining planned duration of the trial [2].

Many different clinical outcomes or events are routinely measured and recorded in

the course of a randomised clinical trial, and the statistical analysis of these events may have several different uses within the context of any given trial. In particular a trial's DMC may use the analysis of these collected clinical events to recommend changes to the future conduct of the trial, or to make an interim or final decision on the overall safety of the treatment. Consequently the analysis of the safety related data, in particular what are termed adverse events (AEs), is extremely important.

The anticipated effect sizes of adverse events in clinical trials are generally small. In order to accumulate the number of events to detect statistically such effect sizes, with a sufficiently high power, either the follow up time has to be very long, or a large number of patients need to be recruited. While the recruitment of large numbers of patients may be both expensive and logistically difficult, a further issue arises from the fact that safety is generally not the primary clinical question of interest in a trial. The recruitment of large numbers of patients for this purpose alone may be considered unethical in that it has the potential to expose them to unnecessary harm.

The statistical analysis of safety or adverse event data from clinical trials is complicated by the large number of different variables recorded. If a hypothesis testing approach is taken, unadjusted significance tests may lead to large numbers of false positive results (Type-I errors). However, simple multiple comparison adjustments risk compromising the already possibly low power to detect important treatment differences. Consideration also needs to be given to the relative importance of false positive and false negative errors (Type-II errors). A false negative result could allow a potentially serious safety issue to go undetected, which in turn could lead to health consequences for patients. In effect we must consider both false positive and false negative errors as equally important when analysing safety data. Recently, a variety of classical (Mehrotra and Adewale [3], Siddiqui [4]) and Bayesian (Berry and Berry [5], DuMouchel [6]) methods have been proposed to address this problem. Although promising, these methods do not yet address the full complexity of the problem in that they are all generally restricted to the analysis of simple incidence data and make no use of the severity or, Siddiqui apart, the timing of adverse events. The methods are also relatively complex to implement and there is, to date, little experience among practitioners in their use. A major part of this study is to review and compare these and other methods for analysing clinical adverse event data, and to consider how they may be extended to interim (longitudinal) data.

In the remainder of this chapter, we discuss in more detail safety analysis in Randomised Clinical Trials. We give a brief overview of clinical trials and define what we mean by an adverse event (§1.2). We discuss the role and function of the DMC (§1.3) and how it relates to safety analysis. We describe the International Council for Harmonisation (ICH) guidelines, particularly those relevant to statistical analyses in clinical trials (§1.4). We look at some possible typical primary and secondary outcomes of RCTs, and at the other types of data that can accrue, particularly safety events, and at how they may be used to evaluate the safety of a treatment (§1.5). Data recording and the categorisation of adverse events by medical dictionaries, for example MedDRA, is discussed in §1.6. The conduct of a clinical trial (GlaxoSmithKline plc. (GSK) Study EGF100151) is discussed in §1.8, particularly with regard to safety. We then look at a categorisation and structure for adverse events which uses an underlying body-system or system organ class (SOC), such as those provided by standard medical dictionaries, to group the adverse events, and discuss how this may be used in a safety analysis (§1.9). The chapter ends with a discussion of the main research questions the project will address (§1.10).

Chapters 2 and 3 provide a review and discussion of some of the statistical methods currently used to control multiplicities in safety data analysis. We also examine in some detail the recent new methodologies from the papers [3], [4], [5], [6] with a view to possibly extending these methods to cover some of the types of analysis not currently generally performed. In Chapter 4 the methods are applied to the safety data from GSK Study EGF100151 and discussed. Chapter 5 is a simulation study on typical trial data with a view to fitting some of the models from Chapters 2 and 3. We wish to gauge their suitability and compare them on data where the underlying model is known. We also wish to assess if the use of the body-system information makes a difference to the correct detection of adverse events associated with treatment. In Chapter 6 we further discuss a number of approaches to modelling adverse event data at trial interim analyses, and define a number of models which may be suitable for this purpose. These are demonstrated on simulated trial data in Chapters 7, and the models which are the most suitable are identified. Chapter 8 summarises the study conclusions.

1.2 Overview

A clinical trial is a prospective study comparing the effect of an intervention (a treatment) against a control (for example a standard treatment or comparator medication) in people [1]. In this project we are primarily concerned with Phase

III trials. These are large trials, often containing several hundreds or thousands of participants randomised to treatment or control groups, whose main aim is to confirm the effectiveness or efficacy of the treatment in a wider population and to monitor the treatment for possible safety concerns. There are many possible designs for Phase III trials (see [1], [2] for further discussion).

A Phase III trial generally has a study protocol which provides the context for the trial and is essentially a roadmap for the overall conduct of the trial. Among other things the protocol will describe the background for the trial, the trial design, the definitions of the study population, the identification of population sub-groupings which may be important, patient enrolment procedures, trial duration, objectives, and follow-ups.

The trial is usually overseen by a Data Monitoring Committee who tend to be an independent group of people with relevant expertise. The DMC meets at regular intervals during the trial and makes decisions regarding the conduct of the trial. The trial itself may be run in a number of different centres or locations, possibly starting at different times, so it is important that the DMC receives timely and accurate data from all centres before it meets.

No treatment is completely safe so there will always be the possibility of safety issues for patients. Safety issues in clinical trials are usually characterised by what are termed adverse events. In order to investigate the relationships between these safety issues and treatments, we need to define exactly what is an adverse event. The United States National Cancer Institute (NCI) provides the Common Terminology Criteria for Adverse Events (CTCAE) which gives the following definition which we will use in this study:[1]

> An *Adverse Event* (AE) is any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medical treatment or procedure that may or may not be considered related to the medical treatment or procedure. An AE is a term that is a unique representation of a specific event used for medical documentation and scientific analyses.

The accurate assessment of adverse events is an important part of the role of the DMC.

[1] CTCAE Version 4.0: `http://evs.nci.nih.gov/`

1.3 Data Monitoring Committees

The Data Monitoring Committee has a number of functions with regard to the conduct of a clinical trial. In particular, it must ensure that the trial is conducted as per the protocol and to the highest ethical standards, and it must ensure that the participants are not unduly harmed [2]. Monitoring trial progress is an ongoing process for the duration of the trial.

The membership of the committee will typically consist of a statistician or statisticians, clinical experts in the area of the trial and other relevant fields and possibly an ethicist. Friedman et al. recommend that the committee members be independent of the participants, trial investigators and sponsors of the trial [1].[2] This will ensure no conflict of interest should the committee make a decision which may be contrary to the wishes of the trial sponsor, for example if the committee decide to end the trial early for safety reasons. The committee may have voting and non-voting members and meet in in both open and closed sessions. Typically, open sessions may be attended by trial sponsors, whereas for closed sessions sponsor attendance may not be allowed, or limited to the principal investigator [1].

The DMC has a number of different priorities. The primary priority is to protect trial participants from harm. The DMC also has responsibilities to the Institutional Review Board or Research Ethics Committee, the trial sponsor, and any concerned regulatory agencies, to ensure trial integrity and that, where applicable, mandatory reports of serious adverse events are made to the relevant authorities.

The committee will meet at regular intervals during the trial to review the accumulated data presented by a study statistician or, in certain cases, from a separate statistical centre. The committee will look at data relating to the primary and secondary objectives of the trial as well as any accumulated safety data. Care must be taken when making decisions, especially early in the trial where the rates of recruitment may be different for the individual arms of the trial resulting in possibly biased data. The committee may look at the blinded data or, if it is necessary, ask for the data to be unblinded. This may occur for example if there is an apparent serious safety issue on one arm of a trial which, if related to treatment,

[2]This is in contrast to Pocock who considers that the principal investigator should be included in the monitoring committee (possibly as chairman)[7]:"...major trials usually need a *monitoring committee* which meets periodically to assess the trial's overall progress. It should include the principal investigator...The monitoring committee should operate in an advisory capacity leaving the principal investigator to implement any decisions."

may require an alteration to, or even termination of, the trial [2]. In addition to safety issues, the DMC may also recommend termination of a trial if significant evidence of beneficial effects is found before the scheduled end of the trial, if it would be futile to continue the trial, if there are unfixable logistical or data issues, or the question the trial is designed to answer has already been decided or is in some sense no longer considered to be important [2].

1.4 ICH Guidelines

The International Council for Harmonisation[3] of Technical Requirements for Pharmaceuticals for Human Use (ICH)[4], created in 1990, brings together regulatory authorities and the pharmaceutical industry to discuss the scientific and technical aspects of drug registration. As part of its remit the ICH produces guidelines for many different aspects of clinical trials. These are divided into four main categories: Quality Guidelines (Q), Efficacy Guidelines (E), Safety Guidelines (S), and Multidisciplinary Guidelines (M). Of particular interest for safety analysis in clinical trials are the Safety and Efficacy Guidelines which cover safety studies and clinical trial safety in some detail.

1.4.1 Long-Term Treatment

Document ICH E1 (*Clinical Safety for Drugs used in Long-Term Treatment*) gives guidelines for the safety evaluation of drugs intended for the long-term treatment (chronic or repeated intermittent use for longer than 6 months) of non-life-threatening diseases and raises some issues with detecting rare adverse events. In particular it states that

> safety evaluation during clinical drug development is not expected to characterise rare adverse events, for example, those occurring in less than 1 in 1000 patients.

However, it is expected that during clinical drug development there should be some characterisation or quantification of the safety profile of a drug over a reasonable duration of time, consistent with its intended long-term usage, and that the safety evaluation should be based on previous experience of the occurrence and detection of adverse events. ICH E1 makes recommendations regarding the

[3]Formerly the International Conference on Harmonisation.
[4]`http://www.ich.org/`

size of cohort and length of treatment for the cases where adverse events occur early in the trial, stating that the number of patients treated for six months at the expected clinical dosage level should enable the pattern of adverse events to be determined. For drugs which may cause adverse events later in the trial, or cause events that increase in severity or frequency over time, the guidelines are for larger or longer-term safety analyses. For approved treatments, post-marketing surveillance of adverse events provides this type of longer-term safety analysis outside of the trial environment.

1.4.2 Pharmacovigilance

Documents ICH E2A-E2F deal with Pharmacovigilance. Document E2A gives the definition of an adverse event, details of how an adverse event's seriousness and severity may be classified (discussed further in §1.6.2), and what events should be subject to expedited reporting. The definition of an adverse event in ICH E2A is almost identical to that given by the NCI (§1.2). In addition to adverse events, ICH E2A defines a number of similar safety related terms. Adverse Drug Reactions are considered to be

> all noxious and unintended responses to a medicinal product related to any dose.

An Unexpected Adverse Drug Reaction is defined as an

> adverse reaction, the nature or severity of which is not consistent with the applicable product information (e.g. Investigator's Brochure for an unapproved investigational medicinal product).

E2D and E2E deal with post-approval safety data management and pharmacovigilance planning. For treatments in the post-marketing phase there are Spontaneous Report Adverse Event Databases for the collection of the occurrences of adverse events. Some regulatory agencies and drug monitoring centres have developed computerised methods for identifying potential reporting relationships in large databases [8], for example the US Food and Drug Administration (FDA) use a Bayesian data-mining approach [9], and the World Health Organization (WHO) use a Bayesian neural network[10].

1.4.3 Reporting and Statistical Analysis

Document ICH E3 deals with the compilation of a clinical study report which would be acceptable to all regulatory authorities of ICH regions and includes guidelines for safety reporting. ICH E3 recommends that safety related data be considered at three levels. Firstly, a measure of the exposure (duration, dose, drug concentration) should be given in order to evaluate to what degree safety can be assessed in the study. Secondly, there should be a summary of the common adverse events, with tabulations of their occurrences and a comparison of the rates between treatment groups. This comparison is often done using crude adverse event rates but ICH E3 also states that not every adverse event needs to be subject to a "rigorous statistical evaluation" but that it may be appropriate to include life tables or similar analyses which may be more informative than the crude rate. Thirdly, serious adverse events, other significant adverse events, related withdrawals and deaths should be identified and tabulated and if any of these events represents a previously unsuspected adverse effect of the treatment then this should be noted. Again, life tables or similar analyses may be used to assess the overall risk.

1.4.4 Statistical Principles for Clinical Trials

Document ICH E9 (Statistical Principles for Clinical Trials) sets out the statistical guidelines for assessing efficacy and safety in clinical trials. ICH E9 §4.5 describes interim analyses and a number of reasons for early stopping in a trial, mainly concentrating on the use of stopping boundaries, and also discusses the possibility that monitoring multiple endpoints may require the adjustment of significance and confidence levels (described in more detail in ICH E9 §5.6). With regard to safety analysis, ICH E9 §6.1 states:

> The incidence of a certain adverse event is usually expressed in the form of a proportion relating number of subjects experiencing events to number of subjects at risk. However, it is not always self-evident how to assess incidence. For example, depending on the situation the number of exposed subjects or the extent of exposure (in person-years) could be considered for the denominator. Whether the purpose of the calculation is to estimate a risk or to make a comparison between treatment groups it is important that the definition is given in the protocol. This is especially important if long-term treatment is planned and a substantial proportion of treatment withdrawals or deaths are expected.

> For such situations survival analysis methods should be considered and cumulative adverse event rates calculated in order to avoid the risk of underestimation.

With reference to the potential multiplicities of adverse events, the guidelines suggest using descriptive statistics with confidence intervals, but if a hypothesis testing approach is taken, then statistical methods to control the multiplicities are appropriate. The issue of "background noise" when accounting for adverse events is also discussed. Methods suggested for dealing with this include ignoring adverse events of mild severity, or requiring that an event should be observed at repeated visits to qualify for inclusion.

1.5 Clinical Trial Outcomes

1.5.1 Primary and Secondary Clinical Trial Outcomes

The primary objective is the main interest for the trial and will often be framed in terms of a hypothesis test where interest lies in whether the treatment has a different outcome than the control. The primary objective is the basis for choosing the size of the study [1] and may be something as simple as a decrease in all-cause mortality between the treatment and control groups. The primary response variable could be the time to a clinical outcome, such as death in this case. There may be a small number of secondary objectives which are additional questions of interest that the study may be designed to help answer. The trial investigator may also be interested in the responses of a number of (protocol defined) population subgroups under treatment versus control. As the trial is sized to answer the primary objective, the power to detect differences for the secondary objectives, or in sub-group analyses, is reduced.

The multiplicity of objectives, possibly combined with a number of planned interim data analyses, even excluding additional hypotheses regarding adverse events, provide a number of issues for the analysis of the trial data. Continued retesting of accumulating data has the potential to inflate the Type-I error. (Group) Sequential Methods with various stopping rules or α-spending functions may be used to control this phenomenon [11], [12], [1]. Similarly, there are a number of procedures for controlling the error rates when testing multiple hypotheses (§2.3).

1.5.2 Safety Outcomes in Clinical Trials

While laboratory investigations and early stage (Phase I) trials may indicate possible safety issues with a treatment, it is not ethically feasible to run a Phase III trial with safety as the primary outcome [1]. However, even though treatment safety is not the primary focus of Phase III trials, it is still extremely important and there are a number of advantages to conducting a detailed safety analysis as part of the trial.

An analysis of the adverse events during the trial, as opposed to post-hoc observations, can be considered a prospective study and adds to the overall credibility of the trial [1]. In addition, as many trials are blinded and balanced with regard to trial arms we may get an unbiased, fair comparison of adverse event occurrences on each arm. Outcomes from safety analysis may lead to further research on adverse events, even if no statistically significant differences between treatment and control are detected. However, there are limitations to what such an analysis can tell us. Trial participants are a selected sample of a particular population based on the trial protocol defined criteria. They may be healthier than the population with the condition the treatment is designed for and certain groups, such as pregnant women or individuals with specific types of medical conditions, may have been excluded. The absence of adverse events in lower risk groups does not necessarily mean that a treatment is safe and it is possible that safety issues may be understated [1].

The lower statistical power of secondary hypotheses means that randomised clinical trials may be unreliable in detecting rare adverse events and if the trial duration is relatively short, but the treatment is for longer term or chronic use, then later occurring adverse events may be missed. Many types of adverse events are found post-approval and, in particular, chronic use treatments need to have a continued safety evaluation. Low power concerns could be overcome by larger and longer trials, however these may have the same ethical issues as running a trial primarily for safety analysis, as well as logistical and economic difficulties, although ICH document E1 does allow for larger or longer safety analyses in certain circumstances (§1.4.1).

A possible alternative is to combine safety data from multiple trials in a meta-analysis, although these types of analysis are not always straightforward. Friedman et al. highlight several issues including, but not limited to, the difficulties finding

and including all relevant studies due to possible publication bias, differences in the treatments administered, different study populations, different follow-up periods, different measures of outcome, and difficulty accessing all relevant data [1, Ch. 17].

1.6 Recording Data in Clinical Trials

On enrolment in a trial a large amount of subject data is recorded. This will generally include non-clinical variables such as sex, age, and other information that the investigator has decided is of interest. In addition, there will be a number of clinical variables which will be used to monitor the patients' responses to treatment and may include other aspects of a patient's clinical history. Some of these variables may not be directly of interest in a statistical analysis, but others may be relevant. Both clinical and non-clinical variables may be used to segment the population into subgroups and hence may be used as covariate information in statistical models for the trial. Pocock classifies patient evaluation into four categories [7] (§3.5):

- Baseline assessment before treatment.
- Principal criteria for patient response.
- Subsidiary criteria, e.g. side-effects (including adverse events).
- Other aspects of patient monitoring.

In many jurisdictions the collection and reporting of adverse events is a regulatory requirement. If individual patients are assessed regularly by a clinician then the timing, duration and severity of any adverse events the patient may have experienced can be recorded. In order for this information to be useful for analysis, both within the trial and for comparison with other trials, a common terminology for the adverse events is required. The NCI CTCAE, or medical dictionaries such as MedDRA (§1.6.1), provide such a reference terminology.

The NCI CTCAE provides a relatively straightforward format for identifying adverse events. It consists of a number of adverse event categories or System Organ Classes (SOCs), 26 as of version 4.0, and individual definitions for identifying the adverse events within a particular SOC. It also provides a mapping of its terms to MedDRA and has the added advantage that it does not require a subscription.

1.6.1 Medical Dictionaries

A Medical Dictionary is a dictionary of the various terms used in medical practice. A suitable dictionary should be usable in clinical settings and related areas. There are a number of medical dictionaries in current use all of which provide similar services.

MedDRA (Medical Dictionary for Regulatory Activities)[5] was developed by the ICH. It is widely used by regulatory bodies, clinical research organisations (CROs), and pharmaceutical companies, and is maintained by the Maintenance and Support Services Organization (MSSO) which releases updated versions twice a year. COSTART (Coding Systems for a Thesaurus of Adverse Reaction Terms) is a dictionary provided by the FDA for the classification of adverse events. It has largely been replaced by MedDRA. WHO-ART (World Health Organisation Adverse Reaction Terminology) is a similar dictionary to COSTART maintained by the Uppsala Monitoring Centre (Sweden)[6] for the World Health Organisation Collaborating Centre for International Drug Monitoring[7].

Both WHO-ART and MedDRA have a similar hierarchical structure consisting of System Organ Classes (SOC) and various grouping and descriptor terms. The hierarchy allows for a medical condition or possible adverse event to be expressed in a number of different ways. A goal of the hierarchical structure is to standardise a medical condition by linking all of its possible descriptions. The WHO-ART dictionary describes a four level hierarchy consisting of a System Organ Class (SOC), High Level Terms (HLT), Preferred Terms (PT), and Included Terms (IT). Figure 1.1 shows an example from the *Heart rate and rhythm disorders* SOC.

MedDRA describes a similar hierarchy to WHO-ART and in fact it includes aspects of both COSTART and WHO-ART. It has a hierarchical structure with five levels consisting of System Organ Class (SOC), High Level Group Terms (HLGT), High Level Terms (HLT), Preferred Terms (PT), and Lower Level Terms (LLT). The PT is a single medical description of a symptom or observation while the LLT is how a patient or data recorder would describe a symptom or observation. Each LLT belongs to one PT and, in general, data will be recorded at the LLT level but reported at the PT level. The MedDRA LLT corresponds to the WHO-ART IT. There are 26 SOCs and over 70,000 LLTs. MedDRA supports Special Search

[5] `http://www.meddra.org/`
[6] `http://www.umc-products.com`
[7] `http://bioportal.bioontology.org/ontologies/WHO-ART`

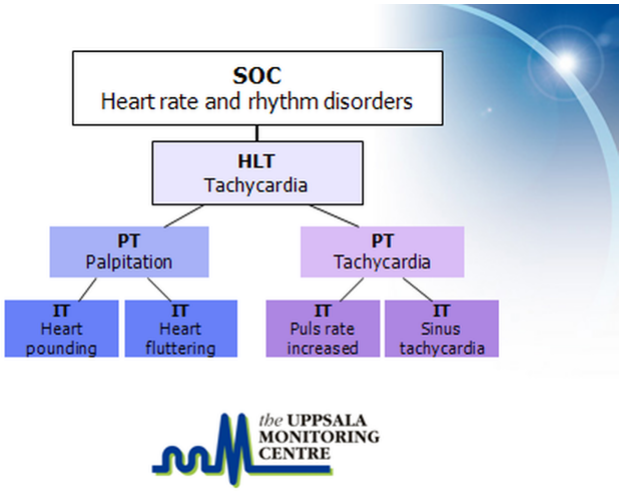


Figure 1.1. WHO-ART hierarchy for *Heart rate and rhythm disorders*.

Image reproduced with permission from the Uppsala Monitoring Centre.

Categories (SSOs) which allow PTs, which have similar causes but which belong to different SOCs, to be grouped in the same category, and has a querying technology, Standardised MedDRA Queries (SMQs), which give groupings of terms that are related to a particular medical condition or syndrome. MedDRA also supports dictionary terms belonging to more than one SOC.

1.6.2 Recording and Classifying Adverse Events

The recording of adverse events may be done in a number of ways. A trial form with a pre-prepared checklist of adverse events may be used, or information can be solicited directly from the patient. Occurrence is only one aspect of an adverse event. The frequency of occurrence of the event and a measure of its severity are also of interest. The exposure time for the patient may be important if a treatment does not cause an adverse event until it has been in use for a certain period, or if the adverse event occurs shortly after a treatment is administered. If a patient is taken off treatment or has their dose reduced, then this could be considered a measure of the severity of the adverse event.

Freidman et al. consider three general categories of adverse event: (1) serious adverse events (SAEs), (2) general adverse events, and (3) adverse events of special interest [1]. Severe adverse events are those which are life threatening, result in hospitalisation, are in some sense irreversible or persistent or result in significant disability, congenital anomaly, or birth defect. There is usually a requirement to

report these to relevant regulatory agencies within a certain time period of their occurrence. This is essentially the definition given in ICH E2A (II.B) (§1.4). Special interest adverse events are events whose occurrence may indicate potential issues with the treatment. These are often identified before the trial commences and defined in the trial protocol. General adverse events are those other events which may be recorded and whose symptoms may range from mild to severe.

The NCI CTCAE includes a simple numerical grading system for the severity of adverse events from 1 to 5, with the explanation of each grade shown in Table 1.1.

Severity Grade	**Description**
1	Mild
2	Moderate
3	Severe
4	Life-threatening
5	Death

Table 1.1. NCI CTCAE: Adverse event severities.

For each individual adverse event in the NCI CTCAE the actual definition of what constitutes a particular grade of adverse event is given. Not all adverse events have all grades defined. As an example, the NCI CTCAE defines *Hyperkalemia* as a disorder characterised by a laboratory test that indicates an elevation in the concentration of potassium in the blood. It is associated with kidney failure or with the use of diuretic drugs. It is in included in the *Metabolism and nutrition disorders* system organ class. The different grades of *Hyperkalemia* adverse events are given in Table 1.2.

Serious Adverse Events as defined in [1] correspond to Grades 4-5 in the NCI CCTAE classification, with General Adverse Events corresponding to Grades 1-3.

The Safety Planning, Evaluation and Reporting Team (SPERT), a technical group of the Pharmaceutical Research and Manufacturers of America formed in 2006, defines a three tier categorisation of Adverse Events, used in a number of publications, as follows: Tier 1 events are those events for which we have a specific hypothesis; Tier 2 events are those other events which are routinely collected during the trial but about which there is no hypothesis (common events); Tier 3 events are rare (possibly serious) events, these may be clinically important and may require

Severity Grade	Concentration of Potassium (millimols/litre); Characterisation
1	ULN[a] - 5.5 mmol/L;
2	5.5 - 6.0 mmol/L;
3	6.0 - 7.0 mmol/L; Hospitalization indicated
4	$>$ 7.0 mmol/L; Life-threatening consequences
5	; Death

Table 1.2. NCI CTCAE: Adverse event severity grade definitions for *Hyperkalemia*.

[a] Upper Limit of Normal.

specific evaluation [13], [14].

While it is possible to use any medical dictionary which provides a classification of adverse events, or syndrome groupings, as the basis for a modelling approach, the models we will investigate, which primarily use hierarchical groupings, do not directly depend on the dictionary, if any, chosen.

1.7 Safety Analyses in Clinical Trials

Safety analysis in clinical trials is mainly concerned with the analysis of adverse events suffered by patients over the course of the trial. Of primary interest is any difference in the occurrences of adverse events on the arms of a trial. For our purposes, following Freidman et al., we define a *safety signal* as a concern about an excess of adverse events on a treatment arm as compared to the control [1]. We wish to determine as far as possible if this safety signal is associated with the treatment or has occurred by chance. It may also be of interest if the occurrence of adverse events in patients under any treatment is higher than in the general population. This will occur naturally if the control arm is a placebo, but it may also be possible to make statements regarding the general population if the background rates of event occurrence are known. Care must be taken as the necessarily restricted treatment population may not be directly comparable with the general population. Within the context of a safety analysis we must be aware that some adverse events may be "anticipated on the basis of known biochemical properties of the investigational product or similar products or possibly from prior preclinical or clinical data" ([15], ICH E2A). For these adverse events the trial may incorporate specific hypotheses about their occurrence, or they may be an accepted

risk for the patients in the trial (Tier 1 events). These adverse events are often considered of special interest (§1.6.2) and a particular focus on the occurrence of these events is required. However, for a large number of the adverse events recorded there will be no pre-defined hypotheses regarding their occurrence, and their role in the safety profile of a particular treatment will be unknown.

The statistical methods suggested for analysing safety data in the ICH Efficacy Guidelines range from tabulations of adverse events and descriptive statistics, to analysis of the crude event rates and lifetable and survival analysis. Whatever methods or models are used, if we take the approach of hypothesis testing the occurrence of adverse events under treatment versus control, we will have the following issues:

- as there are a large number of adverse events there will be a correspondingly large number of hypotheses, particularly if we analyse safety data at interim points in the study, leading to the possibility that adverse events may be flagged as associated with treatment by chance;
- as trials are sized for efficacy of the primary outcome the power of any tests performed will generally be relatively low unless the rate of occurrence of the adverse events is large.

Rather than a direct hypothesis testing approach it can also be considered that a safety analysis is required to provide some sort of assessment of the association of adverse events with a treatment. This is more of an exploratory approach to safety analysis than the hypothesis testing approach which, when unadjusted, and given the large number of potential hypotheses, could in fact be considered a form of data-dredging. In §1.8 we look at both the general conduct of, and safety analysis in, a specific clinical trial.

The recent approaches to the analysis of adverse events in [3], [5] and [6] have used groupings to address the issues noted above. Bayesian methods ([5], [6]) also have the advantage of being able to address directly the multiplicity issues through the choice of prior distributions. These methods are discussed in more detail in Chapters 2 and 3.

1.8 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer

In this section we look at the application of some of the principles described above, particularly with regard to safety analysis, in a clinical trial for women with advanced or metastatic breast cancer.

1.8.1 HER2 Positive Breast Cancer

HER2 (human epidermal growth factor receptor 2) is a protein encoded in the ErbB2 gene and found on the surface of normal breast cells. Overexpression of ErbB2, leading to high numbers of HER2 receptors, plays a role in the progression of certain types of aggressive breast cancer, often referred to as HER2 positive (HER2+) breast cancers. These cases account for approximately 25-30% of all breast cancers and the prognosis for the patients is generally poor [16].

A number of targeted therapies are currently available and research is active in this area [17]. The most common targeted therapy is trastuzumab (Herceptin®), a monoclonal antibody which works by binding on to the HER2 receptor and preventing the cells from dividing and growing. Chemotherapy (anthracyclines, taxanes, and other drugs) may also be given in combination with trastuzumab. However, trastuzumab is associated with cardiac dysfunction ([18]) and many patients do not respond to this treatment, with resistance to it also developing over time [19], [20].

1.8.2 Trial Overview

EGF100151[8] was a GSK sponsored Phase III randomised clinical trial evaluating *Lapatinib and Capecitabine* versus *Capecitabine* in women with refractory advanced or metastatic breast cancer, overexpressing ErbB2, who had prior treatment which included anthracyclines, taxanes, and trastuzumab. The study period was from 29 March 2004 to 18 February 2010. *Capecitabine*, a chemotherapy, is considered to be the control in this trial. Lapatinib is a targeted therapy which inhibits HER2 [21]. There were three clinical study reports (CSRs), two during the study, and a final report at the end of the study, details of which are given in Table 1.3. All

[8]ClinicalTrials.gov identifier: NCT00078572

the data, trial descriptions, and a number of different publications referenced in this section, which reflect various aspects of the trial, are based on these reports, which are available through the GSK clinical study register.[9]

Effective Date	Study Period	GSK ID	Description
09/08/06	29/03/04 - 15/11/05	UM2004/00001/00	Planned interim analysis (§1.8.4).
08/09/06	29/03/04 - 03/04/06	ZM2006/00137/00	Enrolment complete (§1.8.4).
14/03/11	29/03/04 - 18/02/10	2010N107773_00	End of trial.

Table 1.3. Trial EGF100151: Clinical study reports.

1.8.2.1 Primary and Secondary Outcomes

The primary outcome was time to tumour progression (TTP) as assessed by blinded independent review based on imaging data and investigator assessment. Secondary outcomes included overall survival, progression-free survival, 6-month progression-free survival, overall response rate, clinical benefit response rate, time to response, duration of response, and safety measured according to the NCI CT-CAE (Table 1.1).

1.8.2.2 Data Monitoring Committee Role

The Data Monitoring Committee was to review efficacy and safety data during the trial with the possibility of stopping the trial early if:

- there were strong safety issues (harm);
- there was strong evidence of superiority of *lapatinib and capecitabine* over *capecitabine* (benefit);
- there was strong evidence that *lapatinib and capecitabine* would fail to show superiority over *capecitabine* if the trial was to run to completion (futility).

[9] `http://www.gsk-clinicalstudyregister.com/`

1.8.2.3 Interim and Final Analyses

The interim analysis of TTP was planned after 133 events (progressions or deaths due to breast cancer) using O'Brien-Fleming stopping-boundaries for assessing efficacy or futility in the *lapatinib and capecitabine* arm [22]. The final analysis of TTP and secondary endpoints was planned to occur when 266 or more events had occurred, with survival data collected until 457 deaths had been observed, when another analysis of overall survival was to be performed [21].

1.8.2.4 Statistical Methods

The primary outcome was framed in terms of a hypothesis test, with the null hypotheses being that the hazard ratio (HR) for TTP for *lapatinib and capecitabine* compared to *capecitabine*, λ, was greater than or equal to 1, and the alternative hypothesis being $\lambda < 1$. The study was designed to have 90% power to detect a 50% increase in median TTP, estimated to be 3 months for *capecitabine* and 4.5 months for *lapatinib and capecitabine*, and 80% power to detect a 30% improvement in median overall survival, estimated at 8 months for capecitabine and 10.4 months for *lapatinib and capecitabine*. The planned power for the trial required the enrolment of 528 subjects.

Efficacy measures for the primary outcome, overall survival, progression-free survival and progression-free survival at 6 months were based on survival analysis techniques with stratified log-rank tests, estimates of hazard ratios (Pike estimator), Kaplan-Meier curves, and estimates of medians and quartiles all being used. For overall tumour response rate treatment arms were compared using Fisher's exact test. There were no adjustments for multiplicity and no specific statistical analysis was planned for safety other than summary statistics (§1.8.2.6).

1.8.2.5 Study Populations and Randomisation

The study population consisted of non-pregnant, non-lactating women over the ages of 18 with a confirmed diagnosis of invasive, progressive mestastatic advanced breast cancer, with ErbB2 (HER-2/neu) overexpression, who had been exposed to a number of therapies excluding *capecitabine*. A full description of the inclusion/exclusion criteria may be found in the study reports (Table 1.3).

The randomisation of the subjects was stratified into three groups according to stage and site of the disease, with the intention of balancing the groups. The planned, randomised, and actual totals under treatment and control are given in

Table 1.4.

	Lapatinib plus Capecitabine	**Capecitabine**
Planned	264	264
Randomised[1]	198	201
Total Under Treatment	207	201

Table 1.4. Trial EGF100151: Population randomisation.

[1] There was a change to enrolment after the first interim analysis (§1.8.4).

There was a difference in the populations used for the efficacy analyses and the safety analyses. The population used for the efficacy analyses was the intention-to-treat population, while for the safety analyses the population was based on the actual treatment received. The per-protocol population was only used in a supplementary analysis of the TTP data.

1.8.2.6 Safety

Adverse events were coded using MedDRA and grouped by system organ class. They were graded by the NCI CTCAE where applicable. Summaries were provided by frequency, percentage of total subjects, system organ class, and preferred term. Serious adverse events were given additional separate reports.

Phase I studies had indicated that orally administered *lapatinib* is well tolerated in doses up to 1,800 mg daily, with the majority of adverse events reported either grade 1 or 2 including diarrhea,[10] skin rash, fatigue, anorexia, nausea, and vomiting. Some serious adverse events were reported: dyspnoea, dehydration, neutropenia, episodes of interstitial pneumonitis, and decrease in left ventricle ejection fraction [23]. The adverse events in Table 1.5 were identified as being of special interest for the trial.

PPE is a frequent adverse event for *capecitabine* and has its own *capecitabine* specific toxicity rating (UM2004/00001/00).

[10]The spellings diarrhea and diarrhoea are used interchangeably in the CSRs.

Adverse Event	CTCAE System Organ Class
LVEF[1]	Cardiac disorders
PPE[2]	Skin and subcutaneous tissue disorders
Diarrhea	Gastrointestinal disorders

Table 1.5. Trial EGF100151: Adverse events of special interest.

[1] Left Ventricular Ejection Fraction.
[2] Palmar-Plantar Erythrodysaesthesia / Hand-foot syndrome.

In addition to adverse events of special interest, the trial also defined serious adverse events (§1.6.2) to be those which resulted in death, were life threatening, required hospitalisation or prolonged existing hospitalisation, resulted in disability or incapacity, were congenital anomalies or birth defects, were grade 4 laboratory tests, were grade 3+ or 20% decrease from baseline LVEF (cardiac dysfunction), were grade 3+ symptoms of pneumonitis, were hepatobiliary events where Alanine Aminotransferase (ALT) $> 3 \times$ Upper Limit of Normal (ULN) and total bilirubin $> 2.0 \times$ ULN ($> 35\%$ direct). It was a study requirement that SAEs be reported to GSK within 24 hours to allow GSK to fulfil its legal obligations to report adverse events to the relevant regulatory authorities.

1.8.3 Trial Conduct

1.8.3.1 Data Collection and Recording

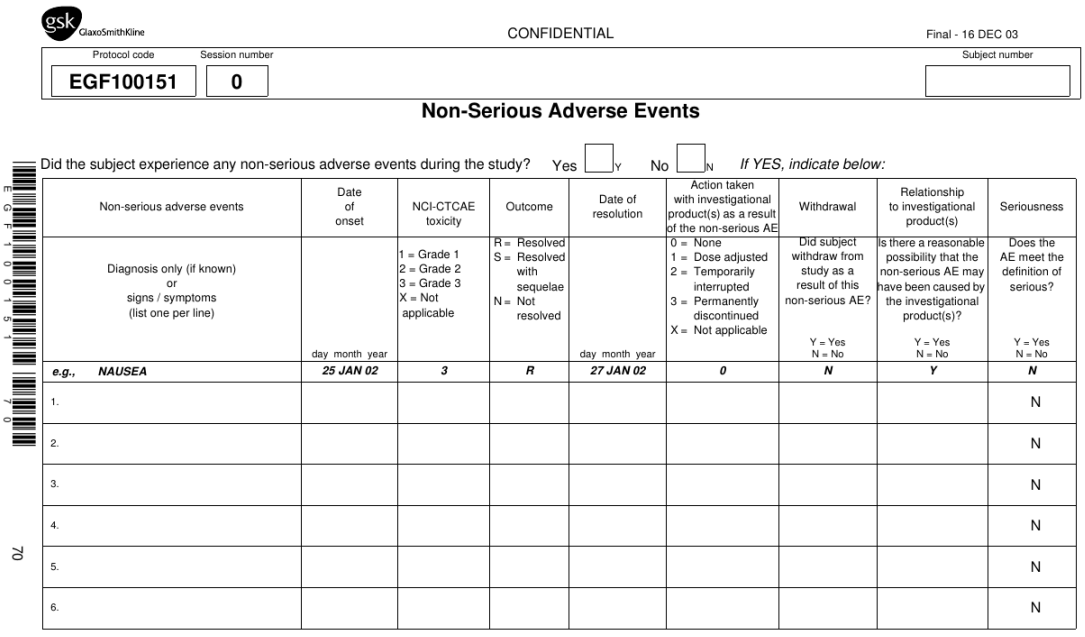
Safety and efficacy assessments were performed every 6 weeks for the first 24 weeks, and then every 12 weeks, and at end of treatment. Additional safety checks were performed on all subjects every 3 weeks. Laboratory and clinical responses were used to determine toxicity and disease progression.

The investigator had responsibility for detection and documentation of adverse events (UM2004/00001/00, pg: 38). Events which were part of the natural course of the disease were excluded. Any abnormal laboratory finding or assessment detected at baseline, or during the study, which significantly worsened and met the definition of an adverse event was recorded. All adverse events were collected from first dose to 30 days after last dose and the intensity of the adverse events as graded by the NCI CTCAE was recorded. Ongoing adverse events were reviewed at subsequent assessments and followed until resolution. The investigator was required to assess the relationship between the treatment and the occurrence of any

adverse event using clinical judgement to determine that relationship with alternative causes being considered. The investigator was also required to identify study defined serious adverse events.

Information was recorded during the study on GSK case report forms (CRFs) which were reviewed for completeness and accuracy and entered into the study database. The CRF contains specific pages for recording general and serious adverse events. The non-serious adverse event form is shown in Figure 1.2.

The importance of serious adverse events to study progress is reflected in the CRF where several pages are available to record safety data. The SAE part of the CRF contains information at the individual patient level, including details of the treatment the patient received, relevant medical conditions or risk factors, concomitant medications, and details of assessments or examinations that were part of the subject's care for the adverse event. The first page of the SAE form is shown in Figure 1.3.



GlaxoSmithKline CONFIDENTIAL Final - 16 DEC 03

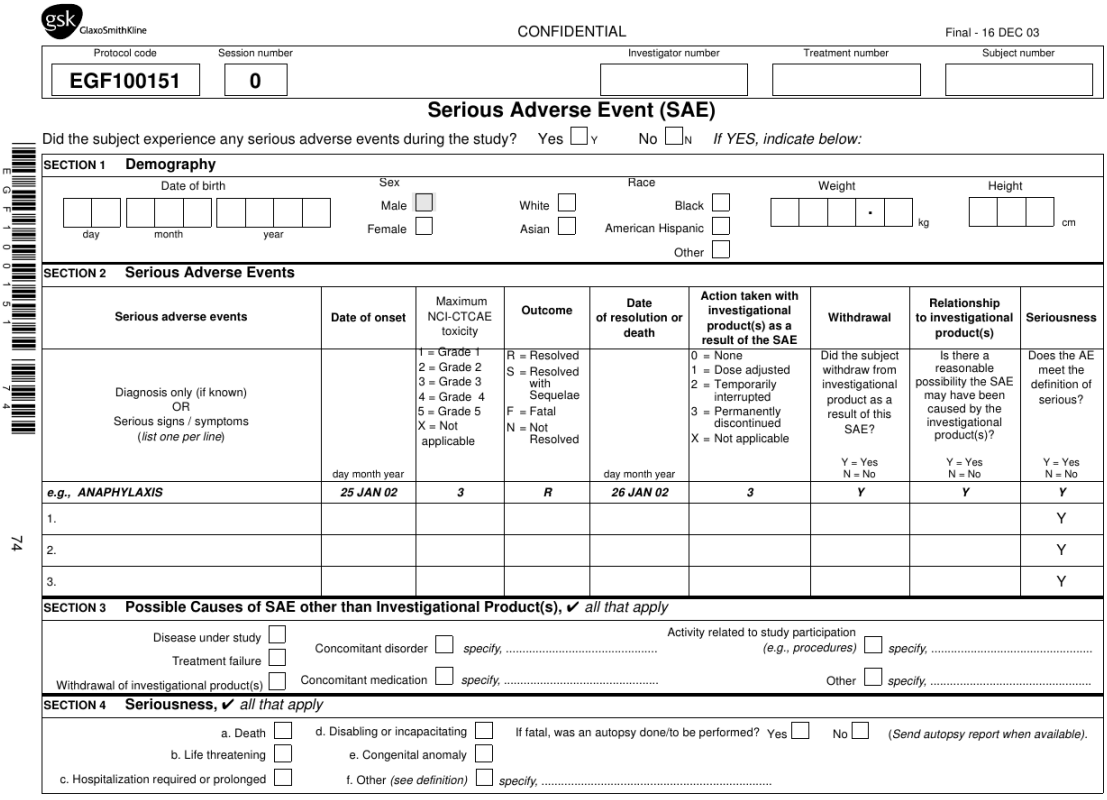
Protocol code: EGF100151 Session number: 0 Subject number:

Non-Serious Adverse Events

Did the subject experience any non-serious adverse events during the study? Yes ☐ Y No ☐ N *If YES, indicate below:*

Non-serious adverse events	Date of onset	NCI-CTCAE toxicity	Outcome	Date of resolution	Action taken with investigational product(s) as a result of the non-serious AE	Withdrawal	Relationship to investigational product(s)	Seriousness
Diagnosis only (if known) or signs / symptoms (list one per line)	day month year	1 = Grade 1 2 = Grade 2 3 = Grade 3 X = Not applicable	R = Resolved S = Resolved with sequelae N = Not resolved	day month year	0 = None 1 = Dose adjusted 2 = Temporarily interrupted 3 = Permanently discontinued X = Not applicable	Did subject withdraw from study as a result of this non-serious AE? Y = Yes N = No	Is there a reasonable possibility that the non-serious AE may have been caused by the investigational product(s)? Y = Yes N = No	Does the AE meet the definition of serious? Y = Yes N = No
e.g., NAUSEA	***25 JAN 02***	***3***	***R***	***27 JAN 02***	***0***	***N***	***Y***	***N***
1.								N
2.								N
3.								N
4.								N
5.								N
6.								N

Figure 1.2. Trial EGF100151: Non-serious adverse event recording form.



GlaxoSmithKline — CONFIDENTIAL — Final - 16 DEC 03

Protocol code	Session number	Investigator number	Treatment number	Subject number
EGF100151	0			

Serious Adverse Event (SAE)

Did the subject experience any serious adverse events during the study? Yes ☐Y No ☐N *If YES, indicate below:*

EGF100151 74

SECTION 1 Demography

Date of birth (day / month / year) — Sex: Male ☐ Female ☐ — Race: White ☐ Black ☐ Asian ☐ American Hispanic ☐ Other ☐ — Weight (kg) — Height (cm)

SECTION 2 Serious Adverse Events

Serious adverse events	Date of onset	Maximum NCI-CTCAE toxicity	Outcome	Date of resolution or death	Action taken with investigational product(s) as a result of the SAE	Withdrawal	Relationship to investigational product(s)	Seriousness
Diagnosis only (if known) OR Serious signs / symptoms (*list one per line*)	day month year	1 = Grade 1 2 = Grade 2 3 = Grade 3 4 = Grade 4 5 = Grade 5 X = Not applicable	R = Resolved S = Resolved with Sequelae F = Fatal N = Not Resolved	day month year	0 = None 1 = Dose adjusted 2 = Temporarily interrupted 3 = Permanently discontinued X = Not applicable	Did the subject withdraw from investigational product as a result of this SAE? Y = Yes N = No	Is there a reasonable possibility the SAE may have been caused by the investigational product(s)? Y = Yes N = No	Does the AE meet the definition of serious? Y = Yes N = No
e.g., ANAPHYLAXIS	***25 JAN 02***	***3***	***R***	***26 JAN 02***	***3***	***Y***	***Y***	***Y***
1.								Y
2.								Y
3.								Y

SECTION 3 Possible Causes of SAE other than Investigational Product(s), ✔ *all that apply*

Disease under study ☐ — Treatment failure ☐ — Withdrawal of investigational product(s) ☐ — Concomitant disorder ☐ *specify,* ..... — Concomitant medication ☐ *specify,* ..... — Activity related to study participation *(e.g., procedures)* ☐ *specify,* ..... — Other ☐ *specify,* .....

SECTION 4 Seriousness, ✔ *all that apply*

a. Death ☐ — b. Life threatening ☐ — c. Hospitalization required or prolonged ☐ — d. Disabling or incapacitating ☐ — e. Congenital anomaly ☐ — f. Other *(see definition)* ☐ *specify,* ..... — If fatal, was an autopsy done/to be performed? Yes ☐ No ☐ *(Send autopsy report when available).*

Figure 1.3. Trial EGF100151: Serious adverse event recording form.

1.8.4 Interim Analysis and Trial Changes

The clinical cut-off date for the planned interim analysis was 15 November 2005 (Table 1.3). By this date investigators had identified 146 TTP events in 321 randomised subjects. However, an independent review committee who were blinded identified only 114 TTP events, lowering the power of the interim analysis and, as a result, the O'Brien-Fleming stopping-boundaries were adjusted. The interim analysis showed a significantly longer TPP for *lapatinib and capecitabine* compared to *capecitabine*, HR = 0.49 (0.34, 0.71) and, based on the recommendations of the data monitoring committee, enrolment to the trial to the trial was discontinued and, as of 3 April 2006, cross-over was offered to subjects receiving *capecitabine*. The 15 November 2015 data was queried and re-validated, resulting in the independent identification of 121 events in 324 randomised subjects, but the difference between the groups remained significant. The median TTP was 36.7 weeks in the *lapatinib and capecitabine* group compared to 19.1 weeks in *capecitabine*.

By the termination of enrolment 399 patients in total had been randomised, 9 were

being screened and were then offered *lapatinib and capecitabine*, giving the totals in Table 1.4. In addition 36 patients crossed-over to *lapatinib and capecitabine*. The early termination of enrolment and cross-over reduced the power to detect differences in overall survival [21].

1.8.5 Efficacy Results

For the primary outcome *lapatinib and capecitabine* increased TTP with a hazard ratio of 0.57 (0.43-0.77). The addition of *lapatinib* to *capecitabine* provides a statistically significant and clinical benefit for subjects with HER-2^+ advanced breast cancer [24].

For overall survival, the median survival times were 75.0 weeks for *lapatinib and capecitabine* versus 64.7 weeks for *capecitabine*, with hazard ratio 0.87 (0.71, 1.08), p-value = 0.210. The trial changes following the interim analysis resulted in a loss of power to detect differences in overall survival, but there were indications of a trend towards better survival with *lapatinib and capecitabine* [21].

While treatments may delay disease progression, toxicity can negatively impact on subjects. Quality of life was assessed during the study using the Functional Assessment of Cancer Therapy-Breast (FACT-B) and EuroQoL (EQ-5D) questionnaires [25]. While no statistically significant differences were detected between the groups, the quality of life (QoL) direction was consistently in favour of the *lapatinib and capecitabine* treatment.

1.8.6 Safety Analysis

The toxicity of the treatments was to be assessed by clinical and laboratory parameters and classified by the NCI CTCAE. If necessary, treatment delays of up to two weeks were possible for both *lapatinib and capecitabine* to allow for reduction in toxicity. Subjects with more than two weeks toxicity were generally withdrawn from *lapatinib*. For the adverse events of special interest (Table 1.5) additional precautions were taken, with subjects with an NCI CTCAE grade 3 or 4 LVEF or interstitial pneumonitis being withdrawn from *lapatinib* (UM2004/00001/00). We have seen in §1.8.3.1 that in addition to the clinical and laboratory analysis investigator judgement was an important part of adverse event analysis.

1.8.7 Interim Analysis Safety Report

Two reports were produced as a result of the interim analysis (Table 1.3). The first report included data up to the interim analysis cut-off date (15 November 2005). The second report included data up to the end of enrolment (03 April 2006). In this section we look at the data from the first interim analysis (15 November 2005). The actual safety population (based on treatment taken) consisted of 164 patients for the *lapatinib and capecitabine* arm and 152 for the *capecitabine* arm. The safety aspects of the report are summarised in [26]. The most common adverse events are shown in Table 1.6 where the p-values are from a Fisher exact test.

Adverse Event	**Lapatinib plus Capecitabine (164)**	**Capecitabine (152)**	***p*-value**
Diarrhea	98	60	<0.001
Nausea	72	64	0.830
Vomiting	43	37	0.800
Stomatitis	24	18	0.570
Abdominal pain	25	32	0.230
Constipation	16	17	0.820
Dyspepsia	18	5	0.014
PPE	80	74	1.000
Rash	45	23	0.011
Dry skin	18	8	0.100
Fatigue	29	41	0.060
Mucosal inflammation	18	19	0.800
Asthenia	10	18	0.110
Headache	15	20	0.340
Pain in extremity	21	13	0.300
Back pain	17	9	0.220
Anorexia	25	30	0.370
Dyspnea	18	10	0.240

Table 1.6. Trial EGF100151: Incidence of the most common adverse events (all grades), 15 November 2005 [26].

The most common adverse events are diarrhea, nausea, vomiting, PPE, fatigue, and rash and most were grade 1, 2, or 3. Looking at the data from a purely hypothesis testing point of view, using a significance level 0.05 and not accounting for multiplicities, we can see that diarrhea (0.0005), rash (0.011), and dyspepsia (0.014) would be considered significant. Applying the Bonferroni correction (§2.3.1.1) leaves diarrhea as the only significant adverse event. The difference in incidence of diarrhea was due to an increased number of grade 1 and 2 adverse events in the *lapatinib and capecitabine* group. For rash the difference was mainly due to grade 1 events.

The proportions of serious adverse events considered by investigators to be related to the study treatments were similar between the groups. For the adverse events of special interest 6 subjects in the *lapatinib and capecitabine* arm and 1 in the *capecitabine* arm experienced a decreased LVEF. All 6 in the combination therapy were considered drug related and 4 were considered serious adverse events. None of these adverse events resulted in subject discontinuation. For PPE, approximately half of the subjects in the each group had an adverse event, with the mean duration shorter in the *lapatinib and capecitabine* group.

The safety conclusions from the interim report were that while diarrhea was more common in the *lapatinib and capecitabine* group, this was due to lower grade events, and the overall incidence of adverse events between the groups was similar, as was the incidence of serious adverse events. Other aspects of safety based on clinical chemical toxicities were considered, and similar incidences were reported for each group. Overall, at this stage of the trial, the conclusion was that *lapatinib and capecitabine* was well tolerated by the subjects.

1.8.8 Final Safety Report

The final clinical study report (CSR: 2010N107773_00) concluded that the incidence and type of adverse events was consistent with previous analyses (§1.8.7), with the majority of adverse events seen being grade 1 or 2. The most common serious adverse events, those experienced by more than 1% of the trial subjects, are listed in Table 1.7. None of these are statistically significant at the 0.05 level using a Fisher exact test. Of the recorded adverse events, 326 affected one percent of the population or less (on both treatment arms).

Adverse Event	**Lapatinib plus Capecitabine (210)**	**Capecitabine (191)**
Diarrhea	15	12
Dehydration	7	5
Vomiting	4	4
Dyspnea	4	4
Ejection Fraction Decreased	5	2
Pulmonary Embolism	4	2
Anemia	3	1
Hypokalemia	3	1
Pyrexia	2	2
Convulsion	2	1
Hyponatremia	2	1
Left Ventricular Dysfunction	0	3
Mucosal inflammation	1	2
Nausea	0	3

Table 1.7. Trial EGF100151: Serious adverse events experienced by more than 1% of subjects, final clinical study report.

For adverse events of any grade Table 1.8 show the most frequent adverse events, the top 10 in each group, along with the p-values from a Fisher exact test.

Adverse Event	Lapatinib plus Capecitabine (210)	Capecitabine (191)	*p*-value
Diarrhea	145	78	<0.001[1]
PPE	130	106	0.223
Nausea	98	85	0.689
Vomiting	63	43	0.112
Fatigue	51	49	0.817
Decreased Appetite	44	42	0.809
Rash	61	20	<0.001[2]
Abdominal Pain	31	30	0.889
Stomatitis	37	23	0.125
Headache	29	30	0.672
Mucosal Inflammation	33	25	0.480
Asthenia	27	24	1.000
Constipation	24	24	0.760
Dyspnea	31	16	0.061

Table 1.8. Trial EGF100151: Most common adverse events (all grades), final clinical study report.

[1] Actual p-value: 0.000000015.
[2] Actual p-value: 0.000003.

We can see from the table that diarrhea and rash have much higher counts on the *lapatinib and capecitabine* arm and this is confirmed by the very small p-values of the Fisher exact test. Both the adverse events remain significant at the 5% level when applying a Bonferroni correction over the adverse events in the table. If we look at most common adverse events by grade on each arm (Tables 1.9 and 1.10), we can see that rash is not among the top six most common adverse events on the *capecitabine* arm and counts consist of mostly grade 1 and grade 2 events, with only 2 grade 3 events recorded, both on the *lapatinib and capecitabine* arm. For diarrhea, again the majority of events are of grade 1 and 2, but there were 2 grade 4 events on the *lapatinib and capecitabine* arm and more grade 3 events, 31 compared to 20. We know from Table 1.7 that diarrhea was the leading serious adverse event for both trial treatments.

Adverse Event	Severity					Total
	1	2	3	4	5	
Diarrhea	64	28	31	2	0	145
PPE	27	67	36	0	0	130
Nausea	62	31	5	0	0	98
Vomiting	36	22	5	0	0	63
Rash	41	18	2	0	0	61
Fatigue	23	21	7	0	0	51

Table 1.9. Trial EGF100151: The six most common adverse events by grade on the *lapatinib and capecitabine* arm, final clinical study report.

Adverse Event	Severity					Total
	1	2	3	4	5	
PPE	23	53	30	0	0	106
Nausea	55	27	3	0	0	85
Diarrhea	32	26	20	0	0	78
Fatigue	21	21	6	1	0	49
Vomiting	26	14	3	0	0	43
Decreased	31	10	1	0	0	42
Rash	16	6	0	0	0	20

Table 1.10. Trial EGF100151: The six most common adverse events, and rash, by grade, on the *capecitabine* arm, final clinical study report.

A number of deaths were determined to have occurred from serious adverse events, 6 on each arm, and a small number of patients experienced LVEF events [21]. Eight of the subjects who experienced LVEF events were considered to have had serious adverse events, but all of them reported as asymptomatic.

The overall conclusion from a safety point of view was, in agreement with the interim analyses, that *lapatinib and capecitabine* is well tolerated. There are higher incidence of some adverse events, particularly diarrhea, but the majority of these were grade 1 and 2.

1.9 Grouping of Adverse Events and the Body-System Approach to Safety Analysis

In §1.8 we have seen a standard approach to safety analysis in a clinical trial. The trial protocol defined special interest adverse events and what constituted a serious adverse event, with clinical judgement playing an important role in the safety analysis. Based on the most common adverse events (Table 1.8) diarrhea appears to be, at least statistically, associated with the *lapatinib and capecitabine* arm of the trial, as is rash. However, it was known from Phase I studies (§1.8.2.6) that diarrhea and rash were adverse events whose appearance might have been expected, and diarrhea was defined in the trial protocol as an adverse event of special interest (Table 1.5). Most diarrhea and rash events were grade 1 or 2, and the conclusion of the safety study was not just that diarrhea or rash were expected adverse events of *lapatinib and capecitabine*, but that the treatment was well tolerated.

Even though the treatment is considered to be well tolerated, the confirmation of diarrhea and rash as expected adverse events is of interest. We have seen that when applying a multiplicity controlling procedure to the most common adverse events, that both remain significant (§1.8.8). However, the correction was applied only over the set of 14 adverse events from Table 1.6. The actual total number of different adverse events (or, more specifically, preferred terms) recorded in the trial was 497 over 23 system organ classes summarised in Table 1.11. In fact even for this large a number of potential hypotheses, when we apply a Bonferroni multiplicity correction, both diarrhea and rash remain significant, such is the strength of the signals associated with these events.

Table 1.12 shows the 10 adverse events which were significant at the 5% level for a Fisher exact test. After the application of the Bonferroni correction, 8 of the 10 adverse events which are individually significant at the 5% level are no longer deemed significant, they are being swamped by the number of potential hypotheses. Even epistaxis and dyspepsia, which have what could be considered small p-values, are not flagged. This issue is further compounded by the fact that of the 497 adverse events considered, 326 had incidences of 1% or less on both treatment arms. When controlling for multiple hypotheses, these very low incidence adverse events make it difficult to flag any but the strongest signals.

In this particular trial, where no multiplicity controls were applied and diarrhea

System Organ Class	Number of Adverse Events
Gastrointestinal disorders	58
Skin and subcutaneous tissue disorders	46
General disorders and administration site conditions	31
Nervous system disorders	37
Musculoskeletal and connective tissue disorders	23
Infections and infestations	53
Respiratory, thoracic and mediastinal disorders	36
Metabolism and nutrition disorders	17
Eye disorders	26
Investigations	42
Psychiatric disorders	11
Blood and lymphatic system disorders	12
Reproductive system and breast disorders	16
Vascular disorders	16
Hepatobiliary disorders	8
Injury, poisoning and procedural complications	17
Cardiac disorders	17
Renal and urinary disorders	12
Ear and labyrinth disorders	2
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	11
Immune system disorders	3
Surgical and medical procedures	2
Endocrine disorders	1
Total	497

Table 1.11. Trial EGF100151: System organ classes and adverse event totals.

System Organ Class	Adverse Event	Lapatinib plus Capecitabine (210)	Capecitabine (191)	*p*-value
Gastrointestinal disorders	Diarrhea	145	78	0.000[1]
Skin and subcutaneous tissue disorders	Rash	61	20	0.000[2]
Respiratory, thoracic and mediastinal disorders	Epistaxis	18	4	0.004
Gastrointestinal disorders	Dyspepsia	24	7	0.004
Skin and subcutaneous tissue disorders	Dermatitis acneiform	8	0	0.008
Musculoskeletal and connective tissue disorders	Muscle spasms	12	3	0.035
Infections and infestations	Localised infection	10	2	0.038
Musculoskeletal and connective tissue disorders	Arthralgia	22	9	0.039
Musculoskeletal and connective tissue disorders	Back pain	27	13	0.047
Skin and subcutaneous tissue disorders	Nail disorder	13	4	0.049

Table 1.12. Trial EGF100151: Adverse events significant at the 5% level, final clinical study report.

[1] Actual *p*-value: 0.000000015
[2] Actual *p*-value: 0.000003

and rash are expected adverse events, this may not be an issue, particularly if any adverse events which may be significant are not clinically interesting. However, if a number of unexpected adverse events had arisen on one particular treatment arm and were possibly of interest, we could consider how we would determine if these events were associated with treatment in an environment where we wish to control for multiplicities. So, for example, epistaxis (nose bleed) may not be of clinical interest in this trial but could be potentially distressing for a subject. The 10 adverse events flagged were from were from 5 different system organ classes, with 3 from *Musculoskeletal and connective tissue disorders*, 3 from *Skin and subcutaneous tissue disorders*, 2 from *Gastrointestinal disorders*, and 1 each from *Respiratory, thoracic and mediastinal disorders* and *Infections and infestations*. This raises the possibility of the existence of potential relationships between the adverse events within a system organ class, and one possible approach is to attempt to use this type of relationship in an analysis.

For EGF100151 the proportion of subjects who experienced at least one adverse event within a particular system organ class is graphed in Figure 1.4.

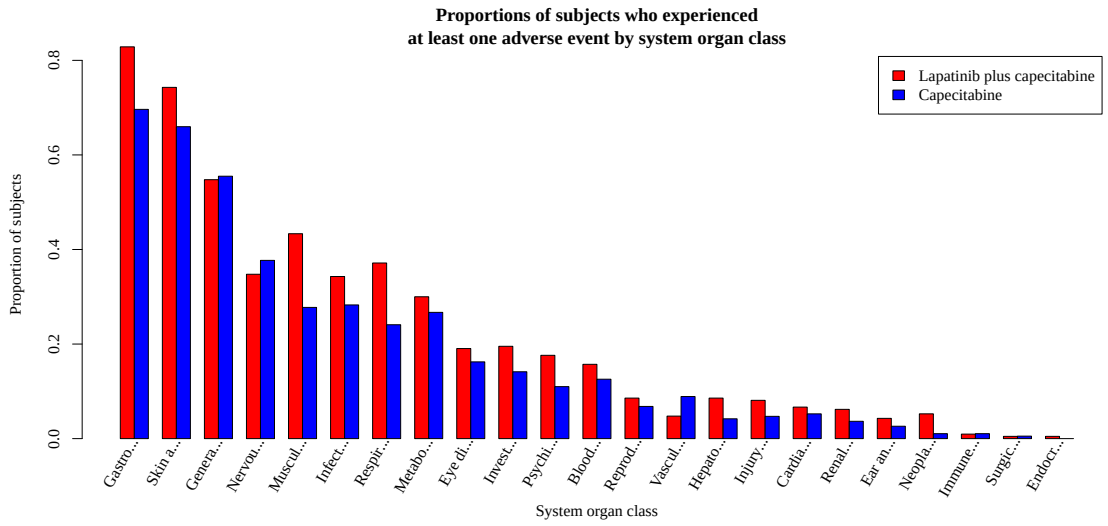


Figure 1.4. EGF10015: Proportion of subjects who experienced an adverse event by system organ class.

Based on Figure 1.4, the system organ classes *Gastrointestinal disorders* (containing diarrhea), *Skin and subcutaneous tissue disorders* (containing rash), *Musculoskeletal and connective tissue disorders*, *Infections and infestations*, and *Respiratory, thoracic and mediastinal disorder* have the largest raised incidences on the

lapatinib and capecitabine arm. The actual incidence numbers and p-values from Fisher exact tests are given in Table 1.13. As it turns out, at the 5% significance level only the following system organ classes are significant:

1. *Musculoskeletal and connective tissue*;
2. *Gastrointestinal disorders*;
3. *Respiratory, thoracic and mediastinal disorders*;
4. *Neoplasms benign, malignant and unspecified (incl cysts and polyps)*;

and applying a multiple hypothesis controlling procedure (Bonferroni) leaves just two:

1. *Musculoskeletal and connective tissue*;
2. *Gastrointestinal disorders*.

We note that even though Rash was an expected adverse event, and its individual p-value was very small, its system organ class, *Skin and subcutaneous tissue disorders*, is not flagged as significant by the tests based on incidence at the system organ class level.

The analysis of adverse incidence at the system organ class level leads us to consider if we can use the system organ class or body-system when performing safety analyses. The body-system approach to safety analysis is to group related adverse events into a single body-system and to use the additional body-system information in a statistical analysis. If a treatment affects a particular body-system then we may expect to see raised adverse event counts for all adverse events in that body-system [5]. We are in effect trying to take advantage of any relationship (biological or otherwise) between the adverse events. The information within the body-systems may also be used as one approach to handling multiplicities, with the additional information available used to shrink non-significant effects towards zero [27].

Choosing which adverse events to group together will have an effect on the results of any analysis. We can see above that the 326 events experienced by less than 1% of the subjects on each arm of EGF10015 swamp some potential safety signals. Possible approaches to this are to include these events as they are, or alternatively

System Organ Class	Lapatinib plus Capecitabine (210)	Capecitabine (191)	*p*-value
Gastrointestinal disorders	174	133	0.002
Skin and subcutaneous tissue disorders	156	126	0.080
General disorders and administration site conditions	115	106	0.920
Nervous system disorders	73	72	0.603
Musculoskeletal and connective tissue disorders	91	53	0.001
Infections and infestations	72	54	0.199
Respiratory, thoracic and mediastinal disorders	78	46	0.005
Metabolism and nutrition disorders	63	51	0.507
Eye disorders	40	31	0.513
Investigations	41	27	0.183
Psychiatric disorders	37	21	0.065
Blood and lymphatic system disorders	33	24	0.393
Reproductive system and breast disorders	18	13	0.577
Vascular disorders	10	17	0.113
Hepatobiliary disorders	18	8	0.103
Injury, poisoning and procedural complications	17	9	0.223
Cardiac disorders	14	10	0.674
Renal and urinary disorders	13	7	0.262
Ear and labyrinth disorders	9	5	0.423
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	11	2	0.022
Immune system disorders	2	2	1.000
Surgical and medical procedures	1	1	1.000
Endocrine disorders	1	0	1.000

Table 1.13. Trial EGF100151: Number of subjects who experienced at least one adverse event in a system organ class.

they could either be ignored in the statistical analysis as suggested in ICH E9 (§1.4.4), or they could aggregated into a common summary event. DuMouchel [6] referencing [28] states:

> The question of how to classify and group adverse drug reaction reports can be controversial because different assignments can change the statistical significance of count data treatment effects, and methods and definitions for comparing adverse drug event rates are not well standardized.

Given this, it is important that adverse event groupings should be made before the trial commences, with medical dictionaries, such as MedDRA or WHO-ART, which provide defined hierarchies or groupings of adverse events, well suited to this task. It would also be beneficial to have an indication that a model which uses a body-system grouping is appropriate for the data. This could possibly be accomplished at an early interim analysis by looking at the incidence counts of the common adverse events by system organ class to see if they are from a small number of SOCs, as we did for EGF10015 above. Finally, the use of a body-system approach in clinical trials has been discussed a number of times in the literature with an early use described in [29].

1.10 Research Questions

There have been a number of clinical trials which failed to pick up safety issues before a drug or treatment received regulatory approval and was released for use in the general population [4]. In the most serious of cases this late detection of safety issues has resulted in drugs being withdrawn completely from use. In less serious cases restrictions on use of the drug, or the inclusion of warnings on the packaging, have been considered sufficient to address the safety issues. In either case the detection of safety issues in the post-marketing phase of a treatment's life cycle, as opposed to the trial phase, can have a serious effect on the health of patients and also a financial impact both for the companies developing the treatments, and the regulatory bodies responsible for overseeing them. Consequently, an improvement in the early or accurate detection of safety issues would be a welcome development. The main research questions that we address in this project are with regard to the use of the body-system in the detection of treatment related adverse events.

The first question addressed is whether the introduction of the body-system is

useful for detecting adverse events. A number of the methodologies from Chapters 2 and 3, which use a body-system approach, are applied to the safety data from EGF10015 in Chapter 4, and then compared in Chapter 5 using a simulation study. The next question looked at is whether we can develop methods using the body-system which are able to take into account the rates and timings of occurrences of adverse events, and, if so, can these methods be used to identify events which are associated with treatment. This is addressed in the remainder of the thesis.

The methods we look at may be considered complimentary to those set out in ICH E9. While Tier 1 events have specific hypotheses, for many of the other adverse events there will be no definite prespecified hypothesis and we are primarily interested in whether we can say anything about these adverse events, while taking into account any relationships which may exist between them at the system organ class level.

Chapter 2

Error Controlling Procedures in Safety Analysis

2.1 Introduction

Two issues which may complicate the analysis of data from clinical trials are the possibility of multiple outcomes, and the interim analysis of accumulated trial data. Both of these are of particular importance in a safety analysis where large amounts of patient safety data are recorded, but about which there may be no specific hypotheses in the trial protocol (§1.6.2). Such data accumulates during a trial's progress and is presented periodically when the trial's Data Monitoring Committee meets. An ability to analyse such data at these interim periods, taking into account possible multiplicity issues, would be a useful aid to the DMC's decision making process. A trial's continuation may be in doubt if there are serious concerns about the safety of a treatment, so it is important that this be assessed correctly. Simple methods for the comparison of adverse event incidence on different trial arms, such as Fisher exact tests, lead directly to the potential issue of multiple hypotheses. In this chapter we review some approaches to safety analysis in clinical trials, concentrating on error rate controlling procedures for multiple hypotheses. Modelling, and in particular Bayesian approaches, are reviewed in Chapter 3.

ICH Guidelines (§1.4.4) suggest a number of approaches for the statistical analysis of safety data, including the use of tabulations, descriptive statistics, crude event rates, and survival analysis. The guidelines state that it is not always clear how to assess event incidence and none of these methods directly address the issue of multiplicities, although controlling for multiplicities is part of the statistical guidelines (ICH E9). Some of these methods are briefly reviewed in §2.2. If we wish to perform multiple hypothesis tests, then many methods are available to adjust

for multiple comparisons. Typically, these methods are used to control the Type-I error rate, and do not say anything about the control of the Type-II error rate. These methods are reviewed in §2.3. For safety and regulatory reasons controlling the Type-I and Type-II error rates often need to be considered of equal importance when analysing adverse event data, and a balance between these found (ICH E9). Consequently alternative approaches are of interest.

One approach to handling multiplicities is to group related adverse events by body-system and use this additional information in the analysis. Grouped methods are the main focus of this study and as part of this review we will look at a number of these which have recently appeared in the literature, including a more in-depth discussion (§2.3.2.5) of the following papers, which use grouped methods:

- **D. V. Mehrotra and A. J. Adewale.** Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals [3].
- **Hu, J. X., Zhao, H., and Zhou, H. H.** False Discovery Rate Control With Groups [30].

The approaches taken in [3] and [30] are different to methods which model safety data in that they directly address multiplicity issues and require p-values, derived from test statistics, to compare the different trial arms.

These methods use an assumed relationship within the data (body-systems or similar groupings) to help directly control the error rates. Apart from this assumption, and the assumption that the Benjamini-Hochberg (BH) False Discovery Rate controlling procedure [31] can be applied (i.e. the hypotheses being tested satisfy the dependency assumptions of [32]), there are no further requirements. It is hoped that this approach controls the Type-II error rate although, as we will see, this is actually not directly addressed. Unlike modelling approaches, which are more in line with the derivation of a safety profile of a particular treatment, based on the data available and possibly an assumed model or structure for the data, statements can be made directly about the Type-I error rate.

2.2 ICH Guidelines

The methods suggested by the ICH guidelines (§1.4.4) are standard descriptive statistics or survival analysis methods. Survival methods applicable for adverse

event incidence comparisons between treatment and control include estimation of the survival curve (e.g. Kaplan-Meier or Cutler-Ederer estimate [1]), comparison of survival curves at a particular time point, comparison of median survival times, and total survival curve comparison by log-rank (Mantel-Haenszel) or Generalised Wilcoxon tests. If the survival curves cross, then the log-rank and generalised Wilcoxon tests may not be reliable [1], and other options, such as the modified Kolmogorov-Smirnov test [33], [34], are available. Semi-parametric approaches, such as the Cox proportional-hazards model, and fully parametric approaches are possible, and these are discussed in Chapter 3. These methods are well understood and are also discussed in a general way in [1, Ch. 15]. The ICH guidelines also recommend the use of crude event rates for comparisons between groups. We define these here for future reference:

The crude incidence rate (CRI) is defined as the number of subjects with a specific event divided by the total number of subjects in the relevant study group [4]:

$$\text{CRI} = \frac{n}{N}$$

where n is the number of subjects in the group having the adverse event, and N is the total number in the group.

For long-term follow up the crude incidence rate has a number of potential problems. It may not be a good measure of occurrence because it does not include a subject's total exposure time in the calculation, and in some trials it may be unrealistic to assume that all patients are followed to study end time. The CRI may therefore be biased if some subjects discontinue their trial participation before the end of the study [4]. Also, the CRI cannot deal with the possibility of multiple occurrences of events per subject. Fisher exact tests as well as normal approximations to the CRI may be used to compare treatment arms. Liu et al. discuss these and a number of other possible approaches [35].

Due to this difficulty in interpreting the crude incidence rate in the presence of subjects who drop out of the trial an alternative, the exposure-adjusted incidence rate, is often used [4].

The Exposure-Adjusted Incidence Rate (EAIR) is defined as the number of subjects with a specific event divided by the total exposure-time among the subjects

in the group:

$$\text{EAIR} = \frac{n}{\sum_i t_i}$$

where n is the number of subjects in the group having the i^{th} type event and t_i is a subject's exposure time until having an i^{th} type event or, if a subject has no event of type i, t_i is the last follow-up time for that subject. If a subject has multiple events of the i^{th} type then t_i is the time of the first event of that type. n is also the numerator in the crude incidence rate.

The total exposure-time of all randomised subjects in a study group is $\sum t_i$. While the EAIR as defined above is for event incidence it could be extended in a straightforward way to recurrent events.

The EAIR may be interpreted as the number of events occurring in a population per unit time, and it is a valid statistic for a treatment comparison when the incidence rate is relatively constant over the study duration [35]. Treatment comparisons using EAIR are biased for events which usually occur early in the study, events whose incidence rates decrease over time, or events which occur on a delayed basis [35]. Liu et al. discuss comparing the EAIR between treatment arms [35].

It may be difficult to assess whether incidence rates are constant over time for any adverse event, particularly at an interim analysis, so both the crude incidence rate and the event-adjusted incidence rate need to be used with caution when evaluating or comparing adverse event incidence during a clinical trial.

2.3 Error Rate Controlling Procedures

Error controlling procedures are generally used to control the overall Type-I error rate when performing multiple hypothesis tests. Let $H_i,\ 1 \leq i \leq m$, be a family of m hypotheses with p_i their associated p-values. Table 2.1, based on Table 1 from [31], is used to describe the characteristics of the hypotheses.

Test Result	Null hypothesis true	Alternative hypothesis true	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 2.1. Null and Alternative Hypotheses: actual status and test outcomes.

In Table 2.1, m is the total number of hypotheses tested, m_0 is the number of true null hypotheses, $R = V + S$ is the number of rejected null hypotheses ("discoveries"), T is the number of false negatives (Type-II errors), S is the number of true positives ("true discoveries"), U number of true negatives, and V is the number of false positives (Type-I errors, "false discoveries").

2.3.1 The Familywise Error Rate

The Familywise Error Rate (FWER) is defined as the probability of making one or more Type-I errors when analysing multiple hypotheses (the "family"):

$$\text{FWER} = \text{P}(V \geq 1) \tag{2.1}$$

The FWER is said to be controlled at a level α if FWER $\leq \alpha$. α is often referred to as the *nominal* significance [12].

2.3.1.1 Controlling the FWER

Possibly the simplest method for controlling the FWER is the Bonferroni correction which rejects H_i if $p_i \leq \frac{\alpha}{m}$ [12]. A simple calculation shows that FWER $\leq \alpha$.[1]

[1] If I_0 is the set of p-values corresponding to true null hypotheses then:

$$\begin{aligned}\text{FWER} &= \text{P}\left(\cup_{p_i \in I_0}\left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{p_i \in I_0} \text{P}\left\{p_i \leq \frac{\alpha}{m}\right\} \\ &\leq \sum_{p_i \in I_0} \frac{\alpha}{m} = m_0 \frac{\alpha}{m} \\ &\leq \alpha\end{aligned}$$

The Bonferroni correction makes no distributional or dependency assumptions about the hypothesis tests being performed, and hence there are no restrictions on its applicability [12]. However, the Bonferroni correction may be considered to be too conservative for many clinical trial needs as, depending on the correlation structure of the p-values, it may control the error rate at a much lower level than the nominal α value [5]. This may also increase the Type-II error rate with consequently reduced power.

A method similar to Bonferroni correction, with the additional requirement of independence of the test statistics, that again controls the FWER at the level α, is the Šidák (or Dunn-Šidák) procedure [36]. In this case each hypothesis is tested at the level:

$$\alpha_{\text{SID}} = 1 - (1 - \alpha)^{\frac{1}{m}}$$

Similar approaches are the Holm or Holm-Bonferroni procedure [37] (no dependency assumptions about the test statistics), and the Hochberg procedure which requires the test statistics be independent, and is also applicable under some forms of positive dependency [38]. Other testing procedures which are applicable in certain situations are Scheffé's method for linear regression [39], the Tukey range test for pairwise comparisons which can be used within an ANOVA approach [40], and Dunnett's test which uses a t-statistic and attempts to exploit correlations which may exist between the test statistics [41].

2.3.2 False Discovery Rate

An alternative to the conservative procedures for controlling the FWER is to control instead what is called the False Discovery Rate (FDR). In their 1995 paper Benjamini and Hochberg consider that often control of the FWER is not needed, and that an alternative rate to control is the expected proportion of errors [31]. Essentially, control of the FDR assumes that when many of the tested hypotheses are rejected it may be preferable to control the proportion of errors, rather than the probability of making even one error, with a potential gain in power associated with controlling the FDR, as opposed to controlling the FWER [31],[32]. Mehrotra and Heyse claim that as the FDR controls the FWER when all null hypotheses are true, and that when not all null hypotheses are true, the FDR has higher power than the FWER, it is suitable for use in a safety context [15].

From Table 2.1 we have that the proportion of errors committed by falsely rejecting

null hypotheses can viewed through the random variable:

$$Q = \begin{cases} \frac{V}{V+S} & V+S \neq 0 \\ 0 & V+S = 0 \end{cases}$$

The FDR is then defined to be:

$$\text{FDR} = Q_e = \text{E}[Q] = \text{E}\left[\frac{V}{V+S}\right] = \text{E}\left[\frac{V}{R}\right] \tag{2.2}$$

2.3.2.1 Control of the FDR by the Benjamini-Hochberg Procedure

The BH-procedure for controlling the FDR is as follows [31]:

Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered p-values and let $H_{(i)}$ be the null hypothesis corresponding to $p_{(i)}$.

1. For a given α find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject (i.e. declare positive discoveries) all $H_{(i)},\ i = 1, \ldots, k$. If no such k exists reject no hypothesis.

In [31] it was shown that this procedure controls the FDR at level α, under the assumption of independent test statistics. In fact in [31] control of the FDR is shown to be tighter:

$$\text{FDR} \leq \frac{m_0}{m}\alpha$$

with equality for continuous test statistics, but as m_0 is generally unknown control can only be claimed to level α.

The BH procedure was shown to be extendable to certain types of dependency conditions by Benjamini and Yekutieli (§2.3.2.2) [32]. Benjamini and Liu [42] provided an alternative procedure for controlling the FDR (under independence) which, under a simulation study where the number of tested hypotheses was small and many of the null hypotheses were not true, was more powerful than the procedure in [31].

2.3.2.2 Extension to Positive Regression Dependent Test Statistics

If $X_1, ..., X_m$ are the test statistics of the hypotheses in Table 2.1, $\mathbf{X}$ is their joint distribution, and I_0 is a subset of $\{1, ..., m\}$, then Benjamini and Yekutieli define

the property *Positive Regression Dependency on each one from a Subset* I_0 (PRDS on I_0) as follows [32]:

For any increasing set D and each $i \in I_0$, $P(\mathbf{X} \in D | X_i = x)$ is non-decreasing in x.

Benjamini and Yekutieli show that if the joint-distribution of the test statistics, $\mathbf{X}$, is PRDS on the subset of test statistics corresponding to the true null hypotheses, then the BH-procedure controls the FDR at level α [32].

The positive association of the test statistics is expressed by the fact that larger values for the test statistics corresponding to true null hypotheses increase the probability that the joint test statistic distribution is in the set D.

2.3.2.3 Adjusted BH *p*-values

An equivalent formulation of the BH-procedure using adjusted p-values is as follows:

$$\tilde{p}_{(m)} = p_{(m)} \tag{2.3}$$

$$\tilde{p}_{(j)} = \min\left(\tilde{p}_{(j+1)}, \frac{m}{j} p_{(j)}\right), \quad j \leq m-1 \tag{2.4}$$

If $\tilde{p}_{(j)} \leq \alpha$ then $H_{(j)}$ is rejected (as are all $H_{(i)},\ i < j$).

2.3.2.4 Further Extensions of False Discovery Rate Control

The original assumptions in [31] and [42] were that the m_0 tests were independent, but Benjamini and Yekutieli were able to extend the procedure to testing under positive dependency [32]. They also introduced an adjusted procedure which allows control of the FDR in all dependency cases, but which is less powerful than the BH-procedure. Benjamini, Krieger and Yekutieli introduced a number of adaptive linear step-up procedures for controlling the FDR in their 2006 paper [43]. An estimate of the number of true hypotheses, m_0, is made as a first part of the process. As the BH-procedure actually controls the FDR at the level $\frac{m_0}{m}\alpha$ ([31]), knowledge of the value of m_0 would allow control of the FDR precisely at, or close to, α, leading to a potential increase in power. However, as in general m_0 is unknown, an estimator is used, and a simulation study does show a gain in power over the original BH-procedure.

Storey introduces the positive FDR $\left(\text{pFDR} = E\left[\frac{V}{R} | R > 0\right]\right)$ as an alternative to

the FDR and explores some of its properties [44]. The pFDR is related to both Bayesian modelling and classification theory. In [45] Storey uses the idea of fixing the rejection region to estimate α, rather than fixing α and estimating k. The idea is to provide point estimates. The data is used to estimate m_0 and the pFDR and FDR are estimated under the assumption of independent test statistics. Storey and Tibshirani extend this to dependent statistics [46], [47]. Storey, Taylor, and Siegmund also show that the point estimate of the FDR may be used to define valid FDR controlling procedures [48]. Efron considers what he calls a local FDR, which is an empirical Bayes version of the BH-procedure [49]. In [50] and [51] Efron et al. relate an empirical Bayes approach to the FDR when analysing data from gene expression experiments. Genovese and Wasserman discuss some of these approaches in [52], [53].

Muller et al. investigate using a Bayesian decision-theoretic approach to control the FDR and derive a Bayes rule which is a variation of the BH-procedure [54] while, in a similar manner, León-Novelo et al. take a Bayesian approach to the False Discovery Proportion (FDP), whose posterior mean is the False Discovery Rate [55]. These methods are discussed further in §3.6.4.

Genovese, Roeder, and Wasserman look at applying a weighting to the p-values in their approach [56], work which is extended by Hu et al. who use a group BH approach (GBH) with a weighting procedure which uses the relative importance of each group [30]. This is essentially an estimate of the number of true hypotheses in the group. They take the view that prior information may allow the hypotheses to be divided into subgroups based on the characteristics of the problem. They discuss a number of adaptive models, including one from [43], but with groupings. The asymptotic properties of the various GBH procedures are discussed. In the case where m_0 is known in advance they show the GBH procedure controls the FDR at the required level and, in other cases, the procedure controls the FDR asymptotically. The GBH is discussed is more detail in §2.3.2.5.

In 2004 Mehrotra and Heyse introduced a double FDR (DFDR) or Mehrotra-Heyse-Tukey approach, which is also a grouped BH method [15]. The BH-procedure is applied at two levels: first at an overall grouping level, where hypotheses were grouped by body-system, and then at the individual hypothesis level. Due to some theoretical issues with the original double FDR it was updated by Mehrotra and Adewale in 2012 [3]. This approach is discussed in detail in §2.3.2.5.

Yekutieli provides a modification of the BH-procedure for testing non-positive

dependent statistics where the set of p-values is divided into separate subsets, each of which have a BH-procedure applied to them (ssBH) [57]. Within each subset there is the assumption that the test statistics are positive regression dependent. This approach is at most as powerful as the BH-procedure.

2.3.2.5 Increasing the Power through Grouped False Discovery Rate Methods

A possible source of additional information about test statistics or p-values lies in groupings of hypotheses. If the p-values within these groupings are related in some way, then it may be possible to use this information to maintain control of the FDR while gaining power over the standard BH-procedure. The Double False Discovery Rate (DFDR) [3], [15] and Group Benjamini-Hochberg process (GBH) [30] are two recent methods which look to take advantage of a grouping structure within the hypotheses being tested to provide some gain in power over other approaches. While the two methods are different, they do have a number of common characteristics in that they either restrict (GBH), or remove (DFDR), the contribution of certain groups of hypotheses from the list of hypotheses to be tested.

The GBH method estimates the proportion of true hypotheses in each group and uses this to apply weights to the p-values. The procedure is shown to asymptotically control the FDR, and increase power compared to the BH-procedure under certain conditions [30]. The GBH procedure is considered to work well when the number of potential signals is small among a large number of hypotheses, and where the proportions of true hypotheses may be different in the different groups. The data discussed and analysed in the paper are from gene expression experiments in which typically a large number of genes are monitored, but only a few are expected to be associated with a disease. Before applying the method the authors use a step to cluster the genes into groups. In contrast, in a clinical trial the groups will be predefined body-systems or system organ classes. Asymptotic results for the GBH are obtained by assuming a finite number of groups, but with the overall number of hypotheses increasing while the proportions of true hypotheses tend to a limiting value. The details of the method are discussed in §2.5.

The Mehrotra and Adewale DFDR method [3] is an adjustment of the original double FDR, introduced by Mehrotra and Heyse [15], to analyse adverse event data which is grouped into defined subsets by body-system or other clinical characteristics. The main focus is the analysis of Tier 2 events with regard to the control of

false discoveries (Type-I errors), however they also discuss the importance of controlling Type-II errors as detailed by the ICH Expert Working Group [58], making the point that an over-stringent adjustment for false positives can lead to false negatives. It is hoped that the less stringent approach of the new DFDR procedure will in some way control the false negative rate, although no attempt is made to control it directly in this approach. The major assumption for the paper is to "assume that the underlying test statistics have a non-negative dependency structure [32] that enables the generation of valid BH FDR". Two simulation studies were used to compare the new procedure to a number of existing methods. A small hypothetical simulation of a clinical trial, and a larger study which made the additional assumption that adverse events in different body-systems are independent. For illustrative purposes the methods were also applied to two examples of real clinical data. The technical details of the approach are discussed in §2.4 but, in summary, the new DFDR method applies the BH-procedure at the body-system level first and gathers all p-values for any body-system deemed significant into a single family F. It then applies the BH-procedure to F. Adverse events are flagged if both the body-system and individual (adjusted) p-values are significant at some level α.

The simulation studies in [3] are interesting to us because they compare the DFR and GBH to unadjusted significance testing (NOADJ) and the standard BH-procedure. The four different methods are applied and the rates tabulated and compared using a significance level of $\alpha = 10\%$ for the DFDR (justified below). For the small clinical trial study the new DFDR correctly identified the 6 adverse events which were known to be different between the control and treatment groups. The other NOADJ and GBH methods also identified these six, plus an additional three adverse events, whereas the BH approach, ignoring the body-systems, performed least well and flagged only one adverse event.

The methods were also compared using a larger simulation study whose goal was to assess the competing approaches (including the original version of the DFDR [15]) and their power properties. The simulation data used were correlated binary random variables representing occurrence or non-occurrence of adverse events as described by Lunn and Davis [59]. Adverse events in different body-systems were considered to be independent and the power was defined to be the expected value of $\frac{C}{T}$, where C is the number of adverse events correctly flagged, and T is the corresponding number of adverse events with underlying true signals. They considered two simulations in total with 5000 simulated trials for each. Again a significance

level of $\alpha = 10\%$ was used for the new DFDR method. The four main conclusions from the simulation study were:

1. The NOADJ approach (with significance level 5%) results in too many false discoveries.
2. The BH approach (one step FDR) and new DFDR approaches were at or below the target α level in all the simulations. The GBH and the original DFDR were above the target level α in many cases.
3. The BH approach was consistently less powerful than the new DFDR approach. The new DFDR is able to capitalise on the groupings by body-system. The power for GBH and the original DFDR were both higher than the new DFDR.
4. The power for the new DFDR was generally comparable to the NOADJ approach when $\alpha = 10\%$, and was higher when $\alpha = 15\%$. For $\alpha = 5\%$ it was not possible to ensure no power loss relative to the NOADJ approach. This is the reason $\alpha = 10\%$ that was chosen for the new DFDR approach.

The authors state that the simulation studies strongly support the new DFDR approach for flagging adverse events as it provides a balance between no adjustment and over adjustment.

The DFDR is an example of what could be considered a more general 2-STEP method where we consider groups or families of hypotheses and we:

1. Select a subset, F, of the families of hypotheses based on some criteria, for example based on a function of the p-values and some threshold value t_F, where t_F may be a function of the data.
2. Apply the BH-procedure at level α to F.

Trivially (§D.1) any general 2-STEP method controls the FDR, however the choice of the final family affects the power of the method. No exact or asymptotic results for the DFDR are given in [3].

Comparisons between the DFDR and GBH methods should take into account the different situations for which they are designed. We should expect each method to perform well in its chosen area. Comparisons between the methods can be made

based on control of the FDR, and the overall power of the procedure or the false non-discovery rate (FNR). Typically, determining the power of such procedures requires knowledge of the distributions of the p-values under the alternative hypotheses. For any 2-STEP procedure this is complicated by the first step which selects a subset of the hypotheses. For example, in the DFDR procedure (defined in §2.4.3) knowledge of the distribution of the order statistics for each group under the null and alternative hypotheses is required, complicated by possible dependencies among the statistics. For this reason the power of a method is often established via simulation studies or, as for the GBH, asymptotically. It is though sometimes possible to set bounds on the power in comparison to other methods. The ssBH is shown to be at most as powerful as the BH-procedure by showing that the number of hypotheses it rejects is a subset of those rejected by the BH-procedure [57]. For the DFDR (and any 2-STEP procedure) the step which reduces the number of hypotheses before applying the BH-procedure to the reduced set introduces an additional random variable, the number of hypotheses in the final set, increasing the difficulty of obtaining analytic results.

2.4 Double False Discovery Rate

In this section we briefly describe the original double FDR approach [15] as it covers the main ideas behind the updated DFDR [3]. We then describe the updated version of the method. As stated above the main focus of the DFDR approach is the analysis of Tier 2 events (§1.6.2) with regard to the control of false discoveries (Type-I errors). Although the method is applicable to any situation involving grouped p-values, Mehrotra and Heyse only consider adverse event incidence data counts, grouped by body-system, in control and treatment groups [15].

2.4.1 Notation

Let there be s body-systems with k_i adverse events in body-system i, and let p_{ij} be the between-group[2] p-value for the j^{th} adverse event within body-system i. So, for example, in the MMR vaccine trial considered in the paper [15], the p_{ij} value is from a two-sided Fisher Exact Test.

2.4.2 Original Double FDR Approach

The original DFDR procedure is defined as follows:

[2]Between treatment and control.

1. Remove adverse events for which the total incidence within body-system (in both treatment and control groups) is so low that a rejection even at the unadjusted 5% level is impossible.

2. Among the remaining adverse events flag those for which the p-value achieves statistical significance after adjusting for multiplicity using a double FDR approach as follows:

   i) Define $P_i^* = \min(p_{i1}, \ldots, p_{ik_i})$ and consider this as the representative p-value among the k_i adverse events in body-system i. This minimum value can be considered to represent the strongest safety signal in body-system i.

   ii) A first level of adjustment is made by applying the FDR to the P_i^* for the s body-systems. A second level of adjustment is applied within each body-system. That is the FDR is applied to $p_{i1}, \ldots, p_{ik_i}$ for each body-system $i = 1, 2, ..., s$. Define $\tilde{P}_i^*$ to be the adjusted BH p-value representative for body-system i and $\tilde{p}_{ij}$ be the FDR adjusted p-value for the k_i adverse events grouped as body-system i.

3. The DFDR flagging rule is to "flag" adverse event (i, j) if $\tilde{P}_i^* \leq \alpha_1$ and $\tilde{p}_{ij} \leq \alpha_2$ for specified values of α_1, α_2.[3]

We consider how the method works in practice. For Step 1, Mehrotra and Heyse consider the removal of low incidence adverse events to be important in reducing the multiplicity problem to the essential number of dimensions ,and they state that adverse events with low incidence can be investigated within Tier 3, i.e. examined by a clinician to see if they are important. However, what may be considered low incidence adverse events in terms of significance varies according to trial size and significance tests. Xia et al. detail this step as being the removal of adverse events with counts of less than 5 in both groups combined [60]. The ICH guidelines (§1.4.4) also allow for the possibility of removal of low severity or low count adverse events. It should be noted that this step removes some information from the procedure and potentially lessens the overall power.

The method does not include a mechanism for determining values for α_1 and α_2 for flagging adverse events such that the FDR is controlled at any particular level. In the paper values for α_1 and α_2 are chosen using a non-parametric bootstrap

[3] (i, j) is the j^{th} adverse event in body-system i.

sampling procedure, where the data from both groups is pooled and then repeatedly re-sampled (with replacement), with patients assigned to either treatment or control group at random, thereby "ensuring a null situation", to simulate many repetitions of the original trial. Mehrotra and Heyse set $\alpha_2 = \alpha$ and then used the bootstrap to determine the largest data-dependent $\alpha_1 \leq \alpha_2$ that ensured FDR $\leq \alpha$, under the hypothesis that the true adverse event incidence profile is the same for both groups.

We may also consider the asymptotic behaviour of the method by holding the number of groupings, s, fixed, but allowing the group body-system sizes, k_i, to increase. With $P_i^* = \min(p_{i1}, \ldots, p_{ik_i}) = p_{i(1)}$, Step 2, ii) in the method applies a BH p-value adjustment at the body-system level. For body-systems with no true alternative hypotheses and continuous independent test statistics, P_i^* is the first order statistic for a set of uniform variables on $[0, 1]$, so $P_i^* \sim \text{Beta}(1, k_i)$, and $\text{E}[P_i^*] = \frac{1}{k_i+1} \to 0$ as $k_i \to \infty$. The flagging rule $\tilde{P}_i^* \leq \alpha_1$ may be ineffective in this case for large body-systems, leaving just the rule $\tilde{p}_{ij} \leq \alpha_2$ to be applied.

In addition to the possibility that for large body-systems the method may not be as effective as hoped, Mehrotra and Adewale report a number of issues with the resampling step [3]. Firstly, it may provide an implementation or computational obstacle to the use of the method, and the suggested ad-hoc approach from [15] to avoid bootstrapping by setting $\alpha_1 = \frac{\alpha_2}{2}$ may in fact cause the actual FDR to be inflated to 2-3 times the size of the target level α. Secondly, based on work in [61], they report that there are a number of potential theoretical issues with the resampling approach.

Confidence intervals for the double FDR are given in [15, §8], where it is proposed that using the significance level corresponding to the largest FDR significant p-value in the family can be used to construct a single interval width for all comparisons. If the double FDR flags j adverse events, then confidence intervals based on $\left(\frac{j}{m}\alpha\right)$ would be computed for all adverse events. This is based on a paper by Williams et al. which addresses confidence intervals for the BH and Hochberg controlling procedures [62].

2.4.3 Double FDR Approach Procedure (2012)

In this section we give the details of the updated DFDR of Mehrotra and Adewale [3]. The new DFDR method addresses the issues with the version in [15] by removing the resampling step and changing the representative p-value for body-system

i to be:

$$p_i^* = \min(\tilde{p}_{i1}, \ldots, \tilde{p}_{ik_i})$$

where $\tilde{p}_{ij}, j = 1, ..., k_i$, are the adjusted BH p-values for body-system i. This allows the elimination of the re-sampling step because p_i^* can be employed as a valid p-value [63], [64]. Mehrotra and Adewale considered three alternatives for p_i^* but the chosen value generally provided more power in their simulation studies. It is interesting to note that because of the definition of adjusted BH p-values, the representative value is weighted by the body-system size, unlike the representative value in the original double FDR. We will see how this affects the behaviour of the DFDR in §D.1.

The new DFDR approach is as follows:

1. Initial dimension reduction step: Remove adverse event types for which the total incidence is so low (rare adverse events) that statistical significance at the conventional 0.05 level is impossible, even without a multiplicity adjustment.[4]

2. Apply a BH adjustment to the p_i^*, $1 \leq i \leq s$, where $p_i^* = \min(\tilde{p}_{i1}, \ldots, \tilde{p}_{ik_i})$, and let $\tilde{p}_i^*$ denote the corresponding adjusted BH p_i^*.

3. Let $F = \{p_{ij} : \tilde{p}_i^* \leq \alpha\}$. Apply a single BH adjustment to the p-values in family F. If $p_{ij} \in F$ then let $\tilde{p}_{ij}^{(F)}$ be its adjusted BH value.

4. The new DFDR approach flags adverse event (i, j) if $\tilde{p}_i^* \leq \alpha$ and $\tilde{p}_{ij}^{(F)} \leq \alpha$.[5]

In summary, the new DFDR method applies the BH-procedure at the body-system level first and, for any body-systems whose adjusted p-values are significant, gathers all their individual adverse event p-values into a single family F. It then applies the BH-procedure to F again generating adjusted p-values. Adverse events are flagged if both the body-system and individual adjusted p-values are significant at level α. There are two main differences between the original and the new DFDR procedure. The first is that the original double FDR has a representative value which is not in any way weighted by the size of the body-system, whereas for the DFDR it does in effect take the overall size of the grouping into account. The second is that original double FDR applies separate FDR adjustments to each of the

[4] In the paper the value 0.05 is specified rather than a significance level α [3].
[5] (i, j) is the j^{th} adverse event in body-system i.

body-systems flagged. In contrast the new DFDR applies a single FDR adjustment to the family of p-values collected from all the flagged body-systems in Step 2.

The authors make the following points regarding the new procedure:

1. The FDR for the new DFDR procedure is at most α at the *body-system level* regardless of how many body-systems contain at least one true signal.

2. While FDR control at the adverse event level is guaranteed under the global null, the actual FDR might theoretically exceed α in some rare cases. This is the same situation as with the original FDR. The authors consider this to be unlikely in practice.

2.5 Group Benjamini-Hochberg

In the Group Benjamini-Hochberg (GBH) procedure hypotheses are grouped and a probability weighting procedure, based on the estimated number of true null hypotheses in each group, is used to weight the p-values, effectively inflating the p-values of hypotheses which are considered not likely to be significant [30].

2.5.1 Notation

Following [30] we consider G families of p-values/hypotheses with m_{g0} the number of true null hypotheses, and m_{g1} the number of true alternative hypotheses, in family g. The total number of hypotheses in family g is then $m_g = m_{g0} + m_{g1}$. Let $p_{g,i}$ be the p-value for hypothesis i in group g.

The total number of true null and alternative hypotheses are given by: $m_0 = \sum_{g=1}^{G} m_{g0}$ and $m_1 = \sum_{g=1}^{G} m_{g1}$ respectively, with m, the overall total number of hypotheses considered, given by $m = m_0 + m_1$.

Let $I_{g,0}$ and $I_{g,1}$ be the sets of true null and true alternative hypotheses in family g respectively, with $I_g = I_{g,0} \cup I_{g,1}$ the set of hypotheses in family g. Let $I = \cup_{g=1}^{G} I_g$ be the total set of all hypotheses.

2.5.2 Group Benjamini-Hochberg Procedure

The GBH procedure is as follows:

1. Estimate the proportion of true null hypotheses in each group ($\hat{\pi}_{g,0}$) and the

overall proportion of true hypotheses ($\hat{\pi}_0$). If $\hat{\pi}_{g,0} = 1$ for all g we accept all the hypotheses and stop.

2. Calculate weighted p-values in each group g as follows:

$$P^W_{g,i} = \frac{\hat{\pi}_{g,0}}{1 - \hat{\pi}_{g,0}} p_{g,i}$$

   If $\hat{\pi}_{g,0} = 1$ we let $P^W_{g,i} = \infty,\ i = 1, ..., m_g$, effectively preventing the null hypotheses in the group from being rejected.

3. Pool and order the weighted p-values: $P^W_{(1)}, \ldots, P^W_{(m)}$. Here we need to maintain a mapping between the ordered weighted p-values $\left\{P^W_{(j)}\right\}$ and the unordered weighted p-values $\left\{P^W_{(g,i)}\right\}$ to ensure we can correctly identify the null hypotheses that the ordered values represent.

4. Find $k_{\text{GBH}} = \max\left\{i : P^W_{(i)} \leq \frac{i}{m}\alpha^W\right\}$ where $\alpha^W = \frac{\alpha}{1-\hat{\pi}_0}$. If such a k_{GBH} exists reject the hypotheses associated with $P^W_{(1)}, \ldots, P^W_{(k_{\text{GBH}})}$, otherwise do not reject any hypotheses.

The method is independent of the estimators chosen for $\hat{\pi}_{g,0}, \hat{\pi}_0$ but two are discussed in [30]. These are Two-Stage Method (TST) [43] and the Least-Slope Method (LSL) [65]. The GBH asymptotically controls the FDR at level α and is asymptotically at least as powerful as the BH-procedure.

We note that in Step 2 the weighted p-values $p^W_{g,i}$ may exceed 1.0 however this does not necessarily mean that the null hypothesis is never rejected as the value it is compared to in Step 4 is also weighted.

We can see that weighting the p-value has the effect of inflating the p-value if $\hat{\pi}_{g,0}$ is large ($1 - \hat{\pi}_{g,0}$ is small). So if the proportion of p-values believed to represent true null hypotheses in the group is large all p-values are inflated, meaning they are less likely to be flagged. The controlling value α, used in the comparison step of p-values for the BH-procedure, is also weighted, and this has the effect of changing the comparison part of the BH-procedure to use the size of those groups which have been "included", similar to the DFDR procedure. Here, it is the values of the estimators for $\pi_{g,0}$ and π_0 which determine how the groups are "included", in contrast to the DFDR, where a step actually excludes the groups.

We can see this more clearly by looking at what Hu et al. call the oracle case [30].

This is the case where we know the actual values $m_g, \pi_{g,0}, \pi_0$:

$$\pi_0 = \frac{1}{m}\sum m_g \pi_{g,0}$$
$$\alpha^W = \frac{\alpha}{1-\pi_0} = \frac{m\alpha}{m - \sum m_g \pi_{g,0}}$$

The threshold for calculating the rejected hypotheses is:

$$\begin{aligned} k_{\text{GBH}} &= \max\left\{ i : p^W_{(i)} \leq \frac{i}{m}\alpha^W \right\} \\ &= \max\left\{ i : p^W_{(i)} \leq \frac{i}{m - \sum m_g \pi_{g,0}}\alpha \right\} \end{aligned} \tag{2.5}$$

In this case $\sum m_g \pi_{g,0}$ is the total number of true hypotheses, so that the value $m - \sum m_g \pi_{g,0}$, used in the denominator in the comparison (2.5), has excluded the number of true null hypotheses. For the case where $\pi_{g,0} = 1$, groups with all true null hypotheses, these groups are excluded in sense that they are part of $\sum m_g \pi_{g,0}$ and their weighted p-values are never significant because they are set to an "infinite" value.

2.6 Comparison Between DFDR and GBH

As both the DFDR and GBH are grouped false discovery rate controlling methods it is of interest to compare them more closely. With m as the total number of hypotheses and m_0 as the total number of true null hypotheses, in GBH we are effectively looking to compare the weighted p-values, $p^W_{(i)}$, with $\frac{i}{m-\hat{m}_0}\alpha$, where as the DFDR process compares the unweighted p-values, $p^F_{(i)}$ in the reduced set of hypotheses F, with $\frac{i}{|F|}\alpha$.

$m - \hat{m}_0$ is an estimate of the number of false null hypotheses out of the total number of hypotheses, whereas $|F|$ is an estimate of the total number of hypotheses in groups that contain at least one false null hypotheses. We may surmise that in general $m - \hat{m}_0 \leq |F|$ and that $\frac{i}{m-m_0}\alpha > \frac{i}{|F|}\alpha$, leading to the possibility that in certain circumstances the GBH is the more powerful approach. This is consistent with the simulation results from [3]. However, the GBH comparison uses weightings which have the possibility to inflate or reduce the p-values, so there may well be circumstances where the DFDR approach may be more powerful. We may compare the methods using the criteria in Table 2.2.

Criterion	DFDR	GBH
Inclusion of Family	$\tilde{p}_g^{*1}$	$\hat{\pi}_{g,0}{}^2$
Effective total number of hypotheses tested	$\|F\|$	m
Denominator of test threshold	$\|F\|$	$m - \hat{m}_0$
p-value for comparison	$p_{(i)}$	$p_{(i)}^W$

Table 2.2. Comparison of DFDR and GBH procedures for controlling the FDR.

[1] Representative group p-value.
[2] Estimated proportion of true null hypotheses in group.

The GBH uses different estimators for each group to effectively decide group inclusion and increase power. In contrast the DFDR uses a single estimator or threshold for group inclusion, hoping to exclude groups which aren't significant and thereby increase the power.

Both GHB and DFDR have properties that may make them useful for analysing clinical trial safety data. However, as GBH is designed for use with sparse data there are certain circumstances under which we may expect to see some error inflation. In particular use of the least-slope method (LSL) for estimating the proportion of true hypotheses in a group of size one always returns the value 1.0, effectively excluding that family from consideration. More generally we may expect to see error inflation in GBH for groupings where most null hypotheses are false. In this case the remaining weighted p-values will be reduced substantially. For example: if the estimated proportion of true null hypotheses is $\frac{1}{5}$ then we have an inflation factor of $\frac{\frac{1}{5}}{1-\frac{1}{5}} = 0.25$. In this case the remaining p-values will be weighted to a quarter of their actual value. In the application of the method in [30], where a cluster analysis has already gathered likely genes together, this may not be a problem. In the case of predefined groupings, such as body-systems in clinical trials, where it may be that quite a number of adverse events in a body-system have raised treatment rates, this reduction in the size of the p-values may result in other adverse events in the group, which do not have raised rates, being flagged.

Where most hypotheses are true in groups we may also expect that the GBH may give relatively poor performance detecting adverse events. In this case there is the possibility that the p-value weighting will overwhelm the remaining hypotheses' p-values. For example, if the estimated proportion of true hypotheses is $\frac{4}{5}$, the p-

value inflation factor is $\frac{\frac{4}{5}}{1-\frac{4}{5}} = 4$, leading to inflated weighted p-values and possibly leading to adverse events with raised treatment not being detected, unless they have a very strong signal.

2.7 Discussion

The procedures we have reviewed control error rates for hypothesis testing so, in a sense, they don't have anything to say directly about clinical safety data. The clinically important issues of severity, timings, and censoring don't have any direct impact on these methods. However, any of the procedures when applied to safety data do directly control for multiple comparisons and provide a definite flagging mechanism for adverse events whose occurrence may be related to treatment.

The extensions to the more common error controlling procedures which use groupings are of particular interest to us. For the DFDR the main idea is that, in some sense, the body-system has to be significant, and an adverse event within a significant body-system, has to be significant, before it is flagged. The method is designed for non-rare Tier 2 events, with rare or serious events (Tier 3) requiring a different type of evaluation. This fact is explicitly encoded into the procedure where very low incidence events are removed before the procedure proper is applied. The GBH approach is similar to the DFDR in that the estimated weights of the p-values are based on the estimated proportions of true hypotheses they contain, effectively eliminating any groups where the estimate is 1.0 or close to it. A number of asymptotic error controlling and power results exist for the GBH.

The DFDR method can be considered to lie somewhere between the original BH-procedure and a grouping/proportion estimating procedure such as the GBH. It is a particular implementation of a general approach where a decision is made to exclude certain groups of hypotheses because there are believed to be not significant. We assume the method of identifying which groups to exclude is based in some way on an analysis of the p-values within those groups, and we are guaranteed that, theoretically, the FDR will be controlled at the required level. The interesting question is whether, and under what circumstances, it provides a gain in power compared to other procedures. Unlike the GBH, which uses estimates of the proportions of true null hypotheses within groups, in their method Mehrotra and Adewale only take into account whether a group may contain significant hypotheses or not. If a group is determined to contain only true hypotheses then that group is excluded. Once this is achieved the groupings are ignored. They do not

make any estimate of the number of true hypotheses in their controlling procedure beyond this inclusion/exclusion criteria.

As stated in §2.4.2, it could also be argued that removing "non-significant" adverse events from the analysis is ignoring important trial information, possibly leading to a reduction in power. This is in contrast to modelling methods, such as the Berry and Berry model (§3.6.1) [5], who also use the body-system and apply their method to Tier 2 events. In these methods there is no need for this restriction and, in fact, the low incidence of some adverse events within a body-system may have an important effect on the models, and hence any conclusions drawn. One solution to the issue of low frequency events is to aggregate within body-system.

Mehrotra and Adewale state that an advantage of the DFDR, and in fact of any error controlling procedure, is that we understand more about the error structure of the method than we do for modelling approaches such as the Berry and Berry model. For Bayesian modelling approaches little is known about the FDR or power properties. This can be seen as an advantage of the DFDR or GBH approach. In fact the Mehrotra and Adewale definition of power could possibly be used to generate a decision rule for the Bayesian modelling approaches of [5] or [6]. Xia et al. have looked at this or similar in their simulation studies [60]. Bayesian and other types of modelling approaches to safety are reviewed in Chapter 3.

In conclusion, the DFDR and GBH are purely error controlling procedures. They do not explicitly take into account any of the properties of clinical trials or adverse events we may wish to include. If we wish to take into account the timing and severity of adverse events, patient censoring, sub-group analysis, or trial interim analysis, then this must all be done in some manner before the method can be used. As the methods require p-values we must have some test statistic or model which can take into account any properties we wish to include. The DFDR or GBH approaches do have the advantage of giving direct rules for flagging adverse events and, based on simulation studies, we may be able to give some sort of estimate of their power.

The DFDR and GBH form part of the simulation study of methods in Chapter 5. There we will compare the power and error rates of the methods with each other and some additional grouped methods based on modelling approaches which we identify in Chapter 3. The DFDR and GBH are also compared directly in Appendix D.

Chapter 3

Modelling Methods for Safety Analysis

3.1 Introduction

Trial safety data may be considered as the occurrence, or the recurrence, of adverse events of varying severity in patients, which may have durations, may be related to each other, to patient censoring, and to terminal events in the trial. In this chapter we review some approaches to safety analysis in clinical trials, concentrating on modelling to analyse the occurrences of adverse events. Similar to the approach in Chapter 2, we look to take into account the issue of multiple types of events, and we are also interested in the possibility of performing interim data analyses.

ICH Guidelines (§1.4.4) suggest survival analysis as a potential approach to analysing safety data, but often common implementations of standard models for survival data tend to consider one type of event only. We are interested in modelling all of the adverse events which occur in a trial. As we will see below, one approach to handling multiple events in models is to group similar adverse events by body-system, and use this additional information to shrink non-significant adverse event effect differences towards zero. Bayesian modelling approaches are often used for this, although other approaches are possible. In fact, for a Bayesian analysis, it is often considered that the appropriate choice of prior distributions is able to provide multiple comparison robustness, thereby removing the need for any further error control [27], [66]. In this review we will discuss a number of methods, including grouped methods, from the recent literature, including a more detailed discussion of the following papers:

- **S. M. Berry and D. A. Berry.** Accounting for Multiplicities in Assessing Drug Safety: A Three-level Hierarchical Mixture Model [5], §3.6.1.

- **William DuMouchel.** Multivariate Bayesian Logistic Regression for Analysis of Clinical Study Safety Issues [6], §3.6.1[1].

- **O. Siddiqui.** Statistical Methods to Analyze Adverse Events Data of Randomized Clinical Trials [4], §3.3.1.1.

Berry and Berry [5] and DuMouchel [6] take a Bayesian approach and are, to quote DuMouchel, "similar in spirit", while Siddiqui [4] takes a classical or frequentist approach. Berry and Berry use the body-system, or system organ class grouping, described in §1.9, when analysing adverse event incidence data. DuMouchel takes a more ambitious approach, including covariates and treatment covariate interactions in his model, stressing that the adverse events included in the analysis should be in some sense medically related, but not directly using a body-system or similar grouping. Siddiqui, on the other hand, analyses adverse event incidence data using the non-parametric Mean Cumulative Function, which also allows groupings of adverse events if required, and is the only one of the three papers which directly addresses the timings and cumulative presence of the adverse events [4].

Before discussing the above papers in detail, it is worth considering some of the different philosophies behind these and other approaches to analysing clinical incidence data, such as the DFDR [3] discussed in Chapter 2. Bayesian modelling approaches, such as those used by [5] and [6], or their frequentist counterparts, fit particular models to the data. In a Bayesian context, they use the data, together with certain prior distribution assumptions, to make statistical statements based on the derived posterior distributions. These types of Bayesian models do not directly address the issue of Type-I or Type-II error rates or power for testing multiple hypotheses about clinical outcomes. They may be considered in some sense exploratory approaches. In a clinical context though, where we need to decide if a particular treatment is associated with raised adverse event levels, the models must be used to help determine if this is indeed the case. Similarly, frequentist modelling approaches, for example multi-level modelling, also make assumptions about suitable model forms, and use the data to estimate the model parameters. Again, while the model may not directly address the issue of multiple comparisons or tests, it must be used in some way to make a decision about the clinical effects of the treatment and the adverse event occurrence rates.

[1] Originally a presentation: Du Mouchel. Multivariate Bayesian Logistic Regression for Clinical Safety Data) at the 4th Seattle Symposium in Biostatistics (2010).

On the other hand, the approaches taken by the methods described in Chapter 2 directly address multiplicity issues. In particular, the DFDR and GBH approaches use an assumed relationship among the data (body-systems) to help directly control the error rates. Apart from this assumption, and the assumption that the Benjamini-Hochberg False Discovery Rate controlling procedure [31] can be applied, there are no further assumptions, and statements can be made directly about the error rates.

Siddiqui takes a different approach, using the non-parametric Mean Cumulative Function (MCF) [4]. This investigates what the average occurrence of adverse events tells us about the differences between treatments. The MCF allows for the analysis of multiple re-occurring adverse events and comparisons of different groupings within the clinical data, including by body-system. However, applying the MCF to multiple subgroups and performing hypothesis testing raises issues of multiple comparisons. Siddiqui suggests that, due to the nature of safety analysis in clinical trials, the approach should be explorative rather than based on hypothesis testing, and the MCF is a suitable tool for this. Following this approach, Siddiqui does not directly address the issue of controlling Type-I or Type-II error rates.

Other methods reviewed in this chapter include approaches based on survival analysis, recurrent event analysis, and longitudinal analysis. Many of the methods we will discuss are designed to be used in an exploratory or confirmatory sense rather than as direct replacements for existing procedures, although Berry et al. [67] advocate an overall Bayesian approach to the analysis of Clinical Trials as they move through their different phases. Advocacy for this type of approach is growing in the literature [68], [69], [70].

Some of the Bayesian approaches we consider here [5], [6], and the approach used by Siddiqui [4], are more in line with the derivation of a safety profile of a particular treatment based on the data available and, possibly, an assumed model or structure for the data, or choice of prior distributions.

Despite the differences in philosophy between modelling approaches, such as [5], and error controlling approaches, such as [3], there is a common underlying assumption that dependencies or correlations that may exist in the data may be used in a statistical analysis.

3.2 Survival Analysis Models

Survival Analysis is the analysis of data in the form of times from a well-defined time origin until the occurrence of some particular event or end-point. Parametric, semi-parametric, and non-parametric approaches are all possible and tend to focus either on the event time, such as the Kaplan-Meier estimate of the survival curve, or the event rate, such as the Cox model [71]. In this section we look at the parametric and semi-parametric approaches to survival analysis.

Two import concepts when modelling survival data are the survivor (survival) and hazard functions. If the survival time of an individual is modelled by a random variable T, then the survivor/survival function $S(t)$ is defined to be [72]:

$$S(t) = P(T > t) \tag{3.1}$$

and the hazard function to be [72]:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T > t)}{\delta t} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t) \tag{3.2}$$

assuming T is a continuous variable.

The survival time of an individual is said to be *censored* when the end-point of interest has not been observed for that individual. The main approach to fitting survival models is through likelihood estimation. If we have non-informative[2] right censored[3] data $(t_1, \delta_1), ..., (t_1, \delta_1)$, where $\delta_i = 1$ for a real event time, and $\delta_i = 0$ for a censored time, the likelihood is:

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} f(t_i)^{\delta_i} P(T > t_i)^{1-\delta_i} \tag{3.3}$$

where $f(t)$ is the density function of T.

Parametric approaches to survival analysis assume a form for $f(t)$, with the Weibull distribution being one such form commonly used for this purpose, as its shape and scale parameters allow it to model many different hazard functions [73].

Assuming a parametric form for T has a number of drawbacks. Among these is the problem of justifying the chosen distribution, leading to corresponding diffi-

[2]The event and censoring time are independent.

[3]Right censored data occurs when the event occurs after the last known survival time.

culties in model checking. Non-parametric and semi-parametric approaches do not have this disadvantage. The Cox proportional hazards model is one such semi-parametric approach which is widely used in survival analysis [71]. Here no probability distribution is assumed. The hazard function is assumed to have the form:

$$h(t) = h_0(t) \exp(\beta_1 x_1 + ... + \beta_n x_n) \tag{3.4}$$

where $x_1, ..., x_n$ are explanatory variables recorded at the time of origin, $\beta_1, ..., \beta_n$ are model parameters to be estimated, and $h_0(t)$ is known as the baseline hazard function (to which all others hazard functions are proportional). Lack of knowledge of $h_0(t)$ means that maximum likelihood estimators for the parameters cannot be calculated, but a partial likelihood approach may be used in its place [74], [75], [76]. A number of extensions to the Cox model have been introduced, including time-dependent covariates and stratification [72]. An example of the use of the Cox model to investigate adverse events is [77] where O'Neill considers the relationship between the occurrences of adverse events (what he terms toxicity) and dose, using a proportional hazard model approach adapted to these competing risks (the competition between toxicity and survival).

Bayesian approaches to (semi-parametric) survival analysis, or the analysis of recurrent events, are discussed by Kalbfleisch [78], Burridge [79], Clayton [80], Sinha [81], Ibrahim et al. [82], Kalbfleisch and Prentice [83], Duchateau and Janssen [84], Dunson and Herring [85], and Shaban and Mostafa [86]. Kalbfleisch considers the proportional hazards model, treating the cumulative baseline hazard function as a "nuisance" parameter by dividing the time domain into a number of disjoint intervals, and using Dirichlet and Gamma processes as priors [78]. Burridge extends this model using an empirical Bayes approach, and, additionally, looks at tied and grouped data [79]. Clayton assumes Gamma priors for the baseline hazard [80], an approach which is also used by Duchateau and Janssen [84]. Sinha [81] considers a similar model to [78] for multiple events but includes a frailty term. Shaban and Mostafa, who look at an additive hazard frailty model, assume a piecewise linear baseline hazard function with a Gamma prior for its parameters [86]. This is a slightly different approach than using a Gamma process as a prior. Dunson and Herring look at the choice between multiplicative and additive models as a Bayesian model selection problem [85]. They concentrate on inference while accounting for model selection. They also choose a piecewise constant baseline hazard with Gamma prior. Kottas, on the other hand, looks at non-parametric Bayesian estimation using a Dirichlet process mixture, with a Weibull kernel, for

modelling the survival distribution [87].

The are a number of approaches to extending survival analysis to multiple events, including random effects and frailty based models, multiple event Cox models, stratified models with different baselines hazards for each event type [88], [89], and similar marginal models [90]. While the random effects and frailty models can account for relationships between the model parameters, the other approaches all require some method of accounting for correlations between the model parameters, for example by an estimator such as the jackknife [88]. In this context multiple events can mean multiple occurrences of the same event (recurrent events) or the occurrence of multiple types of event. Recurrent event analysis is reviewed in §3.3.

There are a number of issues we may need to consider when using these types of models. Many subjects may experience only a small number of adverse events over the course of a trial and, as survival analysis deals with individual subjects rather than summary level data, this may lead to many censored observations. In the case of random effects or frailty models, for large trials there is also the possibility of large numbers of parameters making model fitting difficult. In clinical trials we also have the complication that death is a terminating adverse event for a patient, essentially there are competing risks for the events at the patient level.

3.3 Recurrent Event Analysis

Methods from the theory of recurrent events can also be used to analyse adverse events. There are many methods available for modelling such recurrent data, including, but not limited to, inter-event time models, time to event models, marginal methods based on multivariate failure time data, general intensity modelling, and methods based on event counts. Andersen et al. provides the relevant theoretical background for a large number of the papers in this area [91]. Cook and Lawless describe many of these and other approaches to recurrent data analysis, including two-state and Markov process models [74]. They consider that there are really three general approaches to recurrent event analysis:

1. Count models based on the intensity function (conditioned on the process history) defined as:

$$\lambda(t|H(t)) = \lim_{\Delta t \downarrow 0} \frac{\mathrm{P}(\Delta N(t) = 1|H(t))}{\Delta t}$$

where $N(s,t)$ is the number of events which occur in the interval $(s,t]$, $N(t) = N(0,t)$ for $t > 0$, $\Delta N(t) = N(t + \Delta t^-) - N(t^-)$, and $H(t) = \{N(s) : 0 \leq s < t\}$, $t > 0$, is the history of the process.

2. Gap times models based on hazard functions (conditioned on the process history). If the time between events has a common distribution W with density $f(w)$ and survivor function $S(w) = \mathrm{P}(W \geq w)$, then this is defined in a similar way to the hazard function in survival analysis (3.2):[4]

$$h(w) = \lim_{\Delta w \downarrow 0} \frac{\mathrm{P}(W < w + \Delta w | W \geq w)}{\Delta w} = \frac{f(w)}{S(w)}$$

3. General Intensity functions (conditioned on the process history) which may in some sense be considered a combination of the intensity and hazard function approach.

These cover everything from simple Poisson models and renewal models, to semi-parametric and non-parametric additive and multiplicative models, and accelerated failure time models. Inference is generally carried out using the likelihood function or, for semi-parametric models, the partial likelihood.

Models for multitype recurrent events can be constructed in a straightforward way using the intensity function approach. Incorporating dependencies between the event processes is possible through stratification and time dependent covariates [74]. These types of dependencies can also be achieved by introducing a random effect into the model [74]. So, for example, if there are n subjects and J event types we could model the intensity function as:

$$\lambda_{ij}(t|H_i(t), u_{ij}) = u_{ij}\lambda_{ij}(t|H_i(t)) \quad i = 1, ..., n; j = 1, ..., J$$

where u_{ij} is a random effect.

There are many possible extensions to these general recurrent event formulations. Cook and Lawless consider the case of adverse events where a terminal event for a subject is dependent on the adverse events [92]. Their model is a multivariate counting process which is a joint model for recurring and terminal events. They consider both non-parametric and semi-parametric approaches using rate and mean cumulative functions. Wang et al. [93] and Rosenkranz [94] also con-

[4] A number of slightly different definitions of the survivor function are used in the literature.

sider the case of adverse events where a terminal event for a subject is dependent on the adverse events. They assume a multivariate counting process where the relationship between recurrent events and the terminal event is via a latent variable. The recurrent event process, conditional on the latent variable, is a non-stationary Poisson process, and a semi-parametric form is taken for the rate function. A key assumption is that conditional on the latent variable the terminal event time and the counting process are independent. A multiplicative intensity model is assumed. Rosenkranz [94] describes three models which directly model the dependence: a parametric model based on a method of Heitjan [95], a semi-parametric bivariate local shift model [96], and a copula based dependence model [97]. Cook et al. take an approach which conditions on enough of the event history to render the censoring conditionally independent, and analyse marginal features by averaging over prior event history [98]. This can be achieved by an approach based on multi-state Markov models. They model using marginal rate functions, marginal survivor functions for event times, and partially conditional rate functions employing Markov assumptions. Inverse probability of censoring weighted (IPCW) versions of the estimators are also given to provide robustness to event-dependent censoring. Frailty model approaches to event analysis are discussed in [84], and discussions of the frailty model for recurrent events in the presence of a terminating event are given in [99], [100], [101].

Lawless and Nadeau describe methods for the analysis of recurrent events using a counting process and non-parametric estimation of a common Mean Cumulative Function (MCF), under the assumption of a Poisson process generating the events [102]. They derive an estimate for the variance of the MCF and extend the approach to include covariates. They consider the comparison of two such MCFs and derive a test statistic, an approach which they say is generalisable. Nelson describes a non-parametric approach based around the MCF [103]. Nelson's approach and the MCF in general are described in more detail in §3.3.1. More recently, Wang and Quartey [104], [105] have considered non-parametric and semi-parametric approaches for the Mean Cumulative Duration function (MCD), an approach similar to the MCF but concerned with the duration as well as the occurrence of the adverse events, extending their results to include dependent censoring using the IPCW method [98]. Zhao and Zhou use the MCF approach to model gap times between events [106].

3.3.1 The Non-Parametric Estimate of the Mean Cumulative Function

The non-parametric estimate of the Mean Cumulative Function (MCF) is one approach to exploring the occurrence of adverse events as recurrent events, while making minimal assumptions. Siddiqui provides an example of the use of a non-parametric estimate of the MCF in an analysis of adverse events in clinical trials [4]. Nelson describes the non-parametric MCF in detail [103].

The Mean Cumulative Function (MCF), $M(t)$, is defined as the mean of the distribution of the number of events at time t. More precisely if $N(t)$ is the number of events to have occurred by time t, then $\mathrm{M}(t) = \mathrm{E}[N(t)]$.

The MCF is a representation of a simple counting process of a cumulative number of discrete events. Nelson describes a non-parametric estimator, $\hat{M}(t)$, for $M(t)$, where $\hat{M}(t)$ is the estimator of the mean cumulative number of adverse events up to time t [103]. The estimate involves no assumptions about the form of $M(t)$. At time t, a fraction of subjects have accumulated one occurrence, a fraction two recurrences, and so on. This distribution differs at time t and has a mean $M(t)$. The estimate is the pointwise average of all subjects' cumulative adverse event curves passing through the vertical line at each time t. For a large sample the estimate of the MCF is usually regarded as a smooth curve. Cook and Lawless have applied a similar non-parametric method to analyse recurrent safety data in clinical trials [74, §3.4].

The estimate $\hat{M}(t)$, and its confidence limits for recurrence data, are analogous to the Kaplan-Meier (KM) estimate and Greenwood's variance for life data [4]. Plots of $\hat{M}(t)$ and confidence intervals versus time t yield information such as the number of cumulative events expected by time t, whether the rate of occurrence is increasing, decreasing or constant, and whether the two groups differ significantly in the expected number of events. In contrast, Kaplan-Meier analysis includes only time to the first adverse event. The derivative $\hat{m}(t) = \frac{d\hat{M}(t)}{dt}$ is called the instantaneous recurrence rate, or intensity rate, of an event at time t. The assumptions for the use of the MCF are as follows:

1. The target population is clearly specified and sampled.
2. The sample units are a simple random sample from the target population.

3. Random/Non-informative censoring: the cumulative history functions of all sample units are statistically independent of their censoring ages.

The MCF can be applied to exact and interval age randomly censored data, and is unbiased for exact age data. The MCF for interval age data is biased, but the bias is likely to be small compared to the statistical randomness in most applications [103]. There are a number of possible confidence intervals which can be calculated for the MCF, all of which require certain assumptions (e.g. normal approximation [103]). Nelson suggests the possibility of using the jackknife, bootstrap, or other re-sampling methods to obtain an empirical sampling distribution for the limits, as these do not entail a normal approximation. This has the advantage that for count data, such limits are always positive, whereas normal based confidence intervals may have a negative lower limit.

One interesting property of MCFs is that they are additive, so they can be applied to different combinations of events. This allows us to consider the "total" MCF, and then the individual contributions of each event type to the overall total.

There are a number of methods suggested for comparing the MCFs from two samples. This is similar to comparing two (or more) survival curves. Point-wise approaches include comparison using calculated confidence intervals, permutation tests of the two MCFs at a particular time, all pairwise differences (multiple comparison issues would need to be handled by some form of error controlling procedure), analysis-of-variance comparisons (this requires a normal approximation and results in a chi-squared test), and simultaneous intervals where wider simultaneous intervals which have approximate probability that all difference confidence intervals enclose their corresponding true values. Under this assumption, if an interval does not enclose zero we have stronger evidence of a real difference. This requires a Bonferroni type correction in the interval calculation [103, Chapter 7, pg: 115].

It is also possible to consider an overall MCF comparison using weighted differences over their common age ranges. In this case the sampling distribution of the test statistic must be obtained through an approximation, using permutation methods or a simulation method like boot strapping. The two-sample statistic is analogous to the Hotelling T^2-test, and the k-sample statistic is just an extension of the two-sample approach [103]. Parametric based comparisons are also possible.

We should note that many of the comparison methods make the assumption that the two samples being compared are statistically independent. This may not be

the case if, for instance, we were comparing two adverse events within one body-system. The individual cumulative histories of the subjects may be viewed on an event chart for graphical comparison [107].

3.3.1.1 Using the MCF to Analyse Adverse Event Data from Randomised Clinical Trials

In his 2009 paper, Siddiqui analyses randomised trial data using the non-parametric MCF [4]. While there are several proposed parametric counting process models to analyse recurrent adverse event data, which we have reviewed above, Siddiqui considers that these models may make assumptions that are either unrealistic, or unverifiable, in safety analysis for clinical trials, for example the assumption of constant hazard rate over time. These types of assumption are not required by the non-parametric MCF approach. With a smaller number of assumptions, the non-parametric mean cumulative function makes use of all the adverse event information of all randomised subjects in a trial.

Siddiqui briefly discusses the withdrawal of a number of drugs (specifically the COX-2 inhibitor Vioxx®, Bextra®, and Rezulin®) due to higher risk of heart attack or stroke, and also withdrawals due to significantly more adverse events in women than men [108], possibly for physiological reasons. He considers the possibility of a flawed safety analysis for these drugs being due to low power or short trial duration. The MCF is one possible tool which may be suitable for exploring adverse event occurrences for these types of situations as it may give some idea of the occurrences of events over time.

In any safety analysis we need to consider that recurrences of adverse events of the same or different kinds (which might be correlated with adverse events which have previously occurred) are often seen in clinical trials, and all subsequent occurrences of adverse events might be correlated with the time to discontinuation of a patient in a trial. Recurrent adverse events consist of the inter-event times of repeated adverse events of the same or different type for each subject, and times between adverse events within a subject are not necessarily independent. Further, it is recognised that some clinically important adverse events occur on a delayed basis, and Phase II and Phase III trials may fail to capture these. However, adverse events which do show up early in a trial may individually or collectively be good indicators of these delayed adverse events, and the possibility exists that one type of adverse event, or several adverse events jointly, might lead to another type of delayed adverse event in the future. For example, heart abnormalities leading to

a heart attack. The goal here is to understand, and possibly use in a predictive sense, the cumulative prevalence, as well as the trajectories, of each adverse event type over the study period, rather than just analysing the final totals.

Siddiqui compares the MCF approach to the Crude Incidence Rate and the Exposure-Adjusted Incidence Rate (§2.2) by applying the methods to clinical data from a trial of a COX-2 selective, non-steroidal anti-inflammatory drug (NSAID). The trial was a 12 week double blind trial where the start date and end date for each clinically apparent adverse event was recorded for each subject, some of whom had recurrent adverse events. Both serious and non-serious clinical adverse events, including recurrent events, were analysed together, using the MCF approach to compare the safety profiles of the study drug versus placebo. A comparison of the safety profiles was also carried out for all adverse events related to eight organ systems. This statistical analysis of observed adverse events as they relate to a particular organ system may provide additional information to indicate or detect potential safety issues, and the grouping of adverse events on an organ system basis is very reminiscent of the body-system approaches of [3] and [5].

There is no reference for the trial in the paper and the complete data set is not given. However, Siddiqui reports that among the eight organ systems it was found in the analysis that the study group had a different safety profile of cardiovascular related adverse events compared to the corresponding profile for the placebo group.

The results from the CRI and EAIR analysis, without controlling for multiple comparisons, and with adverse events aggregated for organ class, were that:

1. The CRI and EAIR for all adverse events, and for cardiovascular adverse events, were higher in the study group compared to the placebo group.

2. The rates of cardiovascular adverse events were even higher for females compared to those for males.

3. The rates for organ systems other than cardiovascular were similar between the study drug and placebo, indicating that the overall differences between the arms was due to cardiovascular adverse events.

For all adverse events, the MCF estimate, $\hat{M}(t)$, showed that while there were more events in the drug group than control, the two curves had become parallel after a few days, indicating that intensities of adverse events had become constant and

were similar. The confidence intervals calculated at the end of the trial period for $\hat{M}(t)$ overlapped. However, for the study drug group the MCF estimate indicated that more females than males suffered adverse events, with the confidence intervals at the end of the study separated.

On the other hand, the MCF estimate for cardiovascular adverse events showed a higher intensity rate for treatment than study, with a subject from the study group having more than two times higher cardiovascular related adverse events on average than a patient from the placebo group. Although by the end of the study the confidence intervals for the different curves still overlapped, the trajectories indicated that the confidence intervals would become separated for pro-longed drug use, provided the intensity rates remained similar. A similar pattern existed in the drug study group for females versus males, while there were more adverse events for females the confidence intervals overlapped, but continued use would indicate that they would be come separated.

A weakness of the crude incidence rate and exposure-adjusted incidence rate is that they do not provide the trajectories of adverse event occurrence over the study period. In particular, as they are incidence related statistics, they ignore the occurrence of more than one adverse event of the same type which may occur for the same subject. Siddiqui considers that these are important to understanding the cumulative history of all adverse events over the study period. The MCF analysis has suggested that the intensity of the cardiovascular adverse events has increased with a higher rate for the study group as opposed to the control group, in particular for females. This could also be considered as a future indicator of delayed serious adverse events, such as heart attacks, due to prolonged use of the drug, and further investigation is required here to understand why the rate was higher for these adverse events.

The argument is that the MCF, as a simple approach to representing a stochastic counting process of a cumulative number of discrete events, not overly dependent on statistical assumptions, can be used to understand the safety profiles of a study drug, including gender specific safety profiles. In fact any possible subgroup analysis is possible, and as there is no assumption of constant hazard rate, the intensity of hazard rates can be compared throughout the study period. However, confidence intervals calculated based on a normal approximation may not be appropriate for small frequency or individual adverse events. For safety purposes, the Type-II error rate in clinical trials should be as small as possible, but, due to low power, this is

not always possible to achieve. So this type of exploratory approach to safety analysis, trying to understand the expected safety profile of a drug under prolonged use, rather than performing hypothesis tests, may be of benefit.

3.3.1.2 Further Discussion

The MCF method can be used in multiple contexts. It can be used with any population sub-groupings within the trial, either specified in the trial protocol or not, for an explorative analysis for any adverse event or groupings of adverse events. In common with [3] and [5] it can be used for body-system based analysis.

Use of the MCF does not say anything about control of Type-I or Type-II errors, and Siddiqui does not advocate a hypothesis testing approach to safety analysis, but rather uses the MCF and its graphical representation to gain an insight into differences between the occurrences of adverse events for different groups [4]. Analysis of confidence intervals, and the respective intensity levels of the MCFs as indicated by the graphs, are used to gauge the possibility of differences between the different groupings, and also their future behaviour. For example, with regard to cardiovascular events in the trial considered, Siddiqui says that while the confidence intervals for the two groups overlap on the last day of the study, they are expected to be separated for prolonged use of the drug due to the higher intensity rate of adverse events for females during the study period. This continued intensity could lead to further, possibly serious, adverse events for females in the future. Even though the MCF estimate is non-parametric, and no relationships are explicitly modelled, this type of prediction, and use of possible correlations among the adverse events, is, in a sense, "encoded" in the MCF, just by the inclusion of all the event occurrences, and the chosen sub-groupings analysed. The MCF plot shows an increasing intensity for the study group, and it is this increasing intensity that may indicate future delayed adverse events, such as heart attacks, which may not have occurred by the end of the study, but may occur some time later. In this sense the groupings of the adverse events are important when using the MCF.

The use of the MCF in an explorative manner is not dissimilar to a Bayesian approach in the sense that the data is telling us something about the occurrences of adverse events, but we are not performing any hypothesis tests. However, the Bayesian approach is limited by the definition of a model, choice of covariates, parameters, and prior distributions, whereas the MCF could in theory be applied to any combination of adverse events and subgroups. Even without explicit hypothesis testing, this indiscriminate use of the MCF would be open to the dangers

of any analysis that uses unadjusted multiple comparisons. There is also a possible danger in crossing from exploratory analyses to data-dredging [109]. However, when analysing the clinical data, Siddiqui looks at a limited number of groups: all adverse events, cardiovascular adverse events, and females versus males. Other organ groups, which had similar rates between treatment and control, were ignored. So in a sense his analysis was fairly limited, and it could be said that he effectively dealt with the issue of multiple comparisons by only looking at certain predefined groupings. This type of analysis, limited to certain groups of events, is often specified in a trial's protocol.

The non-parametric MCF is an interesting approach which extends the analysis of count data to include repeated events and their timings. It is primarily exploratory. However, the assumption of random censoring may not be valid for clinical trials. Wang and Quartey use IPCW to extend the MCD approach to dependent censoring [104]. It is possible that a similar approach could be applied to the non-parametric MCF approach of Siddiqui, although we do not do that in this study.

In addition to not catering for non-random censoring, the non-parametric MCF approach has a number of other limitations. It does not provide a framework for handling multiple comparisons, it is mainly descriptive, and while inference is possible, this requires parametric assumptions. It also does not cater for the possibility that a large number of adverse events may be from a single subject. The indications or predictions of future occurrences or re-occurrences of adverse events are based on finding an increased intensity rate in one arm of the study. In order to find such indications, we may need to analyse the MCF of a set of related adverse events. Any indications are then dependent on the choice of groupings (or body-systems) analysed. Consequently, the non-parametric MCF does not provide a straightforward method for flagging adverse events, although it is possible to test hypotheses. However, multiple hypothesis tests are not automatically adjusted by the approach, so care is needed when applying the MCF as described.

3.4 Random and Mixed Effects Models

The introduction of random terms in linear and generalised linear models has greatly increased the scope of these methods. Mixed effects models are particularly useful when the number of potential model parameters is large, but the number of observations per parameter is limited. For the standard General Linear Model (LM) with y an $N \times 1$ vector of responses, X an $N \times p$ design matrix for fixed

effects $\boldsymbol{\beta}$, and Z an $N \times q$ design matrix for random effects $\mathbf{b}$, the Normal linear mixed model is defined as:

$$\begin{aligned} y &= X\boldsymbol{\beta} + Z\mathbf{b} + e \\ e &\sim N(0, \Sigma) \\ \mathbf{b} &\sim N(0, D) \end{aligned}$$

where e and $\mathbf{b}$ are independent, and Σ and D are covariance matrices.

In a similar fashion the Generalized Linear Mixed Model (GLMM) is a straightforward extension of the Generalized Linear Model (GLM) where, conditional on $\mathbf{b}$, the model is a GLM, and if $\mu_i = \mathrm{E}[y_i|\mathbf{b}]$ the link function is given by [76]:

$$h(\boldsymbol{\mu}) = X\boldsymbol{\beta} + Z\mathbf{b} \tag{3.5}$$

These models are well understood. The framework for analysing them is discussed in [110], with associated software packages such as `lme4` [111] and `nlme` [112] available for `R` [113]. An alternative model fitting approach is given in [114]. There are a number of similarities between these models and Bayesian models, although the philosophy is quite different. The use of a standard generalised linear model based approach means that current implementations are limited to exponential families of random variables (e.g. Binomial or Poisson), with few choices for the random effects (usually just Normal). As well as being useful in themselves, these models often form a starting point for longitudinal approaches (§3.5) and more complicated models, which can also be analysed by Bayesian methods. We will see a number of these types of approaches described below when we look at Bayesian approaches to safety analysis (§3.6.2.7). Random effects can also be incorporated into survival analysis. In the `R` package `coxme` [115], Therneau fits the survival model with hazard function given by:

$$\begin{aligned} h(t) &= h_0(t) \exp\left(X\boldsymbol{\beta} + Z\mathbf{b}\right) \\ \mathbf{b} &\sim N(0, \Sigma) \end{aligned} \tag{3.6}$$

where, conditional on $\mathbf{b}$, observations are independent.

For the type of end of trial data that we see in §1.8, a Binomial GLMM could potentially be suitable for analysing this data, taking into account the groupings given by the body-systems or system organ classes. The probabilities of the adverse

events for such a model would have the general form:

$$\begin{aligned} \operatorname{logit} \mathbf{p} &= X\boldsymbol{\beta} + \mathbf{b} \\ \mathbf{b} &\sim \mathrm{N}(0, \sigma^2 I) \end{aligned} \tag{3.7}$$

where $\mathbf{p}$ is a vector of probabilities and X is a design matrix indicating treatment or control. With i indexing body-systems, j indexing adverse events within a body-system, $k \in (0, 1)$ indicating treatment or control, and $x_0 = 0, x_1 = 1$, we can write this more simply as:

$$\begin{aligned} \operatorname{logit} p_{ijk} &= \alpha_j + x_k \beta_j + b_{ijk} \\ b_{ijk} &\sim \mathrm{N}(0, \sigma^2) \end{aligned} \tag{3.8}$$

where α_j is an overall body-system effect, β_j is a change in body-system effect due to treatment, and b_{ijk} are adverse event random effects. This is just one of a number of potential models for this data.

3.5 Longitudinal Analysis

Standard longitudinal data analysis methods, such as those described in [116], are also applicable to safety analysis. In addition to Poisson type regression for counts, other types of analyses are possible. Zeger and Diggle describe a semi-parametric model for analysing the counts of CD4 cell numbers in HIV Seroconverters [116], [117]. They model the mean response non-parametrically using a kernel estimator. Their model is:

$$Y_{ij} = \mu(t) + x_{ij}^T \boldsymbol{\beta} + W_i(t) + Z_{ij} \tag{3.9}$$

where Y_{ij} is the j^{th} measurement on individual i, x_{ij} is a vector of covariate values, $\boldsymbol{\beta}$ is a vector of regression parameters, and $\mu(t)$ is a smooth function of time. The W_i are independent replicates of a zero-mean stationary Gaussian process with covariance function $\gamma(u) = \sigma_w^2 \rho(u; \theta)$. The Z_{ij} are mutually independent measurement errors, each distributed normally as $N(0, \sigma_z^2)$. The data is an array of measurements: $\{y_{ij}(t_{ij})\}$. The Gaussian assumptions are not required for parameter estimation, only for inference. The model is also interesting in that the response variable Y is not a discrete variable even though we are dealing with cell counts. Many different covariance structures have been studied for this type of normal approach, but random effects models are also useful for modelling this, particularly for count data.

A Bayesian hierarchical approach to longitudinal analysis, also for CD4 cell counts, is described by Lange et al. [118]. They model the squareroot of the counts over an interval as a piecewise linear growth curve with random effects. The parameters are themselves random variables.

The model of Schildcrout et al. is another such model, although they look at a continuous response variable (liver enzyme activity with alanine aminotransferase) [119]. Their general model is:

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}) \tag{3.10}$$

where $Y_i(t_{ij})$ is the response of subject i at time t_{ij}, $\mathbf{X}_i(t_{ij})$ is a vector of covariates, $\boldsymbol{\beta}(t_{ij})$ is a time-dependent vector of parameters, and $\epsilon_i(t_{ij})$ is a mean zero error term. They base their inference on a mean model using natural splines, and take into account subject selection, dropout mechanisms, and treatment received.

3.6 Bayesian Approaches to Clinical Trials

In their 2010 book, *Bayesian Adaptive Methods for Clinical Trials* [67], Berry et al. describe some of the advantages and methods available for a Bayesian approach to the analysis of data from clinical trials. In particular, they stress the fact that Bayesian methods use all available evidence, that previous information may be encoded in the prior distributions or that uninformative or minimally informative priors may be used, that inferences depend only on the observed data and choice of prior, are flexible in the sense that they can be updated as more data accumulates, and they allow for prediction and are thus suitable for decision making. This flexibility can also be used to apply Bayesian methods over the various phases of clinical trials, and to meta-analyses of trials [120].

Before proceeding, we consider the following point about multiple comparison procedures, which papers introducing Bayesian approaches often make as a part justification for using Bayesian statistics in place of frequentist or classical approaches [109]:

> Why, from the scientific point of view, should the act of measuring a variable affect inferences about another variable? The issue here is that using a standard Multiple Comparison Procedure (e.g. Bonferroni) changes the significance levels (or adjusts p-values) used for hypothesis testing.

To quote Berry [109]:

> An investigator who carried out only one test might find a significant difference whereas the same difference would not have been significant had the investigator tested enough other variables.

The Bayesian view of this is that the results of the analysis performed seem to depend on the intentions of the investigator rather than on the data. Further, the Bayesian approaches used by Berry and Berry [5] and DuMouchel [6] claim to bring a number of advantages above frequentist methods:

- Multiple Comparison Robustness. The idea that posterior distributions tend towards the true distribution of the parameters provides multiple comparison robustness (shrinkage) [27], [66].
- Trials tend to be sized based on efficacy of treatment rather than safety concerns [4]. Typically safety events are rare leading to what DuMouchel calls the granularity problem [6]. There may not be a large amount of data available on which to base inferences. In the Bayesian approach assumed relationships between the parameters allow us to say more about the individual adverse event rates than we would otherwise (borrowing strength).

From a Bayesian perspective the multiple comparison problem can be considered as an issue of appropriately choosing a prior to account for dependency in multiple, related hypotheses [27], [66]. Scott and Berger also emphasise that multiplicities must be handled through choice of prior [121].

3.6.1 Bayesian Models for Adverse Event Incidence

In their 2004 paper Berry and Berry propose a hierarchical Bayesian model for the analysis of incidence data consisting of counts of occurrences of adverse events in control and treatment groups from clinical trials [5]. They use the 2004 paper of Mehrotra and Heyse as the basis for introducing their model, using the same nomenclature and example data [15].

Bayesian hierarchical models have many applications [122]. They are particularly applicable in areas where there is a natural hierarchical or population/subpopulation structure to the data being modelled [123], where the parameters involved can be regarded as connected or related in some way [122], or where the

data has high dimensionality. Hierarchical models are also appropriate where there is a level of uncertainty about the prior distributions and we wish to include this uncertainty in the Bayesian model [123]. In comparison, non-hierarchical models are usually inappropriate for hierarchical data [122], either they do not fit large data sets accurately, or they overfit the data, leading to inferior predictions.

The method of grouping adverse events by body-system has a natural hierarchical structure (Figure 3.1) around which Berry and Berry base their model. We can consider a body-system to be, in a sense, a dimension of the body, and the adverse events to be different aspects of it. There is no need to discard rare events, as Mehrotra and Heyse [15] and Mehrotra and Adewale [3] do in an attempt to reduce dimensionality, and in fact this would violate one of the main assumptions of Berry and Berry's approach as we will see.

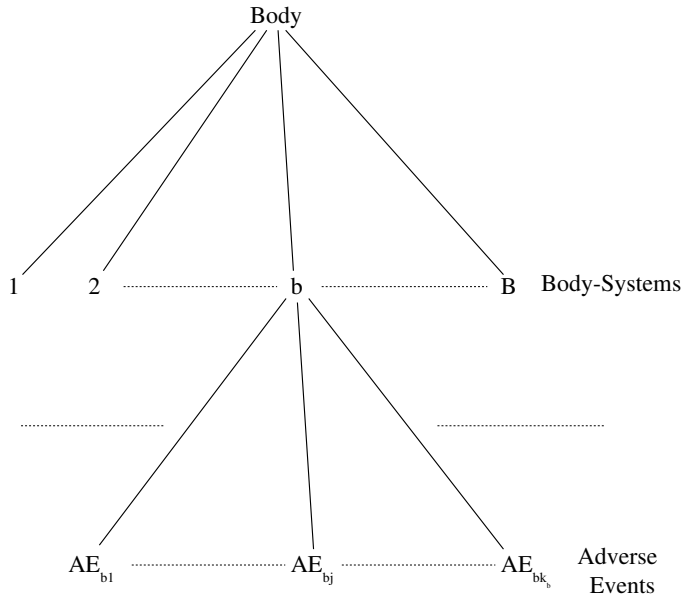


Figure 3.1. Adverse events grouped by body-system.

Bayesian hierarchical models also have a number of properties which make them computationally attractive. The assumption of conditional independence within the model [122], [123] allows the easy factorisation of the joint parameter distribution to derive the complete conditional distributions up to a constant, thereby facilitating the implementation of Markov Chain Monte Carlo (MCMC) fitting

algorithms for Gibbs sampling [124]. Quite often Bayesian hierarchical models use conjugate priors for hyperparameters on the basis that the analysis is robust, which further simplifies the computations required for simulation [123].

Berry and Berry suggest four considerations to be taken into account regarding flagging adverse events [5]:

1. The actual significance levels of the individual adverse events.
2. The total number of types of adverse events being considered.
3. The rates of adverse events not considered for flagging.
4. The biological relationship among the adverse events.

The first two are standard considerations in any frequentist analysis of the data. The latter two are not relevant to a frequentist approach, but are for a Bayesian approach. Removing rare events, as Mehrotra and Adewale [3] do in their method, would not be appropriate to the Bayesian approach that Berry and Berry wish to pursue.

Berry and Berry make the following assumptions regarding the data:

- It is important whether adverse events considered for flagging are in the same body-system. The assumption here is that rates of adverse events are more likely to be similar within a body-system than across body-systems.
- It is also important to take into account in some manner the rates of adverse events in the same body-system which may not be considered for flagging.

The approach is to assess whether a treatment causes an adverse event based on all available information.

DuMouchel introduces quite a similar approach which he calls Multivariate Bayesian Logistic Regression (MBLR) [6]. We will discuss this paper in more detail below (§3.6.3). Here again the emphasis is on adverse events with a clinical or other relationship, modelled by a hierarchical model looking to exploit these relationships to borrow strength.

3.6.2 Berry and Berry Model

The Berry and Berry approach to applying a Bayesian analysis to safety is a model consisting of [5]:

1. A Binomial distribution for the adverse event counts in the control and treatment groups.
2. A logistic model linking the Binomial probabilities to a three stage hierarchical model.

The model, which is described in more detail in below, explicitly includes the possibility of no differences between treatment and control by including a point-mass term.

3.6.2.1 Model Notation

We assume that there are B body-systems, within body-system b there are k_b types of adverse event labelled:

$$AE_{bj}, \quad j = 1, \ldots, k_b$$

There are N_C patients in the control group and N_T in the treatment group.

We let X_{bj} and Y_{bj} be the number of occurrences of AE_{bj} in the control and treatment groups respectively, with the probability of experiencing A_{bj} being c_{bj} for the control group, and t_{bj} for the treatment group.

3.6.2.2 Data Model

The data model is Binomial:

$$\begin{aligned} X_{bj} &\sim \text{Bin}(N_C, c_{bj}) \\ Y_{bj} &\sim \text{Bin}(N_T, t_{bj}) \end{aligned} \tag{3.11}$$

Letting

$$\begin{aligned} \text{logit}(c_{bj}) &= \log \frac{c_{bj}}{1 - c_{bj}} = \gamma_{bj} \\ \text{logit}(t_{bj}) &= \log \frac{t_{bj}}{1 - t_{bj}} = \gamma_{bj} + \theta_{bj} \end{aligned} \tag{3.12}$$

γ_{bj} is the log-odds in the control group, and $\theta_{bj} = \text{logit}(t_{bj}) - \gamma_{bj}$ is the relative increase in this log odds rate in the treatment group, i.e. $\theta_{bj} = \log\left[\frac{t_{bj}(1-c_{bj})}{c_{bj}(1-t_{bj})}\right]$ is the log odds-ratio.

3.6.2.3 First Level

The log-odds in the control group are modelled by normal distributions:

$$\gamma_{bj} \sim \text{N}(\mu_{\gamma b}, \sigma^2_{\gamma b}) \quad b = 1, \ldots, B \;\; j = 1, \ldots, k_b \tag{3.13}$$

From above, the θ_{bj} are log-odds ratios and if $\theta_{bj} = 0$ then the probability of a patient experiencing AE_{bj} is the same in both groups, i.e. $c_{bj} = t_{bj}$. The model assigns a positive probability, π_b, to this possibility by using a mixture of a normal distribution and a point-mass at zero in the prior distribution:

$$\theta_{bj} \sim \pi_b \, \text{I}_{[\theta_{bj}=0]} + (1-\pi_b) \, \text{I}_{[\theta_{bj}\neq 0]} \, \text{N}(\mu_{\theta b}, \sigma^2_{\theta b}) \quad b = 1, \ldots, B, \;\; j = 1, \ldots, k_b \tag{3.14}$$

where I is the indicator function.

3.6.2.4 Second Level

The standard Bayesian hierarchical approach is to assign a prior distribution to the hyperparameters which creates the second stage of the prior structure:

$$\begin{aligned} \mu_{\gamma b} &\sim \text{N}(\mu_{\gamma 0}, \tau^2_{\gamma 0}) \quad b = 1, \ldots, B \\ \sigma^2_{\gamma b} &\sim \text{IG}(\alpha_\gamma, \beta_\gamma) \end{aligned} \tag{3.15}$$

where IG is the inverse-gamma distribution (§A.6).

The probability π_b, that $\theta_{bj} = 0$, is assumed to be the same for all adverse events j in body-system b. The prior chosen for π_b is:

$$\pi_b \sim \text{Beta}(\alpha_\pi, \beta_\pi) \quad b = 1, \ldots, B \tag{3.16}$$

For the hyperparameters of the normal part of the mixture distribution we assume:

$$\begin{aligned} \mu_{\theta b} &\sim \text{N}(\mu_{\theta 0}, \tau^2_{\theta 0}) \quad b = 1, \ldots, B \\ \sigma^2_{\theta b} &\sim \text{IG}(\alpha_\theta, \beta_\theta) \end{aligned} \tag{3.17}$$

3.6.2.5 Third Level

In the third level of the hierarchical model the hyperparameters have distributions:

$$\begin{aligned} \mu_{\gamma 0} &\sim \mathrm{N}(\mu_{\gamma 00}, \tau^2_{\gamma 00}) \\ \tau^2_{\gamma 0} &\sim \mathrm{IG}(\alpha_{\gamma 00}, \beta_{\gamma 00}) \end{aligned} \tag{3.18}$$

The prior distributions for α_π, β_π are truncated exponentials:

$$\begin{aligned} \alpha_\pi &\sim \mathrm{M}(\lambda_\alpha)\, \mathrm{I}(\alpha_\pi > 1) \\ \beta_\pi &\sim \mathrm{M}(\lambda_\beta)\, \mathrm{I}(\beta_\pi > 1) \end{aligned} \tag{3.19}$$

Berry and Berry take $\lambda_\alpha = \lambda_\beta$, this means that 0.5 is the a priori probability that $\theta_{bj} = 0$.

Finally the hyperparameters of the normal prior of $\mu_{\theta b}$ have the distributions:

$$\begin{aligned} \mu_{\theta 0} &\sim \mathrm{N}(\mu_{\theta 00}, \tau^2_{\theta 00}) \\ \tau^2_{\theta 0} &\sim \mathrm{IG}(\alpha_{\theta 00}, \beta_{\theta 00}) \end{aligned} \tag{3.20}$$

The parameters $\lambda_\alpha, \lambda_\beta, \mu_{\gamma 00}, \tau^2_{\gamma 00}, \mu_{\theta 00}, \tau^2_{\theta 00}, \alpha_\gamma, \beta_\gamma, \alpha_\theta, \beta_\theta, \alpha_{\gamma 00}, \beta_{\gamma 00}, \alpha_{\theta 00}, \beta_{\theta 00}$ are assumed to be fixed constants and Berry and Berry give these the values:

$$\begin{aligned} &\mu_{\gamma 00} = 0, \tau^2_{\gamma 00} = 10, \alpha_\gamma = 3, \beta_\gamma = 1, \alpha_{\gamma 00} = 3, \beta_{\gamma 00} = 1, \lambda_\alpha = 1 \\ &\mu_{\theta 00} = 0, \tau^2_{\theta 00} = 10, \alpha_\theta = 3, \beta_\theta = 1, \alpha_{\theta 00} = 3, \beta_{\theta 00} = 1, \lambda_\beta = 1 \end{aligned} \tag{3.21}$$

The importance of fixed values in a hierarchical model is reduced by the model's robustness and the above values were used in this study [123].

In the original implementation of the model the mean parameters $\mu_{\gamma b}$ had a shared variance σ_γ [5]. In the implementation used in this study we follow [60] and [125] and replace σ^2_γ by a body-system variance $\sigma^2_{\gamma b}$.

3.6.2.6 Modelling the Body-System

The directed acyclic graph (DAG) for the hierarchical model is shown in Figure 3.2.

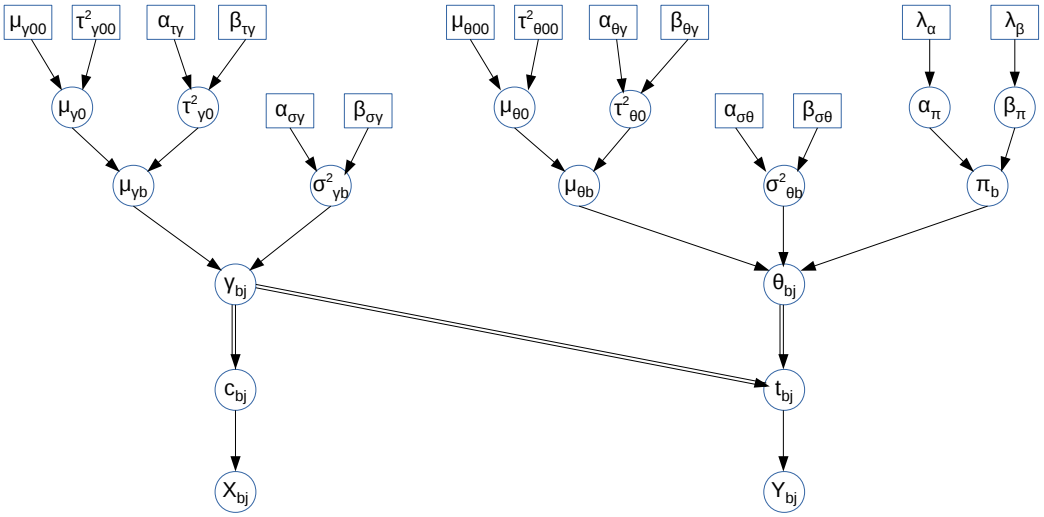


Figure 3.2. Directed acyclic graph for the Berry and Berry Model.

Graphical representations for model DAGs are described in [126, Ch. 2].

We can see the proposed hierarchical Bayesian model is in effect a reflection of the body-system described in Figure 3.1, with the body-system means, $\mu_{\gamma b}, \mu_{\theta b}$, sampled from body means, $\mu_{\gamma 0}, \mu_{\theta 0}$. Uncertainty about the body-means is reflected in the fact that $\mu_{\gamma 0}, \mu_{\theta 0}$ themselves have prior distributions. A key part of the model is the use of a mixture distribution for the θ_{bj}. This explicitly models the possibility that the probabilities of the occurrences of adverse events in both groups is the same. A number of similar models [6], [60] either ignore the point-mass or use mixtures of normal distributions with a small variance as a point-mass equivalent. Implementation of these models is more straightforward. The use of the point-mass is discussed in some detail in [127].

In the simulation study in Chapter 5 we will look at the Berry and Berry model both with and without the point-mass. We expect the presence of the point-mass to reduce the Type-I error rate compared to similar models without the point-mass. A strong safety signal may be required to overcome the effect of including this term in the model. This may result in an increase of the Type-II error rate when differences between the treatment and control are small. For models without the point-mass, we may expect more Type-I errors, particularly if some of the adverse

events have high treatment rates compared to the control, in effect pulling up the overall rates.

3.6.2.7 Discussion

Berry and Berry apply the model to the data from [15, Table 1] and give a brief overview of the results, which are in agreement with those of Mehrotra and Heyse [15]. In addition, they make a small investigation of the effect of moving a "flagged" adverse event to a different body-system. They report that this materially affects the results, indicating that the assignment of adverse events to body-systems is an important consideration.

It is not really possible to say much about the results from the model as applied to the data in [15], the true treatment and control differences for individual adverse events are not known. To investigate further Berry and Berry fitted the model to simulated data with the intention of exploring how well it fits. The main conclusions from the paper are that:

- The inclusion of the point-mass in the model can be considered important in that it is reasonable that some adverse events will not be affected by the treatment.
- The model is sensitive to body-system. Moving an adverse event from one body-system to another can have a (dramatic) effect on the findings.

Berry and Berry further state that it should be possible to model dependencies among adverse events at the patient level, if the data were available, allowing more precise conclusions about treatment effects.

While Berry and Berry concentrate on Tier 2 adverse events, they make no real modelling distinction about rare or serious events. Compared to the Mehrotra and Adewale [3], who consider Tier 3 events to be those rare serious events that require separate clinical evaluation, the Berry and Berry approach has the advantage that this type of event is always included in their model. The model does not require any monitoring in this sense. There is no justification within the paper for the choice of priors or the parameter values given, although a number of citing papers consider the prior choices suitable or discuss alternatives [127].

The Berry and Berry model has limitations. It is useful for marginal incidence data only, it does not take into account timings of events, it does not take into

account the severity of events, no account is taken of patient data or censoring, and no sub-group analysis is possible.

The Berry and Berry model takes adverse events in the same body-system as exchangeable and the assignment to body-system as given. These assignments are critical for making any clinical decisions. However, the paper deals mostly with the suitability and fitting of the model, which is just one part of the procedure. The model must be used to make a decision or at least flag a potential safety signal. Unlike the error controlling procedures discussed in Chapter 2, Berry and Berry do not put forward any clear decision rule. In essence they just examine the posterior probabilities of an increase in incidence and any decisions are based on this. In particular, it is not clear how to determine error rates and power. Simulated studies, with known error rates, could be used to determine thresholds for the posterior distributions to achieve an estimated prespecified power. In a sense Type-I and Type-II error rates become our decision criteria. Xia et al. take this approach [60]. In contrast, it is worth noting that Bayesian decision theoretic approaches to controlling the False Discovery Rate exist [54], [55].

The Berry and Berry model can be considered to be over-parameterised in the sense that that there could possibly be more parameters than data points. In the model there are essentially 4/5 parameters for each body-system. However, over-parameterised hierarchical models have a number of advantages in that they may improve mixing and convergence [126], and they may have enough parameters to both fit the data and to model some dependence between the parameters [122, Ch. 5], in a sense avoiding overfitting. Indeed Gelman et al. state that it is often sensible to fit hierarchical models with more parameters than data-points [122].

3.6.3 Multivariate Bayesian Logistic Regression

DuMouchel [6] introduces Multivariate Bayesian Logistic Regression (MLBR) which, like Berry and Berry [5], is a Bayesian Hierarchical model for the analysis of adverse event incidence data. However, DuMouchel's model includes covariates and treatment-covariate interactions, and allows for subgroup analysis. The data he considers are sparse response data, such as adverse events from clinical trials. In his paper he analyses 10 medically related issues from a pool of 8 studies, so he is performing a meta-analysis. One advantage of working with pooled data is that the number of events is increased.

The predictor variables in the model are assumed to be categorical (or dichoto-

mous) and observable at the time of subject randomisation. The analysis is cross-sectional rather than longitudinal, the timings of the events or observation times are not considered in the analysis. The response data is assumed to arise from individuals enrolled in the studies, the subjects, and for each subject the response is 0 or 1 for each safety issue, depending on whether a subject has experienced that particular safety issue or not, thus giving overall counts of the incidences of various safety events when taken in summary form.

As may be expected, the primary explanatory or predictor variable is Treatment vs. Control, which DuMouchel calls study Arm with values "Treatment" or "Comparator". Other subject level covariates may be included, e.g. gender, age categories, medical history variables, as well as treatment-covariate interactions and, if data is pooled from multiple studies, a study identifier covariate.

DuMouchel gives two main reasons for his approach, which we have previously mentioned:

1. Shrinkage can provide multiple comparison robustness.
2. MBLR fits the same model to each response variable which allows borrowing of strength.

The assumptions here are therefore that the different adverse events should be medically related in some way, such as being part of the same body-system, or all based on the same underlying process. As noted in §1.9, classifying or grouping adverse drug reaction reports may be controversial, and different groupings may change the statistical significance of the adverse event count data [28]. Berry and Berry illustrated this issue in their paper [5].

DuMouchel considers a Bayesian approach as a good compromise between attempting an analysis of safety events that are so rare that reliable comparisons are not possible, and using a single analysis for all adverse events which could potentially submerge a few important safety signals. The model is positioned as an additional support for safety analysis in an exploratory or confirmatory sense when there are many parameters of interest. DuMouchel claims that in the model the Bayesian estimates of the treatment-covariate interactions are conservative, in the sense that estimates are "shrunk" toward null hypothesis values, in order to reduce the false alarm rate due to high variance in small sample sizes. This conservativeness can be considered a form of adjustment for multiple comparisons.

In discussing comparisons of MBLR with other methods, DuMouchel states that a standard logistic regression approach (without treatment-covariate interactions) to analysing the data can fail because the likelihood function has no unique finite maximising set of parameters [128]. Instead, a favourable comparison is made with what he terms Regularized Linear Regression (RLR), a "weak Bayes" method which corresponds to setting certain variance components (that are estimated by MBLR) to values selected to be so large that resulting estimates would be virtually the same as those for standard logistic regression, if the data were not so sparse.

3.6.3.1 The Model

We assume that there are S subjects in total and K types of adverse event. All covariates are assumed to be categorical, subjects are grouped by covariate pattern into m separate groups, with n_i subjects in group i, so that $\sum_{i=1}^{m} n_i = S$. We assume that there are J predictor variables (all categorical), excluding treatment, and that the j^{th} predictor has g_j categories, let $G = \sum_{j=1}^{J} g_j$ be the number of subgroups within these categories. Within group i, N_{ik} subjects experience issue k and T_i is the treatment indicator. For the i^{th} group of subjects the probability of experiencing issue k is:

$$P_{ik} = \frac{1}{1+\exp(-Z_{ik})} \quad (\text{ or alternatively } \operatorname{logit}(P_{ik}) = Z_{ik}) \tag{3.22}$$

where

$$Z_{ik} = \alpha_{0k} + \sum_{1\le g\le G} X_{ig}\alpha_{gk} + T_i\left(\beta_{0k} + \sum_{1\le g\le G} X_{ig}\beta_{gk}\right) \tag{3.23}$$

with the constraints that $\sum_{g=1}^{G} \alpha_{gk} = \sum_{g=1}^{G} \beta_{gk} = 0$, and X is the design matrix of dummy variables for the J covariates (with G total categories over all covariates). The priors for the model are a 2-stage hierarchy:

$$\begin{aligned}
&\alpha_{gk}|A_g \sim N(A_g, \sigma_A^2) && k = 1,\ldots,K; g = 1,\ldots,G\\
&\beta_{0k}|B_0 \sim N(B_0, \sigma_0^2) && k = 1,\ldots,K\\
&\beta_{gk}|B_g \sim N(B_g, \sigma_B^2) && k = 1,\ldots,K; g = 1,\ldots,G\\
&B_g \sim N(0, \tau^2) && g = 1,\ldots,G
\end{aligned} \tag{3.24}$$

The priors for $\{\alpha_{0k}, A_g, B_0\}$ are assumed uniform within $(-\infty, \infty)$. The variances $(\sigma_A^2, \sigma_0^2, \sigma_B^2, \tau^2)$ are assumed to have prior distributions uniform in the 4-dimensional cube $0 \le \sigma_A, \sigma_B, \sigma_0, \tau \le d$. The model is structured in such a way

that:

1. The assumptions that coefficients for the same predictor across multiple issues cluster around predictor specific values $(A_1, \ldots, A_G, B_0, \ldots, B_g)$ is implemented in the hyper-priors for $\alpha_{gk}, \beta_{0k}, \beta_{gk}$. This clustering is dependent on the variances $(\sigma_A^2, \sigma_0^2, \sigma_B^2)$, smaller values giving tighter clustering and larger values leading to no discernible patterns among the coefficients.

2. The hyper-prior for B_g, where $B_g = 0$ means no treatment by covariate interactions (when averaged across responses), ensure that the "null hypothesis" $B_g = 0$ is given priority in the analysis. This is similar to Berry and Berry's use of a point-mass to model the possibility of no differences between control and treatment [5]. The value of τ^2 determines how strongly to shrink the prior means towards 0.

For large G there will be many possible subgroup comparisons and caution is needed in interpreting estimates which are unadjusted for multiple comparisons (i.e. any frequentist fit of the data). The MBLR estimates are designed to be more reliable in the presence of these multiple comparisons because of subgroup-by-treatment interaction shrinkage (towards 0), and the possibility of borrowing strength provided there is an observed similar pattern of treatment and subgroup effects in most of the K issues being analysed. When configuring a MBLR approach issues should be selected for which there is some suspicion of a common medical mechanism involved. This requires that we just analyse one set of adverse events, for example from a single body-system or organ grouping, in contrast to Berry and Berry who include all adverse events in their model [5].

3.6.3.2 Interpretation of Coefficients

The α_{gk}, with the constraint that $\sum_g \alpha_{gk} = 0$, define the risk of issue k for the comparator subjects (i.e. the controls), in the sense that $\alpha_{0k} + \alpha_{gk}$ is the log-odds that a subject in subgroup g will experience issue k, averaged across the categories of other predictors not defined by subgroup g. For treatment effects, $\beta_{0k} + \beta_{gk}$ are the estimated log-odds ratios for the risk of issue k (Treatment vs. Comparator), that a subject in subgroup g will experience issue k, averaged across the categories of other predictors not defined by subgroup g. There is a similar constraint to the α's: $\sum_g \beta_{gk} = 0$.

3.6.3.3 Model Fitting

DuMouchel describes in some detail the process for fitting the model in Section 3 of the paper. In contrast to Berry and Berry ([5]), DuMouchel does not use a direct MCMC simulation approach to approximate the posterior distributions. Instead a discrete approximation approach is taken. The standard deviations $(\sigma_A, \sigma_0, \sigma_B, \tau)$ are assumed fixed and known and the other parameters are estimated, conditional on these values, using Newton-Raphson maximisation of the log of the joint posterior distribution of the parameters. DuMouchel's rational is that he feels inexperienced scientists would have difficulty assessing convergence of high-dimensional MCMC runs, and that certain users might be uncomfortable with the fact that a repeat analysis on the same data would typically lead to slightly different answers (repeatability concerns) for parameter values, even if the MCMC runs had converged.

3.6.3.4 Discussion

DuMouchel briefly discusses his model in comparison to the Berry and Berry model [5]. While Berry and Berry do not include covariates, and concentrate on treatment/control odds-ratios only, they have a more complex model with many more variance components. The inclusion of covariates and treatment-by-covariate interactions in MBLR allows the possibility of detecting vulnerable subgroups that have different responses to treatment. DuMouchel considers that without the "smoothing effect" of Bayesian shrinkage of the interaction terms in the model, estimates of interactions affecting rare events would be so variable as to be useless.

DuMouchel defends his choice of normal priors from the criticism that, since they generate few outliers, using these priors may in fact suppress them. Alternative priors, such as t-distributions or double exponential ("lasso") distributions, could be considered as these tend to shrink outliers less, but the double exponential distribution has properties which make computation difficult, and so DuMouchel does not consider it for his model. Similarly, t-distributions are difficult to handle computationally in a complex model like MBLR. In the proposed implementation MBLR is guaranteed to converge as it has a log-concave posterior density function.

DuMouchel applies MBLR to the analysis of 10 adverse events from the pooled data of 8 studies to investigate:

1. the commonality of the safety issues;

2. the possibility that certain subgroups of subjects may be more (or less) affected;

and uses a simulation study to investigate the statistical properties of MBLR.

He concludes that while safety issues with low frequencies will produce standard logistic regression estimates with wide confidence intervals, MLBR can be seen as a compromise between the analysis of finely distinguished events and a single analysis of a pooled event. It requires a selection of medically related issues, potentially exchangeable with respect to their dependence on treatment and covariates. In fact the key concept is that the set of K issues have been pre-specified as important and likely to be biologically and clinically related.

Although the main data analysis discussed in the paper is a meta-analysis, there is no particular reason why the approach could not be used within a single trial. While not directly using body-systems, and in fact it is not possible to include multiple body-systems in MBLR, DuMouchel ([6, §4]) stresses that the

> selection of which issues to include in an MBLR is important. There needs to be at least a superficial plausibility that all or many of the selected outcome issues might have similar odds ratios with treatment and with the covariates in the model, what Bayesians call exchangeability.

While DuMouchel states that his paper is inspired by, and similar in spirit to, the Berry and Berry model [5], there are a number of differences. He includes grouping by covariate, with the possibility of borrowing strength if there is an observed similar pattern of treatment and subgroup effects in most of the K issues being analysed. This is a separate layer of possible meaningful groupings (by covariate values). However, his model is a restriction of [5] in the sense that his requirement that all adverse events be medically related can be considered to restrict the model in effect to a single body-system, whereas [5] include all body-systems in their model. This is why MBLR does not have a third level to the hierarchical model. We can consider MBLR to be one body-system with covariates, and the Berry and Berry model [5] to be multi-body-systems with no covariates. Further, it would be possible to use a similar approach to $B_g \sim N(0, \tau^2)$ with small τ^2 in the Berry and Berry model, rather than the zero point-mass probability π_b (§3.6.2.6).

In common with the Berry and Berry model, the DuMouchel model has the following limitations: it is useful for marginal (incidence) data only, it does not take into account timings of events, it does not take into account the severity of events, no account is taken of patient data or censoring. However, unlike [5], in MBLR subgroup analyses are possible. Due to their respective similarities any approach to taking clinical decisions based on the model would essentially be the same as those for the Berry and Berry model (§3.6.2.7).

3.6.4 Further Discussion

Both the Berry and Berry model [5] and MBLR [6] have been discussed a number of times in the literature, with both being considered interesting approaches to safety, but also the subject of some criticism, with a number of extensions or alternative models proposed for similar data.

Evans [68], Berry [69], McEvoy et al. [70] and Shaddox et al. [129] all consider MBLR as an important step in the development of statistical methods for the analysis of safety data, particularly with regard to the borrowing of strength between related adverse events. However McEvoy et al. are critical of MBLR as a meta-analysis tool because the model does not preserve the trial specific randomised comparison between the treatment and control groups, and consider that the model could be improved by taking this into account. They present a modified MBLR formulation with trial specific terms, and performed a comparisons between the two models using a fully Bayesian approach (using `OpenBUGS` [126]), rather than the approximation approach used in [6].

The Berry and Berry model has also been criticised in the literature. Crooks et al. [130] consider that, while Berry and Berry have shown that including a Bayesian hierarchy can be important in analysing adverse events, the model is unadjusted and overly simplistic. The Crooks et al. approach is to use a Bayesian model of information sharing as opposed to groupings. They propose a Bayesian hierarchical model which is integrated with methods to allow for confounding and interactions. Their information sharing model was constructed by a practitioner (a gastroenterologist) who defined a 3-level hierarchy. Potential confounding and interactions were assessed by a non-Bayesian analysis, and separate logistic regression models were constructed for the various combinations. The results from this multivariate analysis were included in the Bayesian hierarchy and re-estimated by MCMC methods. It is difficult to assess this approach as the authors do not give the model

details in their paper.

Chen et al. [125] extend the Berry and Berry model [5] to a group sequential method with the aim of early detection of adverse events that might be associated with treatment, while controlling Type-I and Type-II error rates. The approach is a standard Bayesian update of the posterior distributions based on the data available at the interim analyses. They use a decision-theoretic approach to both minimise the posterior expected loss due to the misclassification of an adverse event, and to determine the threshold values for a group sequential signalling process for the safety data. There is no discussion of the possibility that the rate of occurrences of adverse events may not be constant over the lifetime of the trial, or if this is important. However it is an approach that is suitable for interim analyses.

Other similar modelling approaches have also been proposed. Agresti and Klingenberg [131], who briefly discuss [5], use a test statistic which considers the count data in a multivariate manner and applies tests which are the analogue of the Hotelling T^2-test for vectors of binary responses. A not dissimilar multivariate approach is discussed by Chuang-Stein, Mohberg and Musselman [29]. They use clinical and patient information to group safety data into classes by body-system, and a score or grade is assigned at a patient level to the levels of "acceptability" within each class. A multivariate test is performed to determine if the safety profile for treatment and control is the same. Xia et al. offer a `WinBUGS` ([126]) implementation of the Berry and Berry model and discuss five very similar models [60]. Models 1a and 1b in the paper are based on [5], Models 2a, 2b are very similar to [5], but use a Poisson model to take into account the total subject exposure time and incidence rates for the different adverse events. Decisions regarding flagging adverse events are made using cut-off points determined by simulation studies. The fifth model they discuss is a non-hierarchical approach to end of trial count data with fixed parameters.

Goldberg-Alberts and Page take an alternative view on adverse events, grouping them by "constellation" and using a log-linear model to estimate the magnitude of association between them [132]. They do not address the issue of multiplicities. Gould proposes an alternative Bayesian method [133]. He regards the incidences of adverse events as realisations from a mixture of distributions, and looks to find the element of the mixture which corresponds to each adverse event. The actual adverse event counts are considered to be Binomial distributions, with the control probabilities having Beta priors, and the treatment probabilities having mixture

distributions. In this way it is similar to the top level of the Berry and Berry model [5], although it does not use a point-mass or directly take advantage of a body-system relationship. The model is applied to the data in [15] and compared to the Berry and Berry results. Kim et al. propose a number of (non-Bayesian) regression models which model adverse events grouped by body-system over time [134]. Their method is called stratified quasi-least squares (SQLS) and is an extension of the General Estimating Equations (GEE) approach for correlated data. They assume a within-patient correlation between the adverse events, giving a general form for the covariance matrix. The general theory is developed before a simulation study is analysed. The data for the study is generated from Binomial and Poisson variables, and a number of log-linear models are fitted.

Rosenkranz uses an empirical Bayes approach for adverse event data [135]. This is a Poisson model with a subject dependent rate and log-normal priors. The rate is assumed constant leading to a summary Poisson model at the SOC (System Organ Class) level. Schildcrout et al. discuss a general time dependent linear model for longitudinal data which can be used in post-marketing surveillance or for meta-analyses [119]. They do not address the issue of multiplicities and concentrate on a single adverse event only, although this could also be used for multiple occurrences (§3.5). Simo ([136]) presents a Poisson regression multilevel model with random effects, based on a model of Christiansen and Morris [137], and uses it to analyse the data in [15]. The Christiansen and Morris model itself uses "overall" Poisson rates to model the sum of individual Bernoulli trials where the probabilities are small and the data is modelled by Negative Binomial distributions. Southworth and O'Connell [138] discuss a number of approaches to analysing clinical adverse event data, including the Berry and Berry model [5], as alternatives to hypothesis testing. They stress data-mining or explorative approaches and a final analysis based on graphs and summaries which are easy to interpret. The first approach discussed is a logistic regression with a penalised likelihood (including the assumption that the adverse events are independent), the second is what is termed an "inside-out" approach where the adverse events are used to classify the subjects to treatments (a form of machine learning), the third approach discussed is the Berry and Berry model. Their recommendation is that all discussed methods are useful and should be applied where possible. The paper does not contain a full description of the data or the models used apart from the Berry and Berry model [5]. In a 2009 presentation Prieto-Merino et al. ([127]) discuss Bayesian Hierarchical Modelling, with reference to [5], for clinical adverse event data. No new models are introduced,

rather they discuss choosing the prior distributions for use in the model and, in particular, the effect of the point-mass and possible replacements for it, giving as an example a model that is the same as Model 1a from [60]. They also discuss the issue of assigning adverse events to groups.

A number of similar methods have been applied to the area of pharmacovigilance. DuMouchel describes a method using empirical Bayes modelling for data mining Spontaneous Report Adverse Event Databases [9]. Gould ([8]) describes a similar Bayesian method and compares it to both the empirical Bayes approach, and the positive FDR [44]. Bate et al. use a Bayesian neural network method for a similar purpose [10]. Recent work in the area of data mining Longitudinal Observational Databases (LODs) for pharmacovigilance by Shaddox et al. uses the Bayesian Self-Controlled Case Series regression model (BSCCS) [129]. This approach uses a Poisson regression to model multiple outcomes (adverse drug events) for multiple drugs over what are termed drug eras.[5] The data is at the patient level and relationships between all outcomes for any particular single drug are modelled using a hierarchy. In this approach outcomes that are considered related are grouped together, but in this case not by using a body-system. Another property of these pharmacovigilance studies is the absence of control data, the LODs record adverse event occurrence only so no baseline risk estimates are possible.

Not all Bayesian approaches directly model the data as in the Berry and Berry model or MBLR. As mentioned in §2.3.2.4 Muller et al. investigate using a Bayesian decision-theoretic approach to control the FDR [54] and León-Novelo et al. take a Bayesian approach to the False Discovery Proportion (FDP), whose posterior mean is the False Discovery Rate [55]. They consider the decision theoretic problem of controlling the FDR in this Bayesian context, and an alternative decision approach based on a utility function. The data considered are modelled by a Poisson Bayesian hierarchy over three distinct stages (time periods). The utility function approach weights on the size of changes over the stages. They found that strictly using statistical significance for flagging interesting events may in fact be inappropriate and that a criterion which is closer to biological significance may be required.

[5] A Drug Era is defined as a span of time when the Person is assumed to be exposed to a particular drug. Source: Observational Medical Outcomes Partnership (OMOP) (`http://omop.org/`) Common Data Model (CDM) Version 4.0.

3.7 Discussion

Many of the methods discussed above are designed to handle incidence data only, but some include timings in their approach, and also per-patient data. In the case of patient data, generally the individual patients' data is considered to be independent, possibly given some covariate information. However, some of the models which include timings have assumed constant incidence rates and do not take into account the possibility of changing incidence over time. We are looking for relatively simple summary models, easily interpretable, which take into account both the cumulative presence of adverse events and their trajectories, while controlling for multiple comparisons, and none of the methods above provide all these elements.

3.7.1 Bayesian Models

In the Bayesian approaches discussed the incidence data models tend to deal with marginal count data only, typically no patient level data is taken into account. The event timing models tend to either semi-parametric or Poisson based models for counts, and generally rates are assumed constant. Often the duration and severity of the events is not taken into account, nor is the possibility of a single patient being responsible for many adverse events, although some of the per-patient Bayesian models are able to cater for this. The models often ignore censoring. Control for multiple comparisons is automatic, via the choice of the prior, but deciding which adverse events to flag may require a decision theoretic approach. However, Bayesian models are a suitable choice for interim analyses as updating model parameters is part of the Bayesian framework.

3.7.2 Frequentist Models

Frequentist approaches include incidence data models often with patient timings, event duration models, survival models, recurrent event models, and general longitudinal models for counts or rates. They can be parametric, semi-parametric, or non-parametric. Parametric and some semi-parametric models often include assumptions of an underlying Poisson model. The severity of the events is generally not taken into account. The possibility of a single patient being responsible for many adverse events could be handled by a frailty model. Models exist which take into account both random and dependent censoring. There is no straightforward way to control for multiple comparisons although mixed model methods may be

one such approach. Unlike Bayesian models there is no obvious framework for handling interim analyses. These models are usually fitted by maximum likelihood estimation or some variation on it.

3.7.3 Further Extensions of the Above

The use of the non-parametric MCF, compared to the crude incidence rate or event-adjusted incidence rate in §3.3.1.1, shows the value of including event timings where possible. The grouped methods of [3] and [5] also bring advantages when controlling for multiple comparisons. Another advantage of the Bayesian approaches is ability to update parameters at interim time points, such as clinical trial interim analyses. The approach we will look to take will combine adverse event timings and groupings by body-system in easily an interpretable model,suitable for use at interim analyses.

Chapter 4

Adverse Event Detection in GSK EGF100151

4.1 Introduction

In this chapter we apply a number of the methods discussed in Chapters 2 and 3, which are suitable for analysing adverse event incidence data at the end of a trial, to the safety data from GSK EGF100151 (§1.8). The data is taken from Table 8.1[1] in the end of trial study report (Table 1.3).

We are interested primarily in grouped methods for a number of reasons. There may be a relationship between the adverse events, and using groupings which reflect this in a statistical analysis may provide more power for certain error controlling procedures. For Bayesian models, a hierarchical choice of priors may reflect a structure in the data, and provide multiple comparison robustness. Grouped data is also a common way of presenting safety data in clinical trial reports. From the unadjusted test data in Table 1.12, we can see that the adverse events with the smallest Fisher exact test p-values come from a small number of body-systems, and this does appear to indicate that there may be a body-system effect within the adverse events, or that this possibility is worth investigating.

We apply a number of the grouped error controlling procedures to the safety data (DFDR [3], GBH [30]), and also the Bayesian approach of Berry and Berry [5], rather than the more limited approach of DuMouchel [6]. These methods will also form the basis of a simulation study in Chapter 5. The models and method definitions are given in Chapters 2 and 3 and listed in Table 4.1. Some of the results we see below will give pointers to the types of behaviour we may expect to see in the simulation study.

[1] **Table 8.1.** Summary of All Adverse Events (AE attributed to Randomization Phase).

All the methods are implemented in the `c212` package for `R` (Appendix A, [139]). `BUGs` implementations of the Bayesian models were run using `OpenBUGS` ([126]) for comparative purposes (§A.8). For the Bayesian models all results are presented under the assumption that model fitting has reached (approximate) convergence. Model convergence and parameter tuning is discussed in Appendix C.

Method Name	**Description**
c212.BB	Berry and Berry model [5] (model 1b from [60])
c212.1a	Berry and Berry model without point-mass [5] (model 1a from [60])
DFDR	Double false discovery rate [3]
GBH	Group Benjamini-Hochberg [30]

Table 4.1. Trial EGF100151: Methods applied to safety data.

For the Bayesian models the posterior probability that the log-odds ratio of the incidence of adverse events on the treatment arm compared to the control arm, θ, is greater than 0 is used to determine which adverse events may be of interest.

4.2 End of Trial Safety Data

The final clinical report, 2010N107773_00 (Table 1.3), contains the safety data for the trial. Figure 4.1 shows the total incidence for six of the adverse events in the *Gastrointestinal disorders* body-system. Here, `Group` indicates either the control group (1) or treatment group (2), `Count` is the total incidences of the adverse events through the trial on each trial arm, and `Total` is the total number of participants on each trial arm. For example, we can see that 78 out of a total of 191 participants in the control group experienced the adverse event *Diarrhea*, compared to 145 out of 210 in the treatment group.

Body-system/SOC	Adverse Event	Group	Count	Total
Gastrointestinal disorders	Diarrhea	1	78	191
Gastrointestinal disorders	Diarrhea	2	145	210
Gastrointestinal disorders	Nausea	1	85	191
Gastrointestinal disorders	Nausea	2	98	210
Gastrointestinal disorders	Vomiting	1	43	191
Gastrointestinal disorders	Vomiting	2	63	210
Gastrointestinal disorders	Abdominal pain	1	30	191
Gastrointestinal disorders	Abdominal pain	2	31	210
Gastrointestinal disorders	Stomatitis	1	23	191
Gastrointestinal disorders	Stomatitis	2	37	210
Gastrointestinal disorders	Constipation	1	24	191
Gastrointestinal disorders	Constipation	2	24	210

Figure 4.1. GSK EGF100151 - End of Trial Data.

4.3 Berry and Berry Model

In this section we apply the Berry and Berry model with and without the point-mass to the final safety data from EGF100151 using the `c212` and `OpenBUGS` software.

4.3.1 Model with Point-Mass (c212.BB)

The Berry and Berry model results for the top 10 adverse events by posterior probability are presented in Tables 4.2 and 4.3.

We can see that there is very good agreement between the two implementations. The only difference in content is the tenth adverse event which is *Dry skin* in the `R` implementation and *Back pain* in the `OpenBUGS` implementation. In fact *Back pain* has posterior probability 0.879 in the `R` implementation, and is the eleventh highest posterior probability, while for the `OpenBUGS` implementation *Dry skin* has posterior probability 0.875, and is also the eleventh highest posterior probability. The models are fit by random sampling so this type of variation is not unexpected.

Comparing the top 10 events in Table 4.2 to the ten events significant at the 5% level for Fisher exact tests in Table 1.12, we can see that the top 6 adverse events in each are the same although with slightly different ordering, and overall 8 out of the 10 adverse events with significant Fisher exact tests appear in Table 4.2. The two missing are *Localised infection* and *Back pain*, with posterior probabilities of 0.772 and 0.879 respectively. Although, as noted above, *Back pain* appears in place

System Organ Class	Adverse Event	Posterior[1] probability $\theta > 0$
Gastrointestinal disorders	Diarrhea	1.000
Skin and subcutaneous tissue disorders	Rash	1.000
Gastrointestinal disorders	Dyspepsia	0.986
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.980
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.967
Skin and subcutaneous tissue disorders	Nail disorder	0.941
Respiratory, thoracic and mediastinal disorders	Dyspnoea	0.910
Musculoskeletal and connective tissue disorders	Arthralgia	0.905
Musculoskeletal and connective tissue disorders	Muscle spasms	0.892
Skin and subcutaneous tissue disorders	Dry skin	0.890

Table 4.2. Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry point-mass model (c212.BB), `R` implementation.

[1] θ is the log-odds ratio for the adverse event (§3.6.2.2).

of *Dry skin* in the `OpenBUGS` implementation. We can see the body-system effect here with *Localised infection* being the only significant adverse event in the system organ class *Infections and infestations* in Table 1.12, but not having a correspondingly high posterior θ probability. So, overall, the model does reflect the Fisher exact test output, while taking into account the assumed relationships between the system organ classes or body-systems. *Diarrhea* and *Rash* remain the adverse events most associated with *lapatinib and capecitabine*, but other possible adverse events in *Respiratory, thoracic and mediastinal disorders* and *Musculoskeletal and connective tissue disorders* are also indicated. Using a cut-off point of 90% posterior probability we would have flagged 8 events using the Berry and Berry model. For a cut-off of 95% this drops to 5 flagged events.

System Organ Class	Adverse Event	Posterior probability $\theta > 0$
Gastrointestinal disorders	Diarrhea	1.000
Skin and subcutaneous tissue disorders	Rash	1.000
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.985
Gastrointestinal disorders	Dyspepsia	0.984
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.965
Skin and subcutaneous tissue disorders	Nail disorder	0.936
Respiratory, thoracic and mediastinal disorders	Dyspnoea	0.903
Musculoskeletal and connective tissue disorders	Arthralgia	0.903
Musculoskeletal and connective tissue disorders	Muscle spasms	0.889
Musculoskeletal and connective tissue disorders	Back pain	0.880

Table 4.3. Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry point-mass model (c212.BB), `OpenBUGS` implementation.

4.3.2 Model without Point-Mass (c212.1a)

Tables 4.4 and 4.5 give the top 10 adverse events by posterior probability for an analysis using the Berry and Berry model without the point-mass term.

The output from both implementations is almost identical. As is the case for the model with the point-mass, the top 10 adverse events in Table 4.4 are very similar to Table 1.12. Again *Localised infection* is missing, evidence of the body-system effect. Of interest are the higher posterior-probabilities compared to the point-mass model. The effect of the point-mass is to reduce the posterior probabilities that θ is positive.

System Organ Class	Adverse Event	Posterior probability $\theta > 0$
Gastrointestinal disorders	Diarrhea	1.000
Skin and subcutaneous tissue disorders	Rash	1.000
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.999
Gastrointestinal disorders	Dyspepsia	0.999
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.995
Respiratory, thoracic and mediastinal disorders	Dyspnoea	0.995
Skin and subcutaneous tissue disorders	Nail disorder	0.993
Musculoskeletal and connective tissue disorders	Arthralgia	0.993
Musculoskeletal and connective tissue disorders	Back pain	0.991
Skin and subcutaneous tissue disorders	Dry skin	0.989

Table 4.4. Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry model without point-mass (c212.1a), `R` implementation.

System Organ Class	Adverse Event	Posterior probability $\theta > 0$
Gastrointestinal disorders	Diarrhea	1.000
Skin and subcutaneous tissue disorders	Rash	1.000
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.999
Gastrointestinal disorders	Dyspepsia	0.999
Respiratory, thoracic and mediastinal disorders	Dyspnoea	0.995
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.995
Skin and subcutaneous tissue disorders	Nail disorder	0.993
Musculoskeletal and connective tissue disorders	Arthralgia	0.993
Musculoskeletal and connective tissue disorders	Back pain	0.992
Skin and subcutaneous tissue disorders	Dry skin	0.989

Table 4.5. Trial EGF100151: Top 10 adverse events by posterior probability, Berry and Berry model without point-mass (c212.1a), `OpenBUGS` implementation.

4.4 Error Controlling Procedures

Table 1.12 shows the adverse events which are significant at the 5% level for an end of study Fisher exact test. These are unadjusted test results and typically we would consider applying a multiple hypothesis error controlling procedure to this output.

4.4.1 Double False Discovery Rate

Applying the DFDR at the 5% or 10% levels to final safety data from GSK EGF100151, without removing the adverse events referenced in Step 1 of the DFDR procedure, flags the following adverse events as significant:

System Organ Class	Adverse Event
Gastrointestinal disorders	Diarrhea
Skin and subcutaneous tissue disorders	Rash

Table 4.6. Trial EGF100151: adverse events flagged as significant by the Double False Discover Rate.

This is in agreement with the trial conclusions. We can see how the DFDR works, and in particular the effect that Step 1 has on the process, by working through the steps of the procedure (§2.4.3). The overall structure of the adverse events is given in Table 1.11. We let $\alpha = 0.05$ and look at two body-systems, *Gastrointestinal disorders*, which contains 58 adverse events, and *Respiratory, thoracic and mediastinal disorders*, which contains 36. For *Gastrointestinal disorders* the p-values from the Fisher exact test for this body-system are as follows:

$< 0.0000001, 0.0044047, 0.0700341, 0.0917583, 0.1071696, 0.1122317, 0.1252955,$
$0.2184040, 0.2259454, 0.2498130, 0.2498130, 0.2704908, 0.3746945, 0.4763092,$
$0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4763092,$
$0.4763092, 0.4763092, 0.4763092, 0.4998753, 0.4998753, 0.4998753, 0.6070449,$
$0.6248167, 0.6248167, 0.6248167, 0.6284277, 0.6887956, 0.7263064, 0.7596677,$
$0.8894033, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000,$
$1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000,$
$1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000,$
$1.0000000, 1.0000000$

Applying a BH-adjustment, as described in §2.3.2.3, gives the following values:

$0.000001, 0.127736, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$
$1.000000, 1.000000$

where, for example, we have $\tilde{p}_{(2)} = \min\left(\tilde{p}_{(3)}, \frac{58}{2}p_{(2)}\right) = \min\left(1.0, \frac{58}{2}0.0044047\right) \approx 0.127736$. The representative value for *Gastrointestinal disorders*, $p^* = \tilde{p}_{(1)}$, is 0.000001.

The ordered p-values from the Fisher exact test for *Respiratory, thoracic and mediastinal disorders* are as follows:

$$0.0042055, 0.0614218, 0.2262469, 0.2498130, 0.2498130, 0.2885880, 0.4763092,$$
$$0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4763092, 0.4998753,$$
$$0.4998753, 0.4998753, 0.5073516, 0.6248167, 0.6254917, 0.6872757, 0.7132041,$$
$$1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000,$$
$$1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000, 1.0000000,$$
$$1.0000000$$

and the corresponding adjusted p-values are:

$$0.151398, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000$$

The representative p-value here is 0.151 (to 3 decimal places). Overall, the ordered representative values for the body-systems are as follows:

$$\mathbf{0.000001}, 0.000147, \mathbf{0.151398}, 0.356819, 0.666500,$$
$$0.685699, 0.772535, 0.928340, 0.942332, 0.952618,$$
$$0.952618, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000$$

with *Gastrointestinal disorders* and *Respiratory, thoracic and mediastinal disor-*

ders highlighted in bold. Applying the p-value adjustment gives:

$$0.000020, 0.001685, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000, 1.000000, 1.000000,$$
$$1.000000, 1.000000, 1.000000$$

Only two body-systems are included in the final family F at the 0.05 level. These are *Gastrointestinal disorders* and *Skin and subcutaneous tissue disorders*. *Respiratory, thoracic and mediastinal disorders* is not included in the final family. In fact only these two will be included for any α level over 0.001658. The large numbers of adverse events which have incidence values of only 1 or 2 increase the magnitude of the representative p-values. The final step is to apply the BH procedure to F, which gives the results in Table 4.6.

We can repeat the analysis following Step 1 of the procedure if we remove the adverse events with very low counts. Removing the 326 adverse events which effect one percent of the population or less and re-running the analysis gives the same results for the 5% level, but the following results at the 10% level:

System Organ Class	**Adverse Event**
Gastrointestinal disorders	Diarrhea
Skin and subcutaneous tissue disorders	Rash
Gastrointestinal disorders	Dyspepsia
Skin and subcutaneous tissue disorders	Dermatitis acneiform

Table 4.7. Trial EGF100151: Adverse events flagged at 10% significance level by the Double False Discover Rate (low counts removed).

Removing the low count adverse events has made a difference to the overall results. We now have two additional flagged adverse events. Comparing to Table 4.2, for example, we can see that these adverse events are four of the top five flagged by the Berry and Berry model.

4.4.2 Group Benjamini-Hochberg

Applying the GBH at the 5% or 10% level to the final safety data from GSK EGF100151 gives the following adverse events as significant:

System Organ Class	Adverse Event
Gastrointestinal disorders	Diarrhea
Skin and subcutaneous tissue disorders	Rash

Table 4.8. Trial EGF100151: Adverse events flagged as significant by the Group Benjamini-Hochberg procedure.

which is in agreement with original trial conclusions and the DFDR for all safety data. Stepping through the method allows an understanding of how this conclusion is reached. We look at the *Gastrointestinal disorders* and *Respiratory, thoracic and mediastinal disorders* as we did for the DFDR. Using method TST to estimate the value of $\pi_{g,0}$ gives a value of 0.9827586 for *Gastrointestinal disorders* and 1.000000 for *Respiratory, thoracic and mediastinal disorders*. The value 1.000000 effectively excludes *Respiratory, thoracic and mediastinal disorders* from the analysis. Out of the 23 body-system only *Gastrointestinal disorders* and *Skin and subcutaneous tissue disorders* have values less than 1.0, meaning they are they only body-systems considered. The result of the final step of the GBH to is re-weight the p-values as follows:

System Organ Class	Adverse Event	p-value	Weighted[1] p-value
Gastrointestinal disorders	Diarrhea	1.487018e-08	8.476002e-07
Skin and subcutaneous tissue disorders	Rash	3.185761e-06	1.433593e-04

Table 4.9. Trial EGF100151: Adverse events flagged as significant by the Group Benjamini-Hochberg procedure, with re-weighted p-values.

[1] The weighted p-value is defined in §2.5.2.

The p-values have been increased but are still flagged by the GBH. The possibility

of this type of p-value inflation, and how it may affect the method, was discussed in §2.6. Repeating the analysis with the 326 low count events removed gives the following results at the 5% level:

System Organ Class	Adverse Event
Gastrointestinal disorders	Diarrhea
Skin and subcutaneous tissue disorders	Rash
Respiratory, thoracic and mediastinal disorders	Epistaxis

Table 4.10. Trial EGF100151: Adverse events flagged as significant at the 5% level by the Group Benjamini-Hochberg procedure (low counts removed).

and at the 10% level:

System Organ Class	Adverse Event
Gastrointestinal disorders	Diarrhea
Skin and subcutaneous tissue disorders	Rash
Respiratory, thoracic and mediastinal disorders	Epistaxis
Gastrointestinal disorders	Dyspepsia
Skin and subcutaneous tissue disorders	Dermatitis acneiform

Table 4.11. Trial EGF100151: Adverse events flagged as significant at the 10% level by the Group Benjamini-Hochberg procedure (low counts removed).

Unlike the DFDR, the GBH has included an adverse event from the *Respiratory, thoracic and mediastinal disorders* body-system. For the 10% α-level, the five flagged adverse events match the top five from the Berry and Berry model (Table 4.2).

4.5 Discussion

Applying the Bayesian methods leads to a list of potential adverse events associated with *lapatinib and capecitabine* (Tables 4.2 and 4.4). How important these are is a decision for the trial's clinical investigators. We can see that the effect of the point-mass in the Berry and Berry models is to reduce the posterior probability that θ is greater than zero compared to what it would be otherwise. This will control

the Type-I error rate but possibly at the expense of a loss of power. Without the point-mass, more adverse events have higher posterior probabilities, leading to the possibility of incorrectly flagging an adverse event.

For the DFDR and GBH including the numbers of small count adverse events in the analysis has a definite effect on the results. For the DFDR this is part of the procedure, but we can ask should such data really be discarded for the analysis? The Bayesian methods do not require this type of data discard. The GBH and DFDR effectively exclude certain groups from being tested, unlike the Bayesian approach which gives a posterior probability for all adverse events, with the adverse events with larger probabilities (that θ is positive) considered more likely to be associated with the treatment. Overall, the methods gave results similar to the trial conclusions, but with the addition of a number of adverse events for consideration highlighted by some of the methods. Table 4.12 shows the adverse events which look to be of potential interest according to the grouped methods.

System Organ Class	**Adverse Event**	**Method**
Gastrointestinal disorders	Diarrhea	c212.1a, c212.BB, DFDR, GBH
Skin and subcutaneous tissue disorders	Rash	c212.1a, c212.BB, DFDR, GBH
Respiratory, thoracic and mediastinal disorders	Epistaxis	c212.1a, c212.BB, GBH (5%)[1]
Gastrointestinal disorders	Dyspepsia	c212.1a, c212.BB, DFDR (10%)[1], GBH (10%)[1]
Skin and subcutaneous tissue disorders	Dermatitis acneiform	c212.1a, c212.BB, DFDR (10%)[1], GBH (10%)[1]

Table 4.12. Trial EGF100151: Adverse events of interest flagged by method.

[1] Flagged with low count adverse events removed.

In addition to flagging a number of adverse events of potential interest, the application of the grouped methods has shown a number of properties which we will investigate further in the simulation study in Chapter 5.

Chapter 5

Simulation Study

5.1 Introduction

In Chapters 2 and 3 we reviewed a number of approaches to safety analysis in clinical trials. These included both methods for testing multiple hypotheses and methods for modelling clinical data. In Chapter 4 we applied a number of the methods, which used groupings of events by body-system, to real clinical trial safety data. In this chapter we use a simulation study to investigate further how these grouped methods compare to each other, and to standard non-grouped approaches, over a number of different trial scenarios.

The methods we consider may be divided into two distinct categories, error controlling procedures, and modelling approaches. The error controlling procedures we include are control of the False Discover Rate by the Benjamini-Hochberg procedure (BH) [31], the Double False Discovery Rate (DFDR) [3], the Group Benjamini-Hochberg procedure (GBH) [30], the subset Benjamini-Hochberg procedure (ssBH) [57], the Bonferroni correction (BONF) [12], and unadjusted hypothesis testing (NOADJ). While the DFDR and GBH use groupings to attempt to take advantage of certain relationships in the data, the ssBH method uses groupings of hypotheses to extend the range of dependent test statistics to which a BH type FDR controlling procedure can be applied, and still control the FDR at the desired level. It is known to be as or less powerful than the BH-procedure itself in all circumstances, and is included purely for completeness. The models we look at are the hierarchical Bayesian model of Berry and Berry, with and without the point-mass [5], [60]. The methods included in the study are listed in Table 5.1.

The error controlling procedures and unadjusted hypothesis testing require the calculation of p-values. The simulated trial data we use in this chapter is binomial (§5.2) and we will follow [5] and use an exact Fisher two-sided test to calculate the p-values for differences between treatment and control. Direct comparisons between

Method Name	Description
c212.BB	Berry and Berry model [5] (model 1b from [60])
c212.1a	Berry and Berry model without point-mass [5] (Model 1a from [60])
NOADJ	No error controlling procedure
BONF	Bonferroni correction [12]
BH	FDR control by the BH-procedure [31]
DFDR	Double false discovery rate [3]
GBH	Group Benjamini-Hochberg [30]
ssBH	Subset Benjamini-Hochberg [57]

Table 5.1. Methods used in the Simulation Study.

these error controlling procedures are possible. However, direct comparisons with the Bayesian models require that the Bayesian approaches have a defined criteria for flagging adverse events. None of the models we look at have such criteria defined so, in this chapter, when comparing across the different methods, we will use, as the event flagging mechanism for the Bayesian models, nominal threshold values of 95% and 90% posterior probability that the log-odds ratio of adverse event incidence on the treatment arm compared to the control arms, θ, is greater than zero. We will also look at the model fits where appropriate, and discuss the model behaviour for the known underlying model data. For the grouped FDR methods (DFDR, GHB) we will use significance levels of both 5% and 10%.

Comparisons between the methods focuses on the numbers of adverse events correctly identified as having raised treatment rates, and on control of the error rates. We are particularly interested to see if the grouped methods perform better than comparable ungrouped approaches. As we are comparing frequentist and Bayesian methods we use definitions of error rates based on classical approaches. We define as Type-I errors those events which are flagged as having raised rates when the underlying rate in the simulation model is known not to be raised. Type-II errors are defined as those events whose underlying rates in the simulations are known to be raised but which are not flagged by the methods. Despite the classical nomenclature, the error rates for the Bayesian methods are based on Bayesian inference using the posterior distributions of the model parameters. For the Bayesian methods we are also interested in the estimates of the underlying model parameters. In the analyses of the relative performance of the methods we discuss below, the

error rates are given as totals of events over all the simulations included in that analysis. In the most general case, where all simulations are included, this gives an overall view of how the methods compare. However, this approach does have the disadvantage that the performance of the methods and the corresponding error rates in individual simulations is not easily assessed. To address this we also look at a number of individual simulations in more detail (§5.5.1).

The Bayesian models are fitted using the MCMC algorithms described in §A.2. For model c212.1a we performed sampling in two separate ways. The first approach used Metropolis-Hastings (MH) steps for the non-standard distributions, while the second approach used a slice sampler (SL). For c212.BB we used slice samplers for all non-standard distributions, apart from θ, which used an MH step as described in §A.2. The global parameter values required by the MH and SL steps used in the fitting algorithms are given in Table A.4.

All the methods are implemented in the `c212` package for `R` (Appendix A, [139]). For the Bayesian models, all results are presented under the assumption that the model has reach (approximate) convergence. Parameter tuning and model convergence is discussed in Appendix C.

5.2 Simulated Adverse Event Incidence Data Model

The trial data we simulate for the study is marginal trial incidence data. When considering incidence data, only the first occurrence of an adverse event for a particular subject is of interest, multiple occurrences of the same event are not taken into account. The data is marginal in the sense that we consider the overall probabilities of occurrences on each trial arm, rather than for individual trial subjects. We do not take event timings into account. The data generated is similar to that presented in clinical trial safety reports, an example of which is shown in Figure 4.1.

The simulation uses a logistic regression model to generate the trial incidence data. The data is assumed to correspond to the binomial model:

$$\begin{aligned} &\text{Control Group:} && X_{bj} = \text{Bin}(N_C, p_{1bj}) && 1 \leq b \leq B \\ &\text{Treatment Group:} && Y_{bj} = \text{Bin}(N_T, p_{2bj}) && 1 \leq j \leq k_b \end{aligned} \tag{5.1}$$

where N_C, N_T are the number of patients in the control and treatment groups respectively, B is the number of body-systems, body-system b contains k_b adverse events, and p_{1bj} and p_{2bj} are the probabilities of an event occurring in the control and treatment groups respectively.

The logistic model in its most general form is:

$$\text{logit}(p_{tbj}) = \mu_{tbj} + U_{tbj}, \quad t = 1, 2 \tag{5.2}$$

where

- $t = 1, 2$ are the control group and treatment group respectively.
- μ_{tbj} is a fixed underlying adverse event rate for adverse event j in body-system b and treatment group t.
- U_{tbj} is an underlying random effect for adverse event j in body-system b and treatment group t.

The probabilities can then be recovered from:

$$\text{logit}(p_{tbj}) = s \tag{5.3}$$

where s is the simulated value giving:

$$p_{tbj} = \frac{e^s}{1 + e^s} \tag{5.4}$$

The log-odds ratio for AE_{bj} between treatment and control group is:

$$(\mu_{2bj} + U_{2bj}) - (\mu_{1bj} + U_{1bj})$$

For the purposes of the simulation we are interested in detecting increases in the odds ratios or, assuming rare events, the relative risks of adverse events between the two groups.

5.3 Trial Layout and Body-Systems

In order to simulate trial data a structure for the trial must be defined based on the overall simulation goals, the number of simulations planned, and the computational resources available. Adverse events are expected to be quite rare, and

we have seen for GSK trial EGF100151 (§1.8.8) that many adverse events have extremely low occurrence rates, and that only a limited number of body-systems may have adverse events with raised rates. Taking these factors into account, we selected a relatively modest total of 8 body-systems, with the number of events in each body-system varying between 1 and 11, to give some disparity between the body-systems. Table 5.2 give the numbers of adverse events in each body-system.

Body System (b)	**Number of AEs (k_b)**
1	1
2	4
3	7
4	5
5	8
6	11
7	3
8	6

Table 5.2. Simulation body-systems and numbers of adverse events.

As well as the body-system layout, we must consider the total number of patients enrolled in each trial arm. For this study we considered three different trial sizes with participant numbers given in Table 5.3.

Trial Size	**Control Numbers (N_C)**	**Treatment Numbers (N_T)**
Small	110	110
Medium	450	450
Large	1100	1100

Table 5.3. Trial Sizes and numbers of participants used in simulation study.

5.4 Simulation Definition

5.4.1 Simulation Data Model Parameters

We split the fixed part of the data model (5.2) as follows:

$$\mu_{tbj} = \mu + \gamma_t + \delta_{tb} + \alpha_{tbj} \tag{5.5}$$

where μ is an overall mean, γ_t is a trial arm effect, δ_{tb} are body-system effects, and α_{tbj} are individual adverse event effects. For all the simulations, the overall mean (μ) and control trial arm (γ_1) effects have the fixed values given in Table 5.4.

Model Parameter	Value
μ	-4
γ_1	0

Table 5.4. Simulation study: overall mean and treatment arm parameter values.

There are seventeen different simulation scenarios in total, which may be broken into three different groups. In the first group, Table 5.5, there are possible treatment arm effects and body-system effects only. In the second group, Table 5.6, there are possible treatment arm effects and adverse event effects only.In the third group, Table 5.7, there are possible treatment arm effects and both body-system and adverse event effects.

Simulation Name[1]	Description
TDM1	There is a possible treatment arm effect. There are no body-system or adverse event effects.
TDM2	There is a possible treatment arm effect. There is a body-system effect for body-system number 3 for treatment only. There are no adverse event effects.
TDM3	There is a possible treatment arm effect. There is a body-system effect for body-system number 3 for both treatment and control. There are no adverse event effects.
TDM4a	There is a possible treatment arm effect. There is a body-system effect for body-system number 5 for both treatment and control. There is a body-system effect for body-system 3 for treatment only. There are no adverse event effects.
TDM4b	There is a possible treatment arm effect. There is a body-system effect for body-system number 5 for both treatment and control, with additional raised body-system effect for body-system 5 for treatment only. There are no adverse event effects.

Table 5.5. Simulations with trial arm and body-system effects only.

[1] Each set of simulations has a name of the form TDMn, where n is an identifier, and TDM stands for Trial Data Model.

Simulation Name	Description
TDM5	There is a possible treatment arm effect. There are no body-system effects. There is a treatment adverse event effect for one adverse event in body-system 2.
TDM6	There is a possible treatment arm effect. There are no body-system effects. There are treatment adverse event effects for two adverse events in body-system 2.
TDM7	There is a possible treatment arm effect. There are no body-system effects. There are treatment adverse event effects for three adverse events in body-system 2.

Table 5.6. Simulations with trial arm and adverse event effects only.

Simulation Name	Description
TDM8	There is a possible treatment arm effect. There is a body-system effect for body-system 3 for treatment only. There is a treatment adverse event effect for one adverse event in body-system 2.
TDM9	There is a possible treatment arm effect. There is a body-system effect for body-system 3 for treatment only. There are treatment adverse event effects for two adverse events in body-system 2.
TDM10	There is a treatment arm effect. There is a body-system effect for body-system 3 for treatment only. There are treatment adverse event effects for three adverse events in body-system 2.
TDM11	There is a treatment arm effect. There is a body-system effect for body-system 3 for both treatment and control. There is a treatment adverse event effect for one adverse event in body-system 2.
TDM12	There is a treatment arm effect. There is a body-system effect for body-system 3 for both treatment and control There are treatment adverse event effects for two adverse events in body-system 2.
TDM13	There is a treatment arm effect. There is a body-system effect for body-system 3 for both treatment and control. There are treatment adverse event effects for three adverse events in body-system 2.

TDM14	There is a treatment arm effect. There is a body-system effect for body-system 5 for both treatment and control. There is a body-system effect for body-system 3 for treatment only. There is a treatment adverse event effect for one adverse event in body-system 2.
TDM15	There is a treatment arm effect. There is a body-system effect for body-system 5 for both treatment and control. There is a body-system effect for body-system 3 for treatment only. There are treatment adverse events effects for two adverse events in body-system 2.
TDM16	There is a treatment arm effect. There is a body-system effect for body-system 5 for both treatment and control. There is a body-system effect for body-system 3 for treatment only. There are treatment adverse event effects for three adverse events in body-system 2.

Table 5.7. Simulations with trial arm, body-system, and adverse event effects.

The parameters for the fixed parts of the model are given in Tables 5.8, 5.9 and 5.10, where any effect in the tables whose value is not specified is assumed to be zero. The $\gamma, \delta, \delta_1, \delta_2$, and α values in these tables represent trial arm effects, body-system effects, and adverse event effects. These vary over the ranges defined in Table 5.11.

The random part of the model (5.5) used in the simulation is:

$$U_{tbj} = U_b \sim N(0, \sigma_I^2)$$

with $\sigma_I^2 = 0.001$ for all simulations.

Simulation Name	Treatment Arm Effect ($\boldsymbol{\gamma_2}$)	Control Body-system Effect ($\boldsymbol{\delta_{1b}}$)	Treatment Body-system Effect ($\boldsymbol{\delta_{2b}}$)	Adverse Event Effect ($\boldsymbol{\alpha_{tbj}}$)
TDM1	γ	1	1	0
TDM2	γ	1	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	0
TDM3	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	0
TDM4a	γ	$1 \quad b \neq 5$ $\delta_1 \quad b = 5$	$1 \quad b \neq 3, 5$ $\delta_1 \quad b = 5$ $\delta_2 \quad b = 3$	0
TDM4b	γ	$1 \quad b \neq 5$ $\delta_1 \quad b = 5$	$1 \quad b \neq 5$ $\delta_1 + \delta \quad b = 5$	0

Table 5.8. Simulations with trial arm and body-system only effects: fixed parameter values.

Simulation Name	Treatment Arm Effect ($\boldsymbol{\gamma_2}$)	Control Body-system Effect ($\boldsymbol{\delta_{1b}}$)	Treatment Body-system Effect ($\boldsymbol{\delta_{2b}}$)	Adverse Event Effect ($\boldsymbol{\alpha_{tbj}}$)
TDM5	γ	1	1	$\alpha_{221} = \alpha$
TDM6	γ	1	1	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$
TDM7	γ	1	1	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$ $\alpha_{223} = \alpha$

Table 5.9. Simulations with trial arm and adverse event only effects: fixed parameter values.

Simulation Name	Treatment Arm Effect ($\boldsymbol{\gamma_2}$)	Control Body-system Effect ($\boldsymbol{\delta_{1b}}$)	Treatment Body-system Effect ($\boldsymbol{\delta_{2b}}$)	Adverse Event Effect ($\boldsymbol{\alpha_{tbj}}$)
TDM8	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$
TDM9	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$
TDM10	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$ $\alpha_{223} = \alpha$
TDM11	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$
TDM12	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$
TDM13	γ	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$1 \quad b \neq 3$ $\delta_1 \quad b = 3$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$ $\alpha_{223} = \alpha$
TDM14	γ	$\delta_1 \quad b = 5$ $1 \quad b \neq 5$	$\delta_1 \quad b = 5$ $\delta_2 \quad b = 3$ $1 \quad b \neq 3, 5$	$\alpha_{221} = \alpha$
TDM15	γ	$\delta_1 \quad b = 5$ $1 \quad b \neq 5$	$\delta_1 \quad b = 5$ $\delta_2 \quad b = 3$ $1 \quad b \neq 3, 5$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$
TDM16	γ	$\delta_1 \quad b = 5$ $1 \quad b \neq 5$	$\delta_1 \quad b = 5$ $\delta_2 \quad b = 3$ $1 \quad b \neq 3, 5$	$\alpha_{221} = \alpha$ $\alpha_{222} = \alpha$ $\alpha_{223} = \alpha$

Table 5.10. Simulations with trial arm, body-system, and adverse event effects: fixed parameter values.

Parameter	Description	Value Range
γ	Trial arm effect	1.099, 0.693, 0.405, 0.262, 0.095, 0.0
δ	Body-system effect	0.1, 0.5, 1
δ_1	Body-system effect	1.1, 1.5, 2
δ_2	Body-system effect	1.1, 1.5, 2
α	Adverse Event Effect	0.1, 0.5, 1

Table 5.11. Simulation Study: Parameter ranges.

In summary, in addition to possible treatment arm effects, simulations TDM1-TDM4b include only body-system effects, TDM5-TDM7 include only adverse event effects in one body-system, and TDM8-TDM16 include both body-system and adverse effects. For example, TDM2 has raised adverse event rate in body-system 3 compared to control, TDM8 has raised treatment event rates in body-system 3 for all adverse events, and for the first adverse event in body-system 2, and TDM15 has adverse event rate raised in body-system 5 for both treatment and control, body-system 3 for treatment only, and two adverse events in body-system 2 for treatment only. The trial arm effects are chosen to be zero on the control arm for all simulations ($\gamma_1 = 0$, Table 5.4).

For each scenario above simulations were run for each combination of parameters. So, for example, TDM14 has the parameters $\gamma, \delta_1, \delta_2$ and α, which take on $6, 3, 3$, and 3 different values respectively, giving a total of 162 different combinations.

5.4.2 Simulation Structure and Adverse Events with Raised Treatment Rates

The simulation is structured to examine the methods in two separate but related ways. The first simulated data that we look at examines the effect of varying the treatment arm effect (γ_2) over the different values given in Table 5.11. When this is raised, all adverse events in the trial simulation on the treatment arm have raised rates compared to the control arms. As overall there are 17 simulation scenarios (TDM1-TDM16) where the value of γ_2 is varied, effectively repeating many cases where all treatment events in the trial have raised rates, it was felt that it was not necessary to run any repeated simulations for these cases. The numbers of adverse events with raised treatment rates are given in Table 5.12. For example,

for TDM16 where we have 45 adverse events in total, when $\gamma_2 \neq 0$ all 45 adverse events have raised treatment rates, when $\gamma_2 = 0$ we have 10 adverse events with raised treatment rates. As we have 27 different combinations of parameters other than γ, we have $10 \times 27 + 45 \times 27 \times 5 = 6345$ adverse events with raised rates, out of an overall total of $45 \times 27 \times 6 = 7290$ events.

Simulation Name	**Adverse Events with Raised Treatment Rates**	**Total Adverse Events**
TDM1	225	270
TDM2	696	810
TDM3	675	810
TDM4a	2088	2430
TDM4b	2097	2430
TDM5	678	810
TDM6	681	810
TDM7	684	810
TDM8	2097	2430
TDM9	2106	2430
TDM10	2115	2430
TDM11	2034	2430
TDM12	2043	2430
TDM13	2052	2430
TDM14	6291	7290
TDM15	6318	7290
TDM16	6345	7290
Total	39225	45630

Table 5.12. Treatment Arm Effect Simulations: Total numbers of adverse events.

The second simulated data we wish to look at is where there is no treatment arm effect ($\gamma_2 = 0$). This is the case in which we are most interested. In general, in a clinical trial, we expect that there may only be a small number of adverse events with increased rates on the treatment arm, with most rates being the same on both trial arms. For each of these simulation scenarios 500 repeated simulations were run. The total numbers of adverse events with raised treatment rates for these

simulations are given in the Table 5.13. These numbers can be calculated from Tables 5.8, 5.9, 5.10 and Table 5.11. For example, for TDM8 we have 45 adverse events in total, and raised rates for all 7 adverse events in body-system 3, and for one adverse event in body-system 2, giving 8 adverse events in total with raised rates on the treatment arm. We have 9 different combinations of parameters and this is repeated 500 times giving an expected total of $8 \times 9 \times 500 = 36000$ events with raised rates, out of an overall total of $45 \times 9 \times 500 = 202500$.

Simulation Name	**Adverse Events with Raised Treatment Rates**	**Total Adverse Events**
TDM1	0	22500
TDM2	10500	67500
TDM3	0	67500
TDM4a	31500	202500
TDM4b	36000	202500
TDM5	1500	67500
TDM6	3000	67500
TDM7	4500	67500
TDM8	36000	202500
TDM9	40500	202500
TDM10	45000	202500
TDM11	4500	202500
TDM12	9000	202500
TDM13	13500	202500
TDM14	108000	607500
TDM15	121500	607500
TDM16	135000	607500
Total	600000	3802500

Table 5.13. Repeated Simulations (no treatment arm effects): Total numbers of adverse events.

5.5 Results

The simulations were run as described in section §5.4.2, using the parameter values from §5.4.1. We describe the outputs and compare the methods by their overall adverse event detection rates, and by their Type-I and Type-II error rates. We split the analysis of the simulation up into two parts. We first look in detail at how the methods performed in a number of possibly interesting cases as the difference in treatment and control rate for adverse events varies from no difference to larger differences over the trial sizes. We also look at Bayesian model (c212.1a, c212.BB) estimation of the underlying model parameters. This is described in §5.5.1. In §5.5.2 we give the overall simulation results for both the treatment arm effect and repeated simulations.

5.5.1 Individual Simulations and Model Parameter Estimation

In this section we look at a number of simulations where $\gamma_2 = 0$. These are repeated simulations with equal underlying adverse event rates on both arms (5.2), but increased rates for some body-systems and adverse events on the treatment arms. We look at four different cases, described in Table 5.14, one from TDM1 (Table 5.8) where there is no difference in rates between treatment and control, and three from TDM15 (Table 5.10) where the increase in log-odds of an adverse event in the treatment arm compared to control is 1 (High), 0.5 (Medium), and 0.1 (Low). We use the posterior mean, averaged over the number of simulations, as the estimator for the underlying parameters in models c212.1a and c212.BB (in Bayesian decision theory the posterior mean estimator assumes a quadratic loss function). We use a 95% posterior probability threshold for the c212.1a models, and both 95% and 90% thresholds for the c212.BB model.

5.5.1.1 Equal Treatment and Control Event Rates (TDM1, SIM6)

With equal event rates between treatment and control only the Type-I error rate is of interest. For unadjusted testing using a 5% significance level we theoretically expect an error rate of 5%. However, a number of authors consider the Fisher exact test to be conservative, with the possibility that the error rate will be lower than the nominal significance level, particularly for small sample sizes [140], [141]. This type of issue may occur for any test which uses a fixed significance level when dealing with sets of discrete data.

Simulation Name	Simulation Identifier[1]	Increase in log-odds	Number of Repetitions
TDM1	SIM6	0.0	500
TDM15	SIM6	0.1	500
TDM15	SIM84	0.5	500
TDM15	SIM162	1.0	500

Table 5.14. Individual simulation cases.

[1] Each individual simulation within a set, e.g. TDM1, has an identifier of the form SIMn where n is a numeric identifier. For example, in the first row of the table SIM6 is the 6^{th} simulation within TDM1.

The results of the 500 repeated simulations are as follows:

Method	Large[1] Trial	Medium[2] Trial	Small[3] Trial	Total[4] Events
c212.1a(MH)	784(3.5%)	619(2.8%)	243(1.1%)	22500
c212.1a(SL)	776(3.4%)	626(2.8%)	237(1.1%)	22500
c212.BB	0(0.00%)	4(0.0%)	0(0.0%)	22500
c212.BB(90%)	2(0.0%)	5(0.0%)	2(0.0%)	22500
BONF	11(0.0%)	17(0.1%)	3(0.0%)	22500
DFDR(5%)	14(0.1%)	21(0.1%)	7(0.0%)	22500
DFDR(10%)	35(0.2%)	43(0.2%)	22(0.1%)	22500
BH	12(0.1%)	19(0.1%)	3(0.0%)	22500
GBH(5%)	154(0.7%)	126(0.6%)	81(0.4%)	22500
GBH(10%)	312(1.4%)	274(1.2%)	169(0.8%)	22500
NOADJ	840(3.7%)	832(3.7%)	521(2.3%)	22500
ssBH(5%)	11(0.0%)	17(0.1%)	3(0.0%)	22500
ssBH(10%)	31(0.1%)	28(0.1%)	12(0.1%)	22500

Table 5.15. TDM1, SIM6: Type-I error rates (number of events incorrectly declared significant) by trial size.

[1] The number and percentage of events declared significant in the Large trial.
[2] The number and percentage of events declared significant in the Medium trial.
[3] The number and percentage of events declared significant in the Small trial.
[4] There are 500 repeated simulations each containing 45 adverse events giving 22500 events in total.

The Type-I error rates for unadjusted testing (NOADJ) from Table 5.15 are ap-

proximately 3.7%, 3.7%, and 2.3%, for the Large, Medium and Small trials respectively, which is consistent with our expected error rate, but lower than the nominal 5% level as noted above.

The method which performed best was c212.BB, with very low Type-I error rates, even at the 90% threshold. We can see the effect of the point-mass by comparing to c212.1a, which has the second highest error rate, lower than NOADJ, but still considerably higher than the other methods. As expected, NOADJ performed the worst of all the methods considered. All the direct error controlling procedures performed well in comparison to NOADJ, with GBH having the highest error rate here.

5.5.1.2 Low Increase in Treatment Event Rate (TDM15, SIM6)

For small increases in the adverse event rate we expect that all the methods will have trouble correctly identifying significant events. The results are given in Tables 5.16, 5.17 and 5.18.

Here clearly c212.1a has the best results in terms of detecting adverse events with raised rates. For this criterion it performs better that all the other methods with only NOADJ, for the small trial size, coming close. c212.1a has better control of the Type-I error rate than NOADJ, particularly for the small trial, however the Type-I error rate is substantially higher than the other methods, NOADJ apart. For the error controlling procedures, GBH performed the best in terms of event detection. c212.BB is able to detect very few significant events and, along with the BH-procedure, is probably the poorest performing method overall. For these low event rates c212.BB is unable to overcome the effect of the point-mass, even for the medium and large trials.

The underlying model parameters (without the random effects) and the parameter estimates from the Bayesian models are plotted below (Figures 5.1, 5.2 and 5.3). For the large and medium trials c212.1a gives better estimates than c212.BB. We can see in both Figure 5.1 and Figure 5.2 that, compared to c212.1a, c212.BB underestimates the increase in log-odds in both body-system 2 and body-system 3 for treatment. For the small trial, Figure 5.3, the methods struggle to estimate the underlying model parameters, but model c212.1a is closest to the known underlying model log-odds.

Method	Correct[1]	Type-I[2]	Type-II[3]	Raised[4] Rates	Total[5] Events
c212.1a(MH)	489(10.9%)	633(3.5%)	4011(89.1%)	4500	22500
c212.1a(SL)	482(10.7%)	638(3.5%)	4018(89.3%)	4500	22500
c212.BB	3(0.1%)	2(0.0%)	4497(99.9%)	4500	22500
c212.BB(90%)	8(0.2%)	4(0.0%)	4492(99.8%)	4500	22500
BONF	9(0.2%)	13(0.1%)	4491(99.8%)	4500	22500
DFDR(5%)	9(0.2%)	13(0.1%)	4491(99.8%)	4500	22500
DFDR(10%)	15(0.3%)	33(0.2%)	4485(99.7%)	4500	22500
BH	10(0.2%)	13(0.1%)	4490(99.78%)	4500	22500
GBH(5%)	56(1.2%)	95(0.5%)	4444(98.8%)	4500	22500
GBH(10%)	117(2.6%)	230(1.3%)	4383(97.4%)	4500	22500
NOADJ	280(6.2%)	680(3.8%)	4220(93.8%)	4500	22500
ssBH(5%)	9(0.2%)	13(0.1%)	4491(99.8%)	4500	22500
ssBH(10%)	17(0.4%)	25(0.1%)	4483(99.6%)	4500	22500

Table 5.16. TDM15, SIM6: Large trial results.

[1] The total number of adverse events with raised rates that were correctly identified by the model as having a raised rate.

[2] The total number of adverse events without raised rates that were (incorrectly) identified by the model as having a raised rate.

[3] The total number of adverse events with raised rates that were not identified by the model as having a raised rate.

[4] For each simulation in TDM15 there are 9 adverse events with raised treatment rates. There are 500 repeated simulations giving 4500 events in total.

[5] Each simulation contains 45 events. There are 500 repeated simulations giving 22500 events in total.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	261(5.8%)	488(2.7%)	4239(94.2%)	4500	22500
c212.1a(SL)	257(5.7%)	489(2.7%)	4243(94.3%)	4500	22500
c212.BB	0(0.0%)	3(0.0%)	4500(100.0%)	4500	22500
c212.BB(90%)	0(0.0%)	3(0.0%)	4500(100.0%)	4500	22500
BONF	0(0.0%)	9(0.1%)	4500(100.00%)	4500	22500
DFDR(5%)	1(0.0%)	10(0.1%)	4499(100.0%)	4500	22500
DFDR(10%)	7(0.2%)	32(0.2%)	4493(99.8%)	4500	22500
BH	0(0.0%)	9(0.1%)	4500(100.0%)	4500	22500
GBH(5%)	34(0.8%)	111(0.6%)	4466(99.2%)	4500	22500
GBH(10%)	64(1.4%)	237(1.3%)	4436(98.6%)	4500	22500
NOADJ	171(3.8%)	699(3.9%)	4329(96.2%)	4500	22500
ssBH(5%)	0(0.0%)	9(0.1%)	4500(100.0%)	4500	22500
ssBH(10%)	4(0.1%)	18(0.1%)	4496(99.9%)	4500	22500

Table 5.17. TDM15, SIM6: Medium trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	104(2.3%)	176(1.0%)	4396(97.7%)	4500	22500
c212.1a(SL)	107(2.4%)	177(1.0%)	4393(97.6%)	4500	22500
c212.BB	0(0.0%)	1(0.0%)	4500(100.0%)	4500	22500
c212.BB(90%)	0(0.0%)	1(0.0%)	4500(100.0%)	4500	22500
BONF	0(0.0%)	7(0.0%)	4500(100.0%)	4500	22500
DFDR(5%)	0(0.0%)	8(0.0%)	4500(10.00%)	4500	22500
DFDR(10%)	2(0.0%)	18(0.1%)	4498(100.0%)	4500	22500
BH	0(0.0%)	7(0.0%)	4500(100.0%)	4500	22500
GBH(5%)	17(0.4%)	68(0.4%)	4483(99.6%)	4500	22500
GBH(10%)	39(0.9%)	137(0.8%)	4461(99.1%)	4500	22500
NOADJ	100(2.2%)	398(2.2%)	4400(97.8%)	4500	22500
ssBH(5%)	0(0.0%)	7(0.0%)	4500(100.0%)	4500	22500
ssBH(10%)	2(0.0%)	12(0.1%)	4498(100.0%)	4500	22500

Table 5.18. TDM15, SIM6: Small trial results.

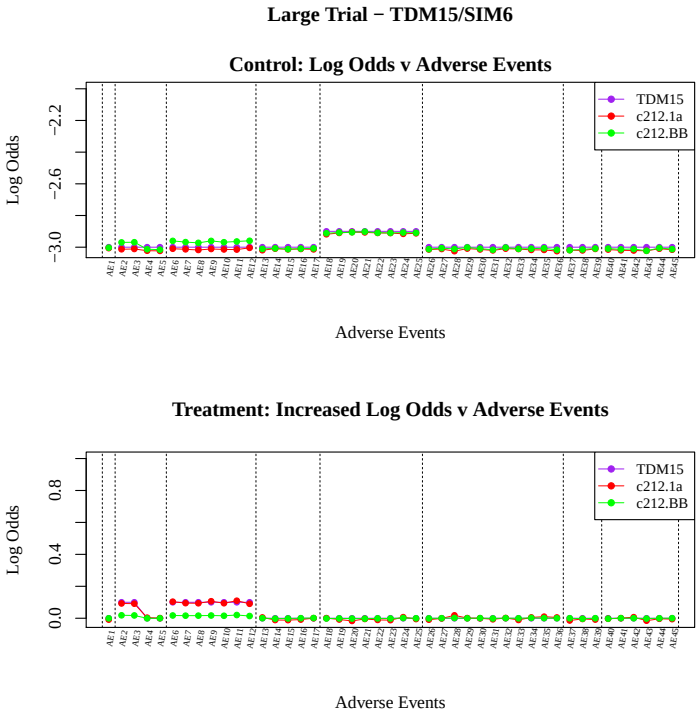


Figure 5.1. TDM15, SIM6: Large trial parameter estimates.

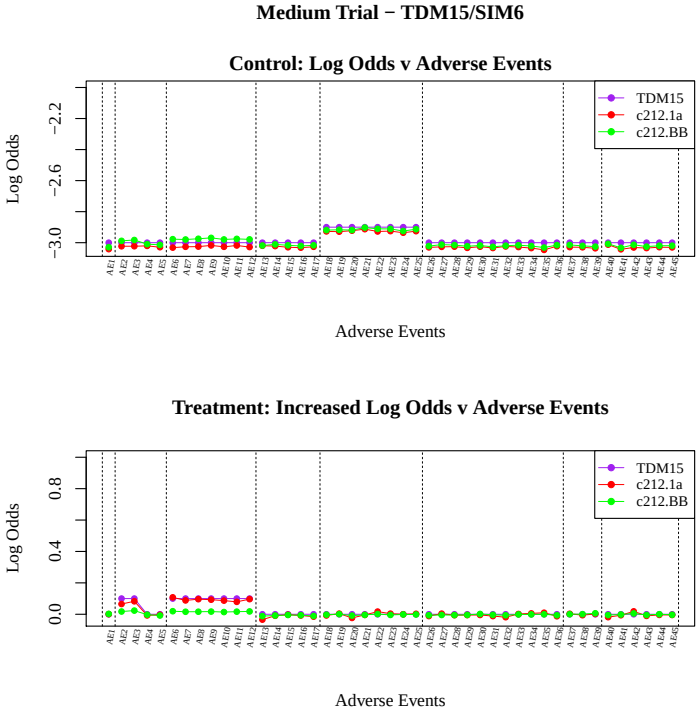


Figure 5.2. TDM15, SIM6: Medium trial parameter estimates.

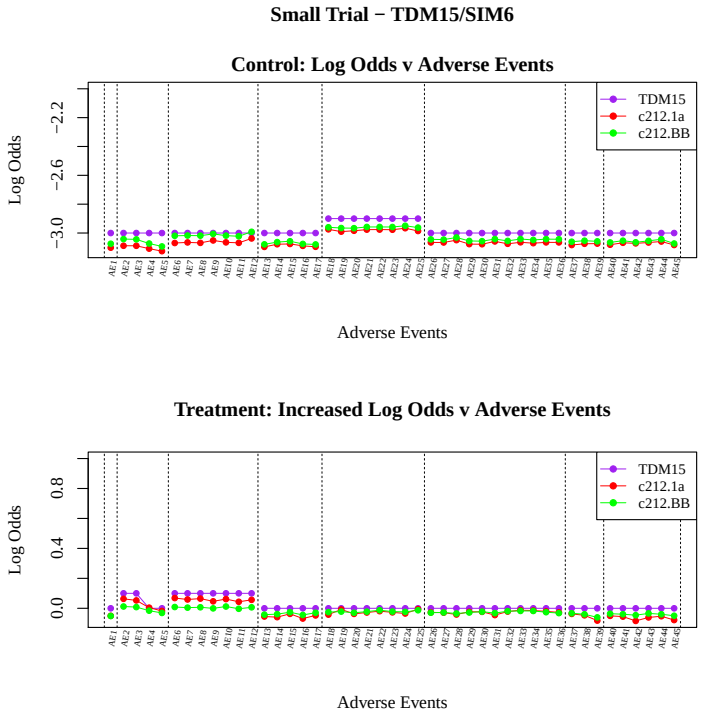


Figure 5.3. TDM15, SIM6: Small trial parameter estimates.

5.5.1.3 Medium Increase in Treatment Event Rate (TDM15, SIM84)

We expect improved performance for all the methods compared to the low increased treatment rate. The simulation results are in Tables 5.19. 5.20 and 5.21.

We can see evidence of the body-system effect in model c212.1a which, even in the small trial, detects many more events than NOADJ, but with smaller Type-I error. Overall, the results are similar to those above for low increase in event rate. c212.BB does not perform well for this lower rate, and even more poorly as the trial size decreases. However, its results are better than the DFDR (5%) for medium and small trials, and overall better than the BH-procedure. For the error controlling procedures, GBH detects most events with raised rates, but with relatively high Type-I error rate compared to the others.

Figures 5.4, 5.5, and 5.6 show how the estimated parameters match the actual underlying model parameters (without the random effects). We can see that for both control and treatment the models have successfully estimated the parameter values for the large and medium sized trials. For the small trial the estimates are not quite as good. In all cases we can also see in the treatment plots that the estimated increase in log-odds for the adverse events in body-system 2 are not as

close to the underlying values as those in body-system 3, even for the large trial size. This is not unexpected as only two of the four adverse events in body-system 2 have increased rates and have "pulled" the estimated values slightly towards each other. This is particularly noticeable for the small trial size where the models, particularly c212.BB, have trouble discerning the differences between treatment and control.

Method	**Correct**	**Type-I**	**Type-II**	**Raised Rates**	**Total Events**
c212.1a(MH)	4128(91.7%)	745(4.1%)	372(8.3%)	4500	22500
c212.1a(SL)	4131(91.8%)	755(4.2%)	369(8.2%)	4500	22500
c212.BB	2483(55.2%)	4(0.0%)	2017(44.8%)	4500	22500
c212.BB(90%)	2966(65.9%)	12(0.1%)	1534(34.1%)	4500	22500
BONF	1344(29.9%)	10(0.1%)	3156(70.1%)	4500	22500
DFDR(5%)	2934(65.2%)	50(0.3%)	1566(34.8%)	4500	22500
DFDR(10%)	3466(77.0%)	153(0.9%)	1034(23.0%)	4500	22500
BH	2099(46.6%)	84(0.5%)	2401(53.4%)	4500	22500
GBH(5%)	3621(80.5%)	145(0.8%)	879(19.5%)	4500	22500
GBH(10%)	4068(90.4%)	365(2.0%)	432(9.6%)	4500	22500
NOADJ	3475(77.2%)	755(4.2%)	1025(22.8%)	4500	22500
ssBH(5%)	1835(40.8%)	10(0.1%)	2665(59.2%)	4500	22500
ssBH(10%)	2274(50.5%)	26(0.1%)	2226(49.5%)	4500	22500

Table 5.19. TDM15, SIM84: Large trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	2891(64.2%)	522(2.9%)	1609(35.8%)	4500	22500
c212.1a(SL)	2886(64.1%)	518(2.9%)	1614(35.9%)	4500	22500
c212.BB	711(15.8%)	1(0.0%)	3789(84.2%)	4500	22500
c212.BB(90%)	1047(23.3%)	7(0.0%)	3453(76.7%)	4500	22500
BONF	250(5.6%)	5(0.0%)	4250(94.4%)	4500	22500
DFDR(5%)	505(11.2%)	20(0.1%)	3995(88.8%)	4500	22500
DFDR(10%)	898(20.0%)	59(0.3%)	3602(80.0%)	4500	22500
BH	334(7.4%)	14(0.1%)	4166(92.6%)	4500	22500
GBH(5%)	1176(26.1%)	102(0.6%)	3324(73.9%)	4500	22500
GBH(10%)	1822(40.5%)	212(1.2%)	2678(59.5%)	4500	22500
NOADJ	1668(37.1%)	626(3.5%)	2832(62.9%)	4500	22500
ssBH(5%)	305(6.8%)	5(0.0%)	4195(93.2%)	4500	22500
ssBH(10%)	477(10.6%)	15(0.1%)	4023(89.4%)	4500	22500

Table 5.20. TDM15, SIM84: Medium trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	1003(22.3%)	228(1.3%)	3497(77.7%)	4500	22500
c212.1a(SL)	1012(22.5%)	226(1.3%)	3488(77.5%)	4500	22500
c212.BB	43(1.0%)	0(0.0%)	4457(99.0%)	4500	22500
c212.BB(90%)	110(2.4%)	2(0.0%)	4390(97.6%)	4500	22500
BONF	14(0.3%)	4(0.0%)	4486(99.7%)	4500	22500
DFDR(5%)	20(0.4%)	4(0.0%)	4480(99.6%)	4500	22500
DFDR(10%)	40(0.9%)	14(0.1%)	4460(99.1%)	4500	22500
BH	14(0.3%)	4(0.0%)	4486(99.7%)	4500	22500
GBH(5%)	101(2.2%)	57(0.3%)	4399(97.8%)	4500	22500
GBH(10%)	188(4.2%)	128(0.7%)	4312(95.8%)	4500	22500
NOADJ	418(9.3%)	401(2.2%)	4082(90.7%)	4500	22500
ssBH(5%)	14(0.3%)	4(0.0%)	4486(99.7%)	4500	22500
ssBH(10%)	28(0.6%)	8(0.0%)	4472(99.4%)	4500	22500

Table 5.21. TDM15, SIM84: Small trial results.

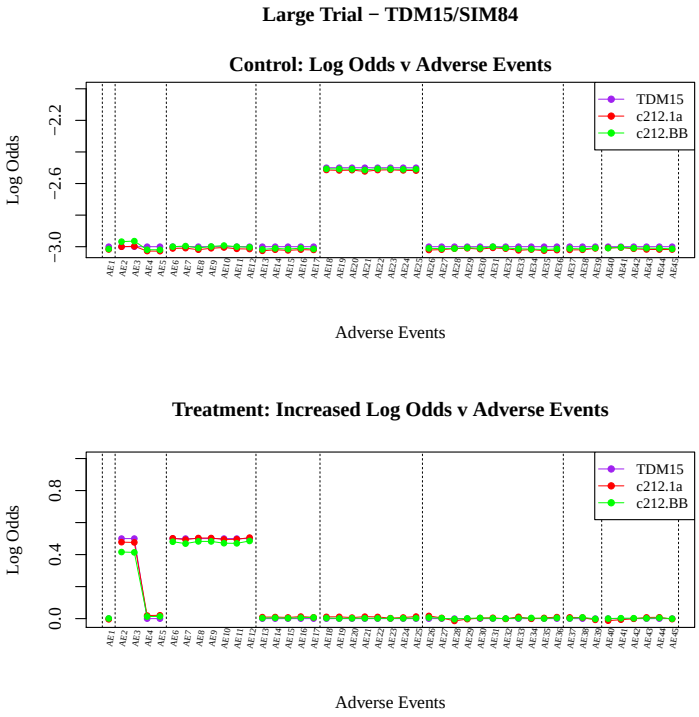


Figure 5.4. TDM15, SIM84: Large trial parameter estimates.

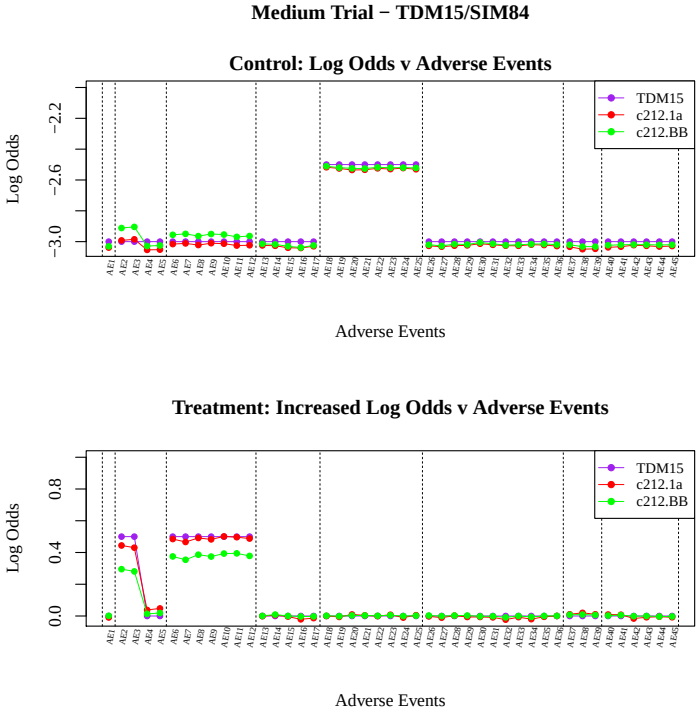


Figure 5.5. TDM15, SIM84: Medium trial parameter estimates.

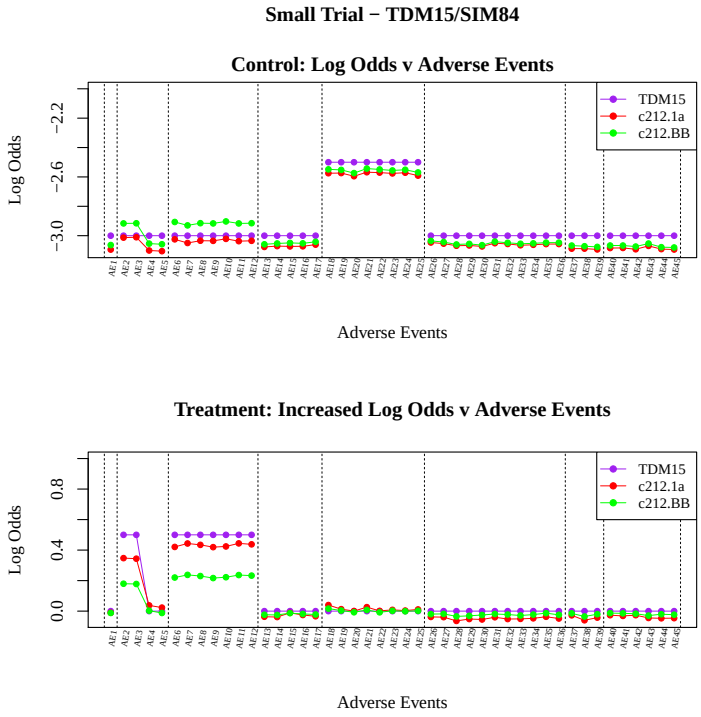


Figure 5.6. TDM15, SIM84: Small trial parameter estimates.

5.5.1.4 High Increase in Treatment Event Rate (TDM15, SIM162)

Here we expect all methods to perform well in terms of detecting significant events. The results are given in Tables 5.22, 5.23, and 5.24.

For the large trial nearly all the methods detected all the adverse events with raised rates, with only BONF, BH, and ssBH failing to detect 100% of the events. c212.BB is the best performing method with the lowest Type-I error rate. For the medium sized trial all the methods did quite well, with c212.1a detecting most events with a Type-I error rate less than NOADJ, but a lot higher than the other methods. c212.BB, DFDR(5%), and GBH(5%) detected high numbers of events, but with better error control than c212.1a. For the small trial c212.1a detected substantially more events than any other method, again with lower Type-I error rate than NOADJ. In terms of event detection, c212.BB was the next best performing, with very tight Type-I error control.

The plots of the estimated and actual parameter values are also shown below (Figures 5.7, 5.8, and 5.9). We can see that for both control and treatment the models have successfully estimated the parameter values for the large and medium sized trials. Again, as in the medium increase in treatment log-odds case, for the

small trial the estimates are not quite as good, and in all cases we can also see in the treatment plots that the estimated increase in log-odds for the adverse events in body-system 2 are not as close to the underlying values as those in body-system 3, even for the large trial size.

Comparing Figure 5.9 to Figure 5.6 we can see very similar patterns in the estimates of the underlying parameters. For body-system 3 model c212.1a has more closely estimated the underlying log-odds rates. For body-system 2 the c212.1a estimates on the treatment arm are larger than those for c212.BB. This means that for the two adverse events with raised rates in body-system 2 the estimates are closer to the underlying log-odds, but for the two which do not have raised rates c212.BB has estimated them more accurately.

Method	**Correct**	**Type-I**	**Type-II**	**Raised Rates**	**Total Events**
c212.1a(MH)	4500(100.0%)	730(4.1%)	0(0.0%)	4500	22500
c212.1a(SL)	4500(100.0%)	733(4.1%)	0(0.0%)	4500	22500
c212.BB	4499(100.0%)	5(0.0%)	1(0.0%)	4500	22500
c212.BB(90%)	4500(100.0%)	13(0.1%)	0(0.0%)	4500	22500
BONF	4492(99.8%)	16(0.1%)	8(0.2%)	4500	22500
DFDR(5%)	4500(100.0%)	71(0.4%)	0(0.0%)	4500	22500
DFDR(10%)	4500(100.0%)	161(0.9%)	0(0.0%)	4500	22500
BH	4499(100.0%)	153(0.9%)	1(0.0%)	4500	22500
GBH(5%)	4500(100.0%)	153(0.9%)	0(0.0%)	4500	22500
GBH(10%)	4500(100.0%)	335(1.9%)	0(0.0%)	4500	22500
NOADJ	4500(100.0%)	725(4.0%)	0(0.0%)	4500	22500
ssBH(5%)	4497(99.9%)	17(0.1%)	3(0.1%)	4500	22500
ssBH(10%)	4500(100.0%)	34(0.2 %)	0(0.0%)	4500	22500

Table 5.22. TDM15, SIM162: Large trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	4490(99.8%)	527(2.9%)	10(0.2%)	4500	22500
c212.1a(SL)	4490(99.8%)	526(2.9%)	10(0.2%)	4500	22500
c212.BB	4280(95.1%)	5(0.0%)	220(4.9%)	4500	22500
c212.BB(90%)	4363(97.0%)	16(0.1%)	137(3.0%)	4500	22500
BONF	3290(73.1%)	10(0.1%)	1210(26.9%)	4500	22500
DFDR(5%)	4324(96.1%)	75(0.4%)	176(3.9%)	4500	22500
DFDR(10%)	4408(98.0%)	151(0.8%)	92(2.0%)	4500	22500
BH	4054(90.1%)	123(0.7%)	446(9.9%)	4500	22500
GBH(5%)	4439(98.6%)	140(0.8%)	61(1.4%)	4500	22500
GBH(10%)	4472(99.4%)	290(1.6%)	28(0.6%)	4500	22500
NOADJ	4369(97.1%)	609(3.4%)	131(2.9%)	4500	22500
ssBH(5%)	3869(86.0%)	10(0.1%)	631(14.0%)	4500	22500
ssBH(10%)	4099(91.1%)	26(0.1%)	401(8.9%)	4500	22500

Table 5.23. TDM15, SIM162: Medium trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	3600(80.0%)	314(1.7%)	900(20.0%)	4500	22500
c212.1a(SL)	3604(80.1%)	311(1.7%)	896(19.9%)	4500	22500
c212.BB	1590(35.3%)	1(0.0%)	2910(64.7%)	4500	22500
c212.BB(90%)	2194(48.8%)	6(0.0%)	2306(51.2%)	4500	22500
BONF	215(4.8%)	5(0.0%)	4285(95.2%)	4500	22500
DFDR(5%)	513(11.4%)	12(0.1%)	3987(88.6%)	4500	22500
DFDR(10%)	954(21.2%)	43(0.2%)	3546(78.8%)	4500	22500
BH	298(6.6%)	11(0.1%)	4202(93.4%)	4500	22500
GBH(5%)	1287(28.6%)	72(0.4%)	3213(71.4%)	4500	22500
GBH(10%)	1999(44.4%)	163(0.9%)	2501(55.6%)	4500	22500
NOADJ	1845(41.0%)	440(2.4%)	2655(59.0%)	4500	22500
ssBH(5%)	268(6.0%)	5(0.0%)	4232(94.0%)	4500	22500
ssBH(10%)	452(10.0%)	10(0.1%)	4048(90.0%)	4500	22500

Table 5.24. TDM15, SIM162: Small trial results.

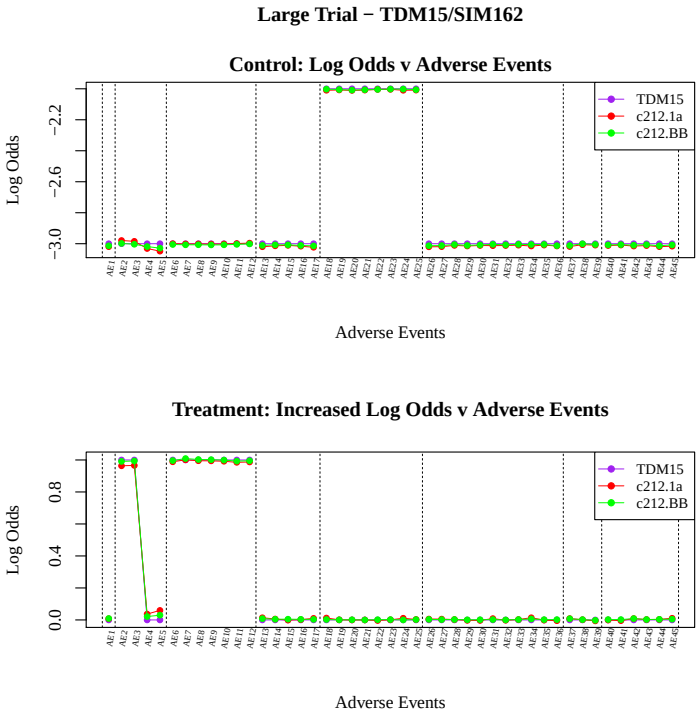


Figure 5.7. TDM15 - SIM162: Large trial parameter estimates.

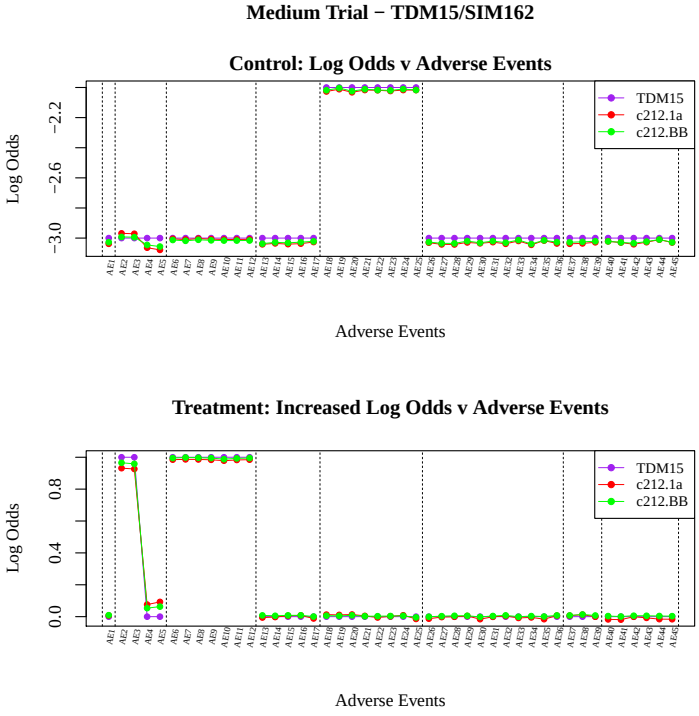


Figure 5.8. TDM15 - SIM162: Medium trial parameter estimates.

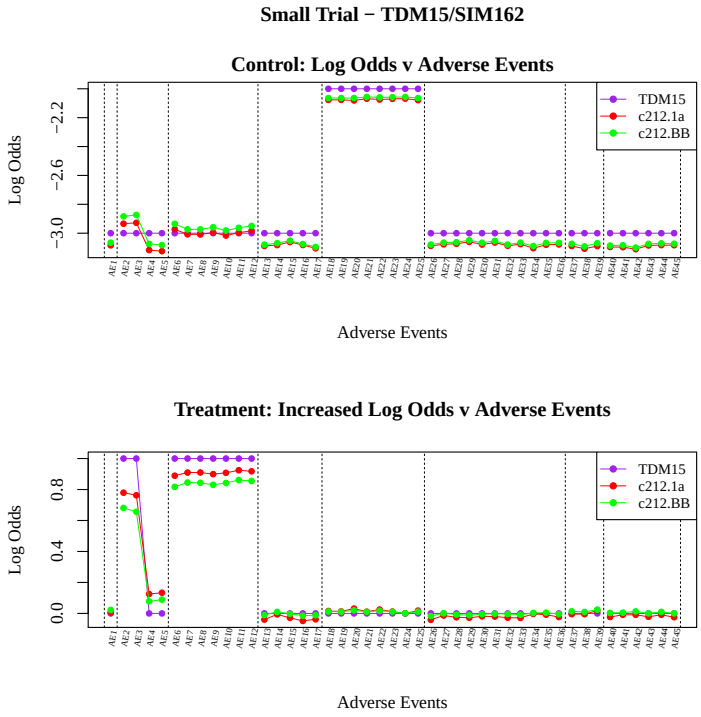


Figure 5.9. TDM15 - SIM162: Small trial parameter estimates.

5.5.1.5 Assessment

Based on this simulation we can see that c212.1a identifies the most significant adverse events in all cases, but that its Type-I error rate is high, sometimes approaching, or on occasion, exceeding that of NOADJ. It performs best of all the methods when the increase in treatment rate is low or the trial size is small. Its Type-I error rate appears to increase as the difference between treatment and control rates becomes larger. This could be anticipated based on the information sharing between body-systems in the model. c212.BB, on the other hand, has the addition of a point-mass to control the Type-I error rate. Based on the above results this works well when the difference between treatment and control rates is high, even for smaller trial sizes, but the model becomes very much less effective as the differences between treatment and control decrease. Here the effect of the point-mass becomes difficult to overcome. Dropping the flagging level for the posterior probability from 95% to 90% does improve the detection rate, but it does not approach that of c212.1a.

For the error controlling procedures, GBH detects most adverse events but with a higher Type-I error rate. DFDR also performs better than non-grouped methods,

but does not appear to be as powerful as the GBH in this simulation. However, it does have better control of the Type-I error rate than GBH.

5.5.2 Overall Results

The total numbers of adverse events and expected significant adverse events for the simulations are given in Tables 5.12 and 5.13.

5.5.2.1 Treatment Arm Effect Simulations

The results of the treatment arm effect simulations (§5.4.2, Table 5.12), in terms of the numbers of adverse events declared to have raised treatment rates or otherwise, are given in Tables 5.25, 5.26, 5.27, and 5.28, and shown graphically in Figure 5.10.

In terms of correctly identifying adverse events we can see from Figure 5.10a that the model c212.1a correctly identifies more adverse events with raised rates than any of the other models, even the unadjusted method (NOADJ). Unsurprisingly the Bonferroni method, as might have been expected, identifies the least number of adverse events, confirming that it may be too conservative for this type of data. The two subset BH methods (ssBH(5%) and ssBH(10%)) also perform poorly. This is expected, they are known to be less powerful than the BH-procedure. The other methods, including model c212.BB, which is similar to c212.1a with an additional point-mass, all perform as well as or better than the standard BH-procedure, indicating that for this data set taking the body-system into account improves adverse event detection performance.

The Type-I error rates in Figure 5.10b show that all approaches have low error rates for the data considered. As expected the unadjusted approach, NOADJ, always has the highest error rate. Model c212.1a, which correctly detects the most significant adverse events, also has high error rate, although it is an improvement on the unadjusted approach, especially for low increase in treatment rates. Of the methods which outperform the BH-procedure in terms of adverse event detection, only c212.BB has a lower Type-I error rate.

Method	Correct[1]	Type-I[2]	Type-II[3]	Raised[4] Rates	Total[5] Events
c212.1a(MH)	28141(71.7%)	237(3.7%)	11084(28.3%)	39225	45630
c212.1a(SL)	28135(71.7%)	232(3.6%)	11090(28.3%)	39225	45630
c212.BB	22285(56.8%)	2(0.0%)	16940(43.2%)	39225	45630
c212.BB(90%)	23685(60.4%)	3(0.0%)	15540(39.6%)	39225	45630
BONF	17853(45.5%)	5(0.1%)	21372(54.5%)	39225	45630
DFDR(5%)	22355(57.0%)	18(0.3%)	16870(43.0%)	39225	45630
DFDR(10%)	24152(61.6%)	37(0.6%)	15073(38.4%)	39225	45630
BH	22313(56.9%)	15(0.2%)	16912(43.1%)	39225	45630
GBH(5%)	23413(59.7%)	46(0.7%)	15812(40.3%)	39225	45630
GBH(10%)	25447(64.9%)	97(1.5%)	13778(35.1%)	39225	45630
NOADJ	24677(62.9%)	238(3.7%)	14548(37.1%)	39225	45630
ssBH(5%)	19173(48.9%)	5(0.1%)	20052(51.1%)	39225	45630
ssBH(10%)	20259(51.6%)	9(0.1%)	18966(48.4%)	39225	45630

Table 5.25. Treatment Arm Effect Simulations: Large trial results.

[1] The total number of adverse events with raised rates that were correctly identified by the model as having a raised rate.

[2] The total number of adverse events without raised rates that were (incorrectly) identified by the model as having a raised rate. There are 6405 events in the simulation which do not have raised treatment rates.

[3] The total number of adverse events with raised rates that were not identified by the model.

[4] Total events with raised treatment rates (Table 5.12).

[5] Total events in the simulation (Table 5.12).

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	24146(61.6%)	185(2.9%)	15079(38.4%)	39225	45630
c212.1a(SL)	24157(61.6%)	184(2.9%)	15068(38.4%)	39225	45630
c212.BB	17579(44.8%)	1(0.0%)	21646(55.2%)	39225	45630
c212.BB(90%)	19110(48.7%)	4(0.1%)	20115(51.3%)	39225	45630
BONF	11571(29.4%)	6(0.1%)	27654(70.6%)	39225	45630
DFDR(5%)	16169(41.2%)	12(0.2%)	23056(58.8%)	39225	45630
DFDR(10%)	17970(45.8%)	22(0.3%)	21255(54.2%)	39225	45630
BH	16118(41.1%)	17(0.3%)	23107(58.9%)	39225	45630
GBH(5%)	17596(44.9%)	48(0.7%)	21629(55.1%)	39225	45630
GBH(10%)	19522(49.8%)	98(1.5%)	19703(50.2%)	39225	45630
NOADJ	18925(48.2%)	230(3.6%)	20300(51.8%)	39225	45630
ssBH(5%)	12956(33.0%)	6(0.1%)	26269(67.0%)	39225	45630
ssBH(10%)	14169(36.1%)	12(0.2%)	25056(63.9%)	39225	45630

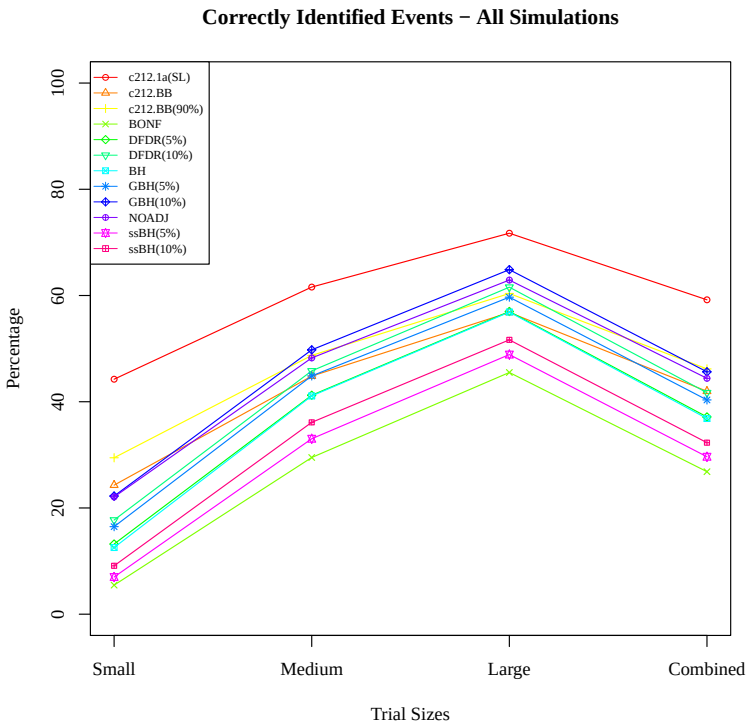
Table 5.26. Treatment Arm Effect Simulations: Medium trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	17364(44.3%)	71(1.1%)	21861(55.7%)	39225	45630
c212.1a(SL)	17353(44.2%)	69(1.1%)	21872(55.8%)	39225	45630
c212.BB	9527(24.3%)	0(0.0%)	29698(75.7%)	39225	45630
c212.BB(90%)	11550(29.4%)	1(0.0%)	27675(70.6%)	39225	45630
BONF	2146(5.5%)	0(0.0%)	37079(94.5%)	39225	45630
DFDR(5%)	5180(13.2%)	0(0.0%)	34045(86.8%)	39225	45630
DFDR(10%)	6953(17.7%)	5(0.1%)	32272(82.3%)	39225	45630
BH	4922(12.5%)	1(0.0%)	34303(87.5%)	39225	45630
GBH(5%)	6466(16.5%)	19(0.3%)	32759(83.5%)	39225	45630
GBH(10%)	8725(22.2%)	39(0.6%)	30500(77.8%)	39225	45630
NOADJ	8657(22.1%)	138(2.2%)	30568(77.9%)	39225	45630
ssBH(5%)	2746(7.0%)	0(0.0%)	36479(93.0%)	39225	45630
ssBH(10%)	3575(9.1%)	1(0.0%)	35650(90.9%)	39225	45630

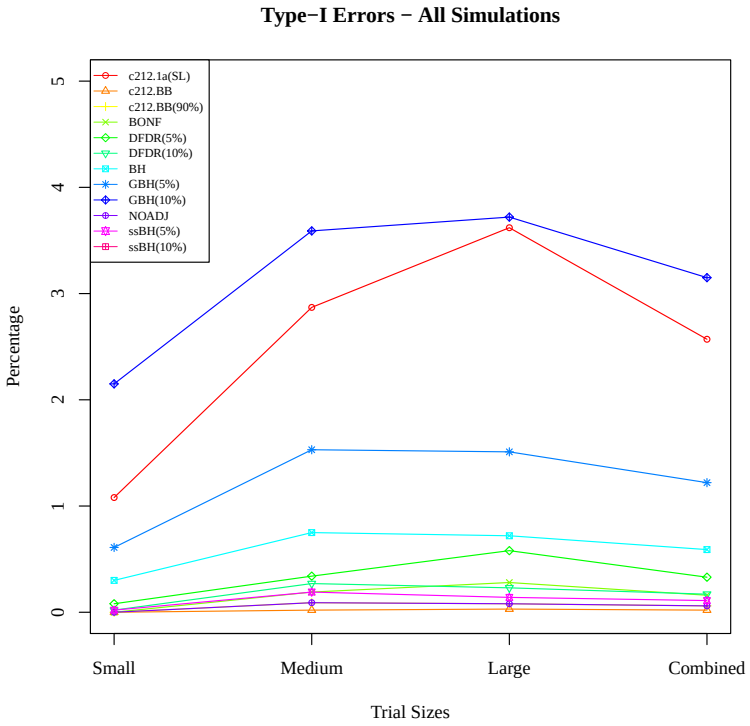
Table 5.27. Treatment Arm Effect Simulations: Small trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	69651(59.2%)	493(2.6%)	48024(40.8%)	117675	136890
c212.1a(SL)	69645(59.2%)	485(2.6%)	48030(40.8%)	117675	136890
c212.BB	49391(42.0%)	3(0.0%)	68284(58.0%)	117675	136890
c212.BB(90%)	54345(46.2%)	8(0.0%)	63330(53.8%)	117675	136890
BONF	31570(26.8%)	11(0.1%)	86105(73.2%)	117675	136890
DFDR(5%)	43704(37.1%)	30(0.2%)	73971(62.9%)	117675	136890
DFDR(10%)	49075(41.7%)	64(0.3%)	68600(58.3%)	117675	136890
BH	43353(36.8%)	33(0.2%)	74322(63.2%)	117675	136890
GBH(5%)	47475(40.3%)	113(0.6%)	70200(59.7%)	117675	136890
GBH(10%)	53694(45.6%)	234(1.2%)	63981(54.4%)	117675	136890
NOADJ	52259(44.4%)	606(3.2%)	65416(55.6%)	117675	136890
ssBH(5%)	34875(29.6%)	11(0.1%)	82800(70.4%)	117675	136890
ssBH(10%)	38003(32.3%)	22(0.1%)	79672(67.7%)	117675	136890

Table 5.28. Treatment Arm Effect simulations: Combined results.



(a) Treatment Arm Effect Simulations: Adverse events correctly flagged.



(b) Treatment Arm Effect Simulations: Type-I error rates.

Figure 5.10. Treatment Arm Effect Simulations: Correctly flagged adverse events and Type-I error rates.

5.5.2.2 Repeated Simulations

The results of the repeated simulations where $\gamma_2 = 0$ (§5.4.2, Table 5.13) are given in Tables 5.29 - 5.32, and shown graphically in Figure 5.11.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	407421(67.9%)	120161(3.8%)	192579(32.1%)	600000	3802500
c212.1a(SL)	407448(67.9%)	120088(3.7%)	192552(32.1%)	600000	3802500
c212.BB	310101(51.7%)	612(0.0%)	289899(48.3%)	600000	3802500
c212.BB(90%)	331023(55.2%)	1427(0.0%)	268977(44.8%)	600000	3802500
BONF	260527(43.4%)	2606(0.1%)	339473(56.6%)	600000	3802500
DFDR(5%)	331331(55.2%)	7617(0.2%)	268669(44.8%)	600000	3802500
DFDR(10%)	355487(59.2%)	17498(0.5%)	244513(40.8%)	600000	3802500
BH	293364(48.9%)	12134(0.4%)	306636(51.1%)	600000	3802500
GBH(5%)	366965(61.2%)	25409(0.8%)	233035(38.8%)	600000	3802500
GBH(10%)	387336(64.6%)	53869(1.7%)	212664(35.44%)	600000	3802500
NOADJ	369017(61.5%)	127726(4.0%)	230983(38.5%)	600000	3802500
ssBH(5%)	283092(47.2%)	2642(0.1%)	316908(52.8%)	600000	3802500
ssBH(10%)	304764(50.8%)	5485(0.2%)	295236(49.2%)	600000	3802500

Table 5.29. Repeated Simulations: Large trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	343723(57.3%)	90053(2.8%)	256277(42.7%)	600000	3802500
c212.1a(SL)	343700(57.3%)	90081(2.8%)	256300(42.7%)	600000	3802500
c212.BB	220915(36.8%)	631(0.0%)	379085(63.2%)	600000	3802500
c212.BB(90%)	241437(40.2%)	1733(0.1%)	358563(59.8%)	600000	3802500
BONF	158820(26.5%)	2052(0.1%)	441180(73.5%)	600000	3802500
DFDR(5%)	216998(36.2%)	4962(0.2%)	383002(63.8%)	600000	3802500
DFDR(10%)	239266(39.9%)	11763(0.4%)	360734(60.1%)	600000	3802500
BH	193068(32.2%)	7514(0.2%)	406932(67.8%)	600000	3802500
GBH(5%)	253302(42.2%)	20710(0.6%)	346698(57.8%)	600000	3802500
GBH(10%)	285172(47.5%)	44459(1.4%)	314828(52.5%)	600000	3802500
NOADJ	280286(46.7%)	112932(3.5%)	319714(53.3%)	600000	3802500
ssBH(5%)	186440(31.1%)	2061(0.1%)	413560(68.9%)	600000	3802500
ssBH(10%)	203764(34.0%)	4291(0.1%)	396236(66.0%)	600000	3802500

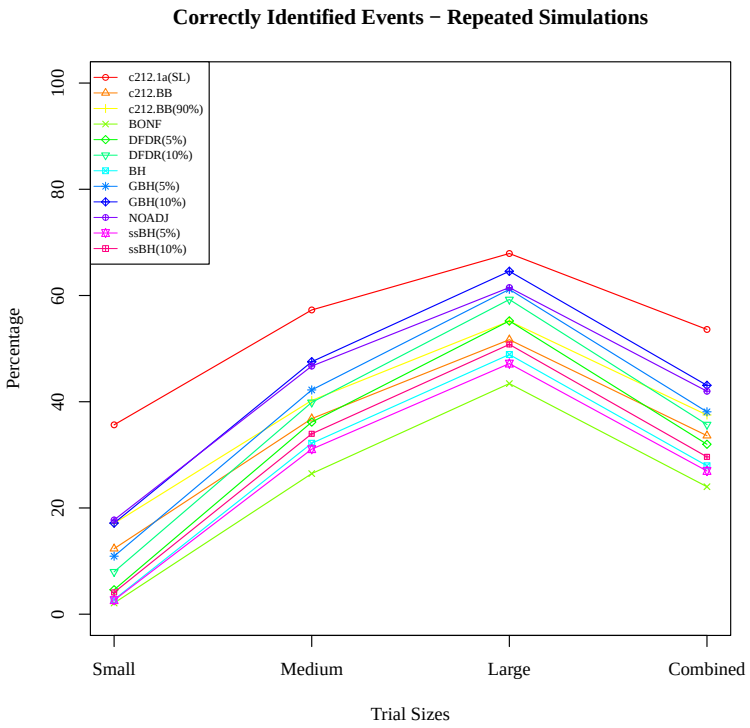
Table 5.30. Repeated Simulations: Medium trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	213809(35.6%)	42250(1.3%)	386191(64.4%)	600000	3802500
c212.1a(SL)	213884(35.6%)	42164(1.3%)	386116(64.4%)	600000	3802500
c212.BB	73995(12.3%)	298(0.0%)	526005(87.7%)	600000	3802500
c212.BB(90%)	102629(17.1%)	1020(0.0%)	497371(82.9%)	600000	3802500
BONF	12577(2.1%)	701(0.0%)	587423(97.9%)	600000	3802500
DFDR(5%)	27575(4.6%)	1113(0.0%)	572425(95.4%)	600000	3802500
DFDR(10%)	47894(8.0%)	3086(0.1%)	552106(92.0%)	600000	3802500
BH	16551(2.8%)	1037(0.0%)	583449(97.2%)	600000	3802500
GBH(5%)	65540(10.9%)	10239(0.3%)	534460(89.1%)	600000	3802500
GBH(10%)	102900(17.2%)	23711(0.7%)	497100(82.8%)	600000	3802500
NOADJ	106421(17.7%)	73353(2.3%)	493579(82.3%)	600000	3802500
ssBH(5%)	15835(2.6%)	710(0.0%)	584165(97.4%)	600000	3802500
ssBH(10%)	24613(4.1%)	1735(0.1%)	575387(95.9%)	600000	3802500

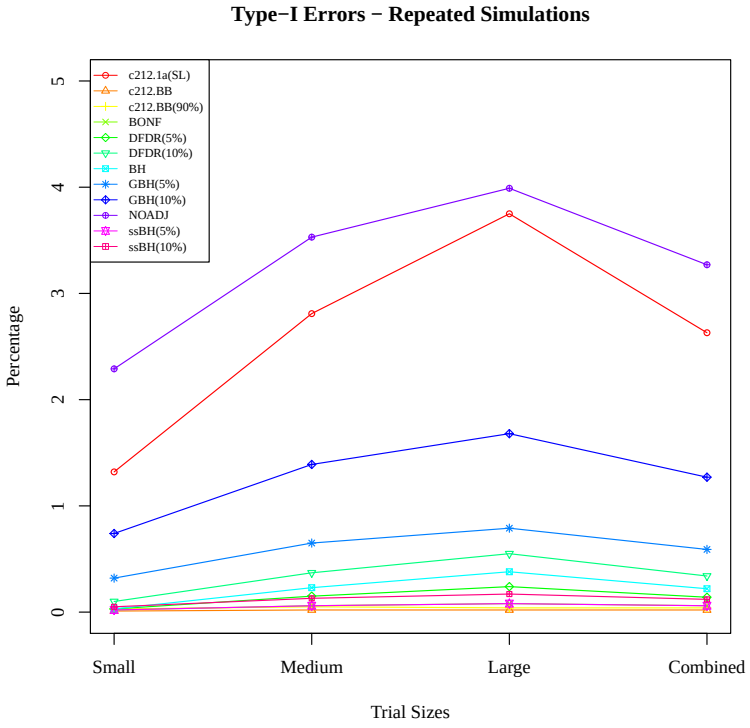
Table 5.31. Repeated Simulations: Small trial results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
c212.1a(MH)	964953(53.6%)	252464(2.6%)	835047(46.4%)	1800000	11407500
c212.1a(SL)	965032(53.6%)	252333(2.6%)	834968(46.4%)	1800000	11407500
c212.BB	605011(33.6%)	1541(0.0%)	1194989(66.4%)	1800000	11407500
c212.BB(90%)	675089(37.5%)	4180(0.0%)	1124911(62.5%)	1800000	11407500
BONF	431924(24.0%)	5359(0.1%)	1368076(76.0%)	1800000	11407500
DFDR(5%)	575904(32.0%)	13692(0.1%)	1224096(68.0%)	1800000	11407500
DFDR(10%)	642647(35.7%)	32347(0.3%)	1157353(64.3%)	1800000	11407500
BH	502983(27.9%)	20685(0.2%)	1297017(72.1%)	1800000	11407500
GBH(5%)	685807(38.1%)	56358(0.6%)	1114193(61.9%)	1800000	11407500
GBH(10%)	775408(43.1%)	122039(1.3%)	1024592(56.9%)	1800000	11407500
NOADJ	755724(42.0%)	314011(3.3%)	1044276(58.0%)	1800000	11407500
ssBH(5%)	485367(27.0%)	5413(0.1%)	1314633(73.0%)	1800000	11407500
ssBH(10%)	533141(29.6%)	11511(0.1%)	1266859(70.4%)	1800000	11407500

Table 5.32. Repeated Simulations: Combined results.



(a) Repeated Simulations: Adverse events correctly flagged.



(b) Repeated Simulations: Type-I error rates.

Figure 5.11. Repeated Simulations: Correctly flagged adverse events and Type-I error rates

As we have seen in the case of the treatment arm simulations (§5.5.2.1), the model c212.1a performs the best in terms of correctly identifying adverse events with raised treatment rates (Figure: 5.11a). Apart from the Bonferroni correction and ssBH(5%), all the methods outperform the BH-procedure, again indicating that including the body-system in the analysis is appropriate for this data.

The Type-I error rates (Figure 5.11b) have similar structure to those in Figure 5.10b, with the unadjusted approach, NOADJ, having the highest (Type-I) error rate, and model c212.1a having the second highest rate, so again it is an improvement on the unadjusted approach. Of the methods which outperformed the BH in terms of significant adverse event detection, both the DFDR(5%) and the c212.BB have lower (Type-I) error rates than the BH-procedure. ssBH again performs comparatively poorly. The (Type-I) error rates appear to be increasing with trial size for many of the methods, most obviously NOADJ and C212.1a. For NOADJ this is a reflection of the type of test being performed [140], [141]. For c212.1a though, for large trial sizes, we have seen this type of behaviour occur before in §5.5.1, the reason being that for adverse events with high treatment rates, the information sharing between body-systems in the model tends to pull up the posterior probabilities of other adverse events, and this is more evident at larger trial sizes. The point-mass in c212.BB controls this behaviour at the expense of lower adverse event detection (increased Type-II error).

5.6 Conclusions

There is an obvious need to be careful when drawing conclusions from simulation studies, due both to the nature of the data generating process used for the simulation, and the choice of cut-off points for determining adverse event significance. Analytic results regarding the FDR are well known ([31], [32]), and for the direct error controlling procedures GBH and ssBH, asymptotic and exact results are given in [30] and [57] respectively. The DFDR controls the FDR at both the event and body-system level (§D.1, §2.4.3). Mehrotra and Adewale recommend a cut-off value of 10%, which does improve the event detection rate without over-inflating the Type-I error rate in the simulations that we have studied [3].

The 95% and 90% posterior probability cut-offs chosen for the Bayesian models, c212.1a and c212.BB, are somewhat arbitrary. Moving from 95% to 90% improves the performance of the c212.BB model somewhat, without inflating the Type-I error rate, particularly for higher rate adverse events. For low treatment and

control differences the effect of the point-mass is felt most strongly, and lowering the threshold does not lead to a very large increase in the numbers detected. Xia et al. use a simulation approach to determine a suitable cut-off value for adverse event significance [60], whereas Chen et al. use a decision theoretic approach for the same purpose [125]. For our purposes the use of the 90% and 95% cut-offs allows suitable comparisons to be made between the methods, and the determination of a cut-off is not the main goal of the simulation study.

Given this we can tentatively make the following conclusions:

- For smaller sizes and lower treatment rates, c212.1a performed best of all the methods, with more events detected, and Type-I error rate lower than unadjusted testing.
- The larger the trial sizes the better all the methods and models perform in terms of event detection, although for model c212.1a there was increased Type-I error rate.
- The simulations have indicated that, for data where there are believed to be relationships between the adverse events, using groupings (body-systems) does appear to make a difference to the results. All of the group methods, with the exception of ssBH as noted in §5.1, detect more adverse events with raised treatment rates than the BH or BONF methods. However, there is a price to be paid for some of these methods in that the Type-I error may become inflated in comparison to the other methods, e.g. c212.1a. However the Type-I error rate is not excessive (less than 5% in all cases studied) and is generally lower than the unadjusted error rate. In particular, the body-system as described by Berry and Berry [5] looks to be a worthwhile structure to consider for use when modelling data.
- The point mass in the model c212.BB makes a quantitative difference to the model results. The effect of the point-mass is most noticeable for smaller trials with low rate differences between treatment and control compared to the model without a point-mass (c212.1a). This effect is perhaps larger than might have been anticipated, given that the models are closely related with one nested within the other. Although Berry and Berry [5] consider the point mass an important part of their model, in a related paper when considering hierarchical models, the authors actually choose model c212.1a (ignoring the

point mass) [67]. We can see in our simulations that with the same cut-off point for both models, the effect of the point mass is to both reduce the numbers of correctly detected adverse events and the Type-I error rate. c212.BB performs best when the differences between control and treatment are large and the trial size is large. In this case it is able to control both the Type-I and Type-II errors. For smaller rate increases c212.BB cannot detect differences between treatment and control as well as c212.1a. There is a trade-off to be made. Which model is actually more useful may depend on whether the anticipated difference, or a clinically important difference, is large, and on the relative importance of Type-I and Type-II errors. However, even with this is mind, model c212.1a appears to have more attractive properties for detecting adverse events, while still performing better than unadjusted or group adjusted error controlling procedures.

- There was no appreciable difference between using the slice sampler and MH step in fitting model c212.1a.

- For the error controlling methods, such as the DFDR, GBH, and ssBH, it may objectively be difficult to pick a method of analysing the data before the trial. The original papers proposing these methods give some asymptotic and exact results, or results based on simulations, and claim the methods are designed for specific purposes. The methods give similar results in certain cases and overall for our data, as expected, ssBH performed the worst of all these methods.

While concluding that the models and methods compared above are useful, there are also a number of areas that they do not address. None of the methods take into account timings of the events, event recurrence, or total exposure time for patients. Indeed it is not clear how best to extend the methods to handle longitudinal data and the necessity of early decision making in a clinical trial context. The severity of the adverse events (§1.6.1) is not taken into account. Individual patient data is not accounted for, and population subgroup analyses are not possible. We will look in more detail at some of these issues in the following chapters.

Chapter 6

Methods for Interim Data Analysis

6.1 Introduction

An interim analysis is a data analysis performed before data collection is complete. In the context of a clinical trial, where data is continually recorded over the trial's duration, the interim analyses are generally planned in advance. The trial's Data Monitoring Committee, which typically meets at predefined times during the trial, may use the results of the interim analyses to make decisions regarding the trial's progress. Many trial designs include mechanisms for early stopping if the data indicates large differences between treatment groups (§1.4.4). For safety data, with potentially large numbers of hypotheses and low power to determine differences between groups, the decision to alter or stop a trial is not straightforward, especially early in a trial when recruitment may not be complete and fewer events may have occurred. In this chapter we will look to extend or develop some of the methods discussed in Chapters 3 and 5 to handle interim or final study data, using a body-system approach. The questions we wish to answer are whether any particular adverse event is associated with a treatment, how early in the trial can we see this happening, and how does this relationship change over time?

6.2 Adverse Event Data

Before we consider models for the analysis of interim data, we first define exactly what data is available. As mentioned in §1.6, subjects in a clinical trial are regularly monitored and occurrences of adverse events recorded. Many of the events that may occur to a subject within a clinical trial have a duration, e.g. a rash which lasts for a number of days. Case report forms are used to collect this data (examples for non-serious and serious adverse events are shown in §1.8.3.1). We assume we

have the following data for each subject in either exact or approximate form:

D1 The patient enrolment time.

D2 The numbers, severity, and occurrence times of each adverse event, for each patient, unless that patient's data has been censored.

D3 If relevant, a censoring time for the patient.

We expect that at specific points in time (e.g. the interim analyses) we will know the exact total of all adverse events that have occurred, and the total time in study for each patient. However, in reality patients may leave the trial, miss appointments, or otherwise be unobserved. We discuss this in more detail in §6.3.1. An example of the type of raw data which we assume is available for the trial is given in Figure 6.1.

Patient Id	Time of Occurrence (days)	Body-system	Adverse Event	Group	Severity
1	113.55626497276	Bdy-sys_1	Adv-Ev_1	1	3
2	392.25414706485	Bdy-sys_1	Adv-Ev_1	1	1
3	3356.37957151388	Bdy-sys_1	Adv-Ev_1	1	1
4	132.45600207784	Bdy-sys_1	Adv-Ev_1	1	1
5	64.82706422834	Bdy-sys_1	Adv-Ev_1	1	1

Figure 6.1. Sample raw trial data.

Here the patients are identified by a numeric identifier, and the patient treatment group, time, severity, body-system, and name of each adverse event is recorded. The times are assumed to be in days relative to the date of the patients' recruitment or start of treatment.

In a recurrent event process we may characterise events as incidental or non-incidental, where non-incidental events are ones which alter the process generating the events for the future. For example, if a subject had a myocardial infarction they may now be more likely to have another one at some later stage. For these events there may be a dependence on the previous event history, with the risk of recurrence related to having previously experienced an event over a certain time period. In addition, subjects may not always be at risk from all adverse events, or there may be certain time periods where the subject is not at risk. For instance,

if an adverse event has occurred then there may be a certain amount of time before the subject is at risk of a recurrence, or, if an adverse event has a duration, then the subject may not experience that adverse event again during this period, although it may be possible that its severity increases. Additionally, there may be some events which if they have once occurred may not reoccur. We may also have additional information, either from clinical knowledge, for example due to an understanding of the treatment mechanism, or from early trial results, which may lead to an expectation that particular adverse events may have raised rates on certain treatment arms.

6.3 Analysing Interim Data

The reviews in Chapters 2 and 3 highlighted a number of approaches to analysing safety data. Many of the Bayesian approaches proposed (e.g. [5], [6] and [60]) may be considered end of study or summary methods. In contrast to this we wish to understand the cumulative presence and the trajectories, rather than just the totals of adverse events, or, in certain instances, the time to the first (possibly severe) adverse event. Models for this type of analysis need to include event timings. For frequentist approaches, parametric and semi-parametric models are common with inference generally made using maximum likelihood or partial likelihood methods, although the Mean Cumulative Function (MCF) approach of Siddiqui [4] is a non-parametric frequentist approach to this type of analysis, and [84] describes a semi-parametric Bayesian approach.

When deciding on an approach we need to establish exactly what we consider to be of importance in determining the occurrence of adverse events, how our approach will take any assumptions we make into account, and finally how the approach will allow us to flag or give some indication that an adverse event is associated with treatment. We focus mainly on treatment effects, the timings of adverse events, and the relationship between the adverse events as expressed by the body-system as the main determining features. We make three initial assumptions for the models we will introduce in this chapter:

A1 Adverse events occur in continuous time.

A2 We consider an adverse event to have occurred at a single point in time.

A3 The adverse events are related by a body-system in the sense that if a treatment affects a particular body-system then we may expect to see raised adverse event

counts for all adverse events in that body-system.

6.3.1 Censoring, Observation, Terminal Events, and Time Scales

The issues of censoring, subject observation, and the existence of terminal events may generally be considered to be related to the type of trial being performed. In certain trials the subjects may be considered to be under observation for their whole participation in the trial, in others we must consider if the subject can temporarily cease to be under observation.

There may be one or more terminal events, such as death, which may or may not be related to the recurrent events. Patients may also leave the trial, or miss an appointment, due to the occurrence of certain types of adverse events. In this case the occurrence of the events may be correlated with the censoring time of a patient. For a given patient, leaving the trial will end the event process for that patient, but, at a summary level, the death or loss to follow-up of an individual patient will not end the potential occurrence of these type of events. Any dependency between the adverse events and the terminal event will either need to be explicitly handled or not accounted for. Another important consideration is whether we may assume that the event and observation processes are conditionally independent, given the process history. We will look at the effect of patients missing from the trial due to one particular approach to censoring in §7.8.2.

There are a number of time scales which may be considered for use in the analysis. The most obvious are calendar time and study time. Many of the methods considered in Chapters 2 and 3 are based either on the time since randomisation in the study or the start of treatment. The actual time units used (e.g. hours, days, weeks etc.) will be dependent on the trial being analysed. In addition to the timescale used, there is also the possibility that time since last treatment or time since last event may be of importance. For trials where treatments are administered at time intervals, we could consider the possibility that events may occur more frequently in the immediate aftermath of a treatment, and then possibly less frequently until the next treatment is administered.

6.3.2 Body-System

The implementation of assumption A3 is dependent upon the modelling approach chosen. In a number of the models we consider we will use a body-system approach closely related to that described in the Berry and Berry model [5]. The parameters in the models will be random samples from distributions whose parameters are themselves random variables. This is a hierarchical approach which matches an assumed corresponding clinical view of how the body behaves under treatment, with regard to adverse events. Another way of looking at it is that we have exchangeability between the parameters within the body-systems at each level [142], [126]. This will be further discussed below §6.5.2.

The body-system is one means of expressing a relationship between the adverse events. However, it is essentially static in the sense that it is an assumed fixed relationship structure. There are other possible relationships between the adverse event rates which may need to be taken into account. For example, as mentioned in §6.3.1, once an adverse event has occurred it may be more likely to reoccur, and this may have an impact on what we choose to model. If we were to model occurrences of all adverse events we may need to consider the effect on the rates of adverse events in the same body-system. If we are modelling event incidence (first occurrence of an adverse event) then this will not have to be taken into account.

6.3.3 Detecting Raised Adverse Event Rates

We are primarily interested in determining which adverse events have raised rates in the treatment group as opposed to control, and when this occurs. Typically, when monitoring data in a clinical trial, all available data is analysed at each interim analysis. For example, data that has been analysed at the first interim analysis is included in the second interim analysis. As we are interested in the trajectories of the adverse event occurrences, as well as their cumulative presence, comparisons between the numbers of adverse events that have occurred over an interim analysis period are of interest. Increased rates or numbers of events on a treatment arm may be interpreted as a safety signal.

The detection of raised adverse event levels could also be considered an indicator for prediction of future adverse events. Siddiqui considers the possibility that it may be possible to use higher rates of certain adverse events to predict later occurring adverse events [4]. Similarly, adverse events may not occur for a long time. This could happen if a treatment cumulatively damages an organ but it takes a

long time for this to become apparent. These are examples of the toxicity of a treatment, where we would relate increased numbers of adverse events to dose incidence (higher doses or longer exposure implying more adverse events). Alternatively, adverse event rates may be raised after treatment, but then decline with time to non-significant levels. For example, allergic reactions or rashes may occur early in the trial but then not reappear.

6.4 Modelling Approach

We look for a modelling approach based on the assumptions of §6.3 which is relatively general and computationally tractable. Given that we assume we have access to patient level data, we could choose to model the occurrences of adverse events either at the patient level or at a summary level. As well as considering whether to model at patient or summary level, we also need to decide what aspects of the adverse events we wish to model. For example, we could consider either the total incidences of adverse event occurrence, or the total number of adverse event occurrences, as a measure of difference between two arms of a trial. Both approaches are of potential interest to trial investigators.

6.4.1 Patient Level Models

General approaches to modelling count or event data at the patient or subject level were discussed in §3.2 and §3.3. In these models each individual subject would need to be treated as a multi-event process. Information regarding the exact timings and durations of adverse events, and the periods when the patient was under observation, should be available for each individual. Incorporating this additional patient level information into the model should allow more accurate modelling of the data compared to summary models. Further, patient level models are capable of handling the case where one subject is responsible for a large number of what are otherwise rare adverse events, for example through the use of frailty terms. Censoring mechanisms are also more readily incorporated into these models in ways which may not be possible with summary models, where a censored individual does not prevent the adverse event generating process from continuing.

There are a number of drawback to using such patient level models. Multi-event processes are complex and not all subjects will experience a wide range of adverse events, particularly for rare events. Handling an over-proliferation of zero counts in this type of process may not be straightforward. There may be thousands of

individuals in any particular trial who may require explicit inclusion in the model through subject level parameters. This over-parameterisation may in effect make the model useless. We will see in Chapter 7 that even for summary level models there are occasions when this is also the case, and the models are unable to detect any events with raised rates in certain circumstances.

6.4.2 Summary Models

Summary level models require that the data for individual patients be gathered in some suitable format. Incorporating the timings of events in summary models is complicated by the fact that they occur among multiple individuals. The same is also true of censoring and numbers at risk. Two pieces of information which are important in this type of modelling are therefore the time in study for each patient, and the total number of patients in the study at any particular time-point.

For recurrent events a natural statistic for answering the question of whether any adverse event rates are the same or different in both groups over any interim period, which incorporates the event counts, numbers of subjects, and times, is the average rate of occurrence (or incidence) over that period, for both treatment and control. These values are naturally paired by the time periods, and also have the advantage of being relatively easy to explain to clinicians. As would also be the case with patient level models, in order to formulate an approach for analysing the data, we may need to take into account possible correlations between time periods for the adverse event generating processes [116], and the (assumed) body-system relationship among the adverse events. Perhaps the simplest such model is the Poisson model discussed in [60]. In this model we have a single analysis point, the end of trial, and the counts are assumed to be Poisson based. A single parameter, θ_{bj}, is used to indicate an increased relative-risk in the treatment arm.

Summary models are more in keeping with the ICH guidelines (§1.4.3) than their equivalent patient level models, and, given that they both model the same data, we may expect that summary models and patient level models should give broadly similar results in many situations.

6.4.3 Incidence and Recurrent Event Analysis

The time to the first event, or time to the first event of a certain severity, is considered an important measure of the safety of a treatment. A severe adverse event on the treatment arm in a much shorter time than on the control arm may be

an indicator of a potential safety issue. We may consider the incidence of adverse events as a survival process, where a subject drops out of the risk set for an adverse event when they experience that event. In §6.5 and §6.6 we look in more detail at a number of models based on the relationship between the Poisson process and proportional hazards models for the interim analysis of summary data, which includes the Poisson models discussed in [60] as a subset.

An alternative to looking at event incidence is to consider all events which occur over the duration of a trial, a recurrent event analysis. As discussed in §3.3.1.1 this has the advantage of giving an overall view of the trajectories of all adverse events over the clinical trial, and not just the first occurrences.

6.4.4 Sequential Analysis of Interim Data

One of the aims of interim analyses is to try to identify as early as possible in the trial if more adverse events are accumulating on the treatment arm, and if this is related to the treatment as opposed to being by chance. One approach for data sets consisting of a single type of event is to use α-spending functions with higher initial threshold or boundary earlier in the trial, and then lower boundaries as the trial goes on, but with over all (Type-I) error rate of α, where typically $\alpha = 5\%$ [11]. For multiple types of adverse events using this or similar approaches is complicated by the multiplicity of events, and their possible relationships. One advantage to using a Bayesian approach to modelling is that models with body-system structures or hierarchies have the possibility of providing a level of multiple comparison robustness through the choice of prior [27]. Another advantage is that a Bayesian approach is naturally suited to this type of sequential analysis, where more and more data arrives over time. We expect the power to detect adverse events to be quite low (§1.5.1) and so it may be only with the accumulation of data over time that any relationships between the treatment and the adverse events will become apparent.

6.5 Models for Interim Adverse Event Analysis

In this section we introduce models for analysing counts of events as they occur over the duration of the trial. For reasons given in §6.4.1 and §6.4.2 we look at summary models. We are particularly interested in assessing the incidences of adverse events at particular time points, both in terms of calendar dates, such as for the interim analysis meetings of the DMC, and also relative to the start of

the trial, such as the number of events which occur in the first six months under treatment as opposed to control.

We model the occurrences of adverse events as stochastic processes[1] generated by the individuals in the study. One way of doing this is to consider counts of the incidence of adverse events over time as a survival process as follows:

1. when a patient is enrolled to the study they become a part of the set of individuals who can generate a particular adverse event - a risk set for the adverse event;

2. when a patient experiences an adverse event (possibly of a particular severity) they are removed from the corresponding risk set for that adverse event, only the first adverse event for that patient is counted.

Referring to Figure 6.1 the patient with identifier 1 will leave the risk set of `Adv-Ev_1` after approximately 113.6 days have elapsed.

This type of survival-counting procedure is compatible with, and comparable to, the approach of [5], where adverse event incidence is compared between trial arms, for example at the end of a trial. It removes the possibility of one patient being responsible for a large number of adverse events on a particular arm of a trial, and also the possibility that the occurrence of the event in a patient may in some way change later occurrences of the same event (non-incidental adverse events). A typical frequentist approach to this type of problem would be via a proportional hazards model. We wish to introduce a body-system so we look to fit a Poisson model which should provide a good approximation to the survival type approach [143], [144], [145], [146]. These Poisson models are also interpretable in there own right. Indeed there is an exact correspondence between proportional hazard models (and many different types of recurrent event models) and Poisson models [76], [147]. In particular, for piecewise constant baseline hazard models, the correspondence with a Poisson model with piecewise constant rates is exact.

As well as providing a useful approximation for more complicated hazard models, the Poisson process is often considered the canonical approach for modelling count data via an intensity function [74]. For adverse event data there is the possibility that events in the same body-system may be related in some systematic way, and that the independent intervals property may not hold. Random or mixed models,

[1] One for each adverse event and covariate pattern.

for example the hierarchical model proposed in [5], are one way of extending the Poisson approach. As we are interested in analyses at certain time points, in the models we discuss here, we break the overall trial duration into a number of non-overlapping intervals, and attempt to use this approach to understand differences between different arms of the trial. This requires that raw trial data, such as that in Figure 6.1, be summarised by interval. Figure 6.2 gives an example of this type of summarised data for all event severities. Here, for each interval, body-system, adverse event, and treatment group, the total number of events (Count) and the total time at risk in the interval for all patients (Exposure) in each particular interval has been tabulated.

Interval	Body-system	Adverse Event	Group	Count	Exposure (days)
0.0-180.0	Bdy-sys_1	Adv-Ev_1	1	87	160133.6919932150
0.0-180.0	Bdy-sys_1	Adv-Ev_1	2	103	163224.6442895000
180.0-360.0	Bdy-sys_1	Adv-Ev_1	1	80	145054.3643045380
180.0-360.0	Bdy-sys_1	Adv-Ev_1	2	63	149107.1982650570
360.0-540.0	Bdy-sys_1	Adv-Ev_1	1	71	114792.3812445560
360.0-540.0	Bdy-sys_1	Adv-Ev_1	2	61	120634.9665069450
540.0-720.0	Bdy-sys_1	Adv-Ev_1	1	34	74648.9007535142
540.0-720.0	Bdy-sys_1	Adv-Ev_1	2	47	76697.2180531145

Figure 6.2. Sample summary trial data (all event severities).

6.5.1 Poisson Process Models for Adverse Events

We divide the trial duration, follow-up, and later periods into intervals $I_1, ..., I_{H+1}$ with $t_1, ..., t_H$ the end times for the first H intervals. Relatively speaking an individual joins the trial at the start of interval I_1 and, if not lost to follow-up, leaves the trial at the end I_H. For incidence models, if an individual first has an adverse event AE_{bj} in interval I_h, then we consider the individual is no longer at risk of this type of adverse event from this time point onward.

The model is a piecewise constant conditional Poisson model over the intervals or, to look at it another way, we have an assumption of constant hazard rate over the intervals [145], [146]. The data model is defined as follows:

Let there be B body-systems with k_b adverse events in body-system b, and let AE_{bj} be the j^{th} adverse event in body-system b. If there are C different covariate

patterns[2] among the data, let $\mathcal{R}^{(c)}_{bj,h}$ be the set of patients at risk of the adverse event AE_{bj} at the start of the interval I_h with covariate pattern c, let t_{ih} be the length of the time individual i spends in interval I_h, $\mathbf{x}_{(c)}$ be a vector of covariates, and $\boldsymbol{\theta}_{bj,h}$ a vector of parameters. The model is:[3]

$$\begin{aligned} X^{(c)}_{bj,h} | \lambda^{(c)}_{bj,h} &\sim \text{Poisson}\left(\lambda^{(c)}_{bj,h} T^{(c)}_{bj,h}\right) \\ T^{(c)}_{bj,h} &= \sum_{i \in \mathcal{R}^{(c)}_{bj,h}} t_{ih} \\ \log \lambda^{(c)}_{bj,h} &= \gamma_{bj,h} + \mathbf{x}^T_{(c)} \boldsymbol{\theta}_{bj,h} \\ c &= 1, \ldots, C \\ h &= 1, \ldots, H \\ b &= 1, \ldots, B \\ j &= 1, \ldots, k_b \end{aligned} \tag{6.1}$$

$T^{(c)}_{bj,h}$ is the total time spent in interval h for all subjects with covariate pattern c who have not yet experienced the adverse event j in body-system b. $X^{(c)}_{bj,h}$ is the count of events in interval h for all subjects with covariate pattern c who have not yet experienced the adverse event j in body-system b, and $\lambda^{(c)}_{bj,h}$ is the corresponding underlying rate parameter.

6.5.2 Independence and Exchangeability

In order to construct likelihoods or joint probability distributions and perform inference some assumptions regarding the dependence of the response variables, in our case the counts or timings of events, must be made. Exchangeability is a formal expression of the idea that we find no systematic reason to distinguish between individual variables.[4] Exchangeability is closely related to the idea of independently and identically distributed random variables in that a sequence of identical random variables which are independent, conditional on an underlying

[2]Typically there will be just 2: treatment and control.

[3]If the actual times of the events are not recorded accurately then we could, for example, let $T^{(c)}_{bj,h} = |\mathcal{R}^{(c)}_{bj,h}||I_h|$. In this case we have made the assumptions that the events occur at the end of the interval I_h.

[4]Exchangeability may be defined as follows: Let Y_i be a sequence of random variables, then the Y_i are exchangeable if the joint density of $Y_{i1}, \ldots, Y_{in}$ is invariant under a permutation of the indices $i1, \ldots, in$.

distribution, is exchangeable. Exchangeability is a key concept for Bayesian analysis as well as classical analysis. If we consider the data model in §3.6.1, the counts in the model are considered independent given the model parameters. This is a common assumption in hierarchical models. The likelihood function or joint probability density corresponding to the model (6.1) (assuming independence of the variables and fixed effects) is:

$$L = \prod_{c=1}^{C} \prod_{h=1}^{H} \prod_{b=1}^{B} \prod_{j=1}^{k_b} \frac{\exp\left(-\lambda_{h,bj}^{(c)} T_{h,bj}^{(c)}\right) \left[\lambda_{h,bj}^{(c)} T_{h,bj}^{(c)}\right]^{x_{h,bj}^{(c)}}}{x_{h,bj}^{(c)}!} \tag{6.2}$$

A fixed effect modelling approach means treating all counts as either independent of each other over the intervals, or requires the specification of a correlation structure for the data. The model can be extended to account for possible relationships between the counts within different adverse event groupings (body-systems), and also over the time intervals, by introducing random effects. When the $\lambda_{bj,h}^{(c)}$ are themselves a hierarchy of random variables there are a number of possible approaches for fitting the model. Taking a Bayesian approach, we can choose a hierarchy of prior distributions and, assuming conditional independence in the model, fit the parameters using Markov Chain Monte Carlo (MCMC) methods, for example using a Gibbs sampler [124]. Bayesian fitting is not the only possible approach to fitting random effect models. Generalised Linear Mixed Model (GLMM) techniques, using likelihood approaches, are also possible, although they do not have the flexibility of the Bayesian approach, and, in particular, it is not clear how they would handle a point-mass term such as that in [5].

6.5.3 Modelling Body-System and Longitudinal Relationships

The (hierarchical) relationship proposed by [5] is one approach for within body-system correlation. To model possible correlations between the counts in the different intervals a dependency between interval counts could be introduced. We consider this type of three-level hierarchical model in §6.6.

In Chapter 5 we have seen that the Berry and Berry model with point-mass does not perform as well as some of the other modelling approaches when the rate differences between trial arms is low, or the trial size is small. For a full three-level hierarchical implementation, including point-mass, we may expect similar problems,

particularly when we consider the larger number of parameters used to model the rates over the intervals. We can reduce the number of parameters in the models by removing part of the hierarchy. This may be achieved in a number of ways, the most straightforward of which is to remove the lowest level of random variables (§3.6.2.5), the parameters $\left\{\mu_{\gamma 0}, \mu_{\theta 0}, \tau^2_{\gamma 0}, \tau^2_{\theta 0}, \alpha_\pi, \beta_\pi, \alpha_{\sigma\gamma}, \beta_{\sigma\gamma}, \alpha_{\theta\gamma}, \beta_{\theta\gamma}\right\}$, leaving a two-level hierarchy, but still containing a body-system grouping. We look at these reduced parameter models in §6.7.

For modelling possible relationships between the different intervals we consider three levels of dependence as follows: level 0, where counts in different interval are considered to be independent; level 1, where there is a common body-system mean across the intervals; and level 2, where, in a three-level hierarchy, there is a common set of random parameters at the lowest level in the hierarchy.

The models we investigate are listed in Table 6.1 and their definitions given below. We consider models with and without a point-mass, and use the notations BB and 1a respectively to refer to the various versions of these models in the remainder of this study. We use subscripts to indicate the number of levels in the hierarchy and the level of dependency assumed in the model. For example, BB_{31} is a point-mass model, with a three-level hierarchy, and common body-system means across the intervals. The joint distributions and complete conditional distributions are given in Appendix B. The most general case is the three-level model with a point-mass where counts in each interval are considered to be independent (level 0) (§6.6.1).

Method[1]	Hierarchy[2]	Interval Dependency[3]	Description
BB_{30}	3-level	level 0	Point-mass; Independent intervals (§6.6.1)
BB_{31}	3-level	level 1	Point-mass; related intervals (§6.6.2)
BB_{32}	3-level	level 2	Point-mass; weakly related intervals (§6.6.3)
$1a_{30}$	3-level	level 0	No point-mass; independent intervals (§6.6.4)
$1a_{31}$	3-level	level 1	No point-mass; related intervals (§6.6.4)
$1a_{32}$	3-level	level 2	No point-mass; weakly related intervals (§6.6.4)
BB_{20}	2-level	level 0	Point-mass; independent intervals (§6.7.1)
BB_{21}	2-level	level 1	Point-mass; related intervals (§6.7.3)
$1a_{20}$	2-level	level 0	No point-mass; independent intervals (§6.7.4)
$1a_{21}$	2-level	level 1	No point-mass; related intervals (§6.7.4)

Table 6.1. Hierarchical methods for interim analyses.

[1] BB methods are those models with a point-mass term. 1a methods are those models without a point-mass term. The subscripts on the methods refer to the number of levels in the hierarchy and the level of dependence between the intervals. For example, BB_{31} is a three-level hierarchy with level 1 dependence between the intervals.

[2] The models may implement either two-level or three-level Bayesian hierarchies.

[3] Level 0 dependence means different intervals are independent. Level 1 dependence has common body-system means across the intervals. Level 2 dependence has relationships between the intervals at the lowest level of the hierarchy, where applicable.

6.6 Poisson Bayesian Models: Three-Level Hierarchies

The models we consider in this section use a body-system based on the model structure from [5] described in §3.6.1.

6.6.1 $\mathbf{BB_{30}}$ Poisson Point-mass Model (level 0)

This is the most general model we consider which includes a point-mass and a body-system hierarchy based on [5]. There is an individual hierarchy for each trial interval, and each interval is considered to be independent of the other intervals. This is the most straightforward extension of the Berry and Berry model to multiple interval data [5], [60].

$$\begin{aligned} X_{bj,h}^{(c)} &\sim \text{Poisson}(\lambda_{bj,h}^{(c)} T_{bj,h}^{(c)}) \\ T_{bj,h}^{(c)} &= \sum_{i \in \mathcal{R}_{bj,h}^{(c)}} t_{ih} \\ \log \lambda_{bj,h}^{(c)} &= \gamma_{bj,h} + x_{(c)} \theta_{bj,h} \end{aligned} \tag{6.3}$$

$$h = 1, \ldots, H, \quad b = 1, \ldots B_h, \quad j = 1, \ldots k_{bh}$$
$$c = 1, 2; \;\; x_{(1)} = 0, \;\; x_{(2)} = 1$$

The priors for the model parameters and hyperparameters are given in equations (6.4) - (6.6). As this is a three-level hierarchical model we have three levels of priors.

$$\gamma_{bj,h} \sim N(\mu_{\gamma b,h}, \sigma_{\gamma b,h}^2) \qquad \theta_{bj,h} \sim \pi_{b,h} \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1 - \pi_{b,h}) N(\mu_{\theta b,h}, \sigma_{\theta b,h}^2) \tag{6.4}$$

$$\begin{aligned} \mu_{\gamma b,h} &\sim N(\mu_{\gamma 0,h}, \tau_{\gamma 0,h}^2) & \mu_{\theta b,h} &\sim N(\mu_{\theta 0,h}, \tau_{\theta 0,h}^2) \\ \sigma_{\gamma b,h}^2 &\sim \text{IG}(\alpha_\gamma, \beta_\gamma) & \sigma_{\theta b,h}^2 &\sim \text{IG}(\alpha_\theta, \beta_\theta) \\ \pi_{b,h} &\sim \text{Beta}(\alpha_{\pi,h}, \beta_{\pi,h}) \end{aligned} \tag{6.5}$$

$$
\begin{aligned}
\mu_{\gamma 0,h} &\sim N(\mu_{\gamma 00}, \tau^2_{\gamma 00}) & \mu_{\theta 0,h} &\sim N(\mu_{\theta 00}, \tau^2_{\theta 00}) \\
\tau^2_{\gamma 0,h} &\sim \mathrm{IG}(\alpha_{\gamma 00}, \beta_{\gamma 00}) & \tau^2_{\theta 0,h} &\sim \mathrm{IG}(\alpha_{\theta 00}, \beta_{\theta 00}) \\
\alpha_{\pi,h} &\sim M(\lambda_\alpha)\,\mathrm{I}(\alpha_{\pi,h} > 1) & \beta_{\pi,h} &\sim M(\lambda_\beta)\,\mathrm{I}(\beta_{\pi,h} > 1)
\end{aligned} \tag{6.6}
$$

The following model hyperparameters all have common values over the intervals based on the values used in [5]:

$$
\begin{aligned}
&\mu_{\gamma 00} = 0, \tau^2_{\gamma 00} = 10, \alpha_\gamma = 3, \beta_\gamma = 1, \alpha_{\gamma 00} = 3, \beta_{\gamma 00} = 1, \lambda_\alpha = 1 \\
&\mu_{\theta 00} = 0, \tau^2_{\theta 00} = 10, \alpha_\theta = 3, \beta_\theta = 1, \alpha_{\theta 00} = 3, \beta_{\theta 00} = 1, \lambda_\beta = 1
\end{aligned} \tag{6.7}
$$

Let

$$
\begin{aligned}
X_{bj,h} &= X^{(1)}_{bj,h}, \quad T^{(1)}_{bj,h} = C_{bj,h} \\
Y_{bj,h} &= X^{(2)}_{bj,h}, \quad T^{(2)}_{bj,h} = T_{bj,h}
\end{aligned}
$$

From the parameters

$$
\begin{aligned}
\log \lambda^{(1)}_{bj,h} &= \gamma_{bj,h} \\
\log \lambda^{(2)}_{bj,h} &= \gamma_{bj,h} + \theta_{bj,h}
\end{aligned}
$$

we have that $\theta_{bj,h}$ is the log of the relative risk for AE_{bj} in interval h. A positive value for $\theta_{bj,h}$ indicates an increased risk of adverse event AE_{bj} occurring on the treatment arm over interval I_h. The joint probability distribution for the model and its complete conditionals are given in §B.12.

6.6.2 $\mathbf{BB_{31}}$ Poisson Point-mass Model (level 1)

We may expect that the occurrences of any particular adverse event over different intervals may be similar in some way, for example the rates may be raised over all trial intervals or for early intervals in the trial, and we may wish to take this into account in our model. Introducing a random effect into a standard Poisson model, such as that described in [92, Ch. 2], introduces a correlation between the intervals. We can introduce a similar relationship by restricting the parameters in §6.6.1 as follows:

$$
\begin{aligned}
&\mu_{\gamma 0,h} = \mu_{\gamma 0}, \quad \mu_{\theta 0,h} = \mu_{\theta 0} \\
&\tau^2_{\gamma 0,h} = \tau^2_{\gamma 0}, \quad \tau^2_{\theta 0,h} = \tau^2_{\theta 0} \\
&\mu_{\gamma b,h} = \mu_{\gamma b}, \quad \mu_{\theta b,h} = \mu_{\theta b} \\
&\sigma^2_{\gamma b,h} = \sigma^2_{\gamma b}, \quad \sigma^2_{\theta b,h} = \sigma^2_{\theta b} \\
&\alpha_{\pi,h} = \alpha_{\pi}, \quad \beta_{\pi,h} = \beta_{\pi} \\
&\pi_{b,h} = \pi_b
\end{aligned}
\tag{6.8}
$$

In this case we have common body-means across the intervals, and a common hierarchy below them, with corresponding changes to the joint distribution and complete conditional distributions. These are given in §B.13.

6.6.3 BB_{32} Poisson Point-mass Model (level 2)

Parameter relationships, such as those specified in BB_{31}, are not the only ones that can exist within this type of model. Due to the model's hierarchical nature, we can also consider a weaker relationship between the intervals by restricting the model BB_{30} in §6.6.1 as follows:

$$
\begin{aligned}
&\mu_{\gamma 0,h} = \mu_{\gamma 0}, \quad \mu_{\theta 0,h} = \mu_{\theta 0} \\
&\tau^2_{\gamma 0,h} = \tau^2_{\gamma 0}, \quad \tau^2_{\theta 0,h} = \tau^2_{\theta 0} \\
&\alpha_{\pi,h} = \alpha_{\pi}, \quad \beta_{\pi,h} = \beta_{\pi}
\end{aligned}
\tag{6.9}
$$

with the definitions of other model parameters remaining unchanged. Compared to BB_{31}, the relationship between the parameters is "lower" in the hierarchy, and may be expected to provide a weaker correlation. The joint distribution and complete conditionals are given in §B.14.

6.6.4 Poisson Models Without Point-mass

The models without the point-mass ($1a_{3l}$, $l = 0, 1, 2$) may be defined by setting $\pi_{b,h} = 0$ in §6.6.1, §6.6.3, and $\pi_b = 0$ in §6.6.2. The complete conditional distributions may be derived in a similar fashion. The joint and complete conditional distributions are given in §B.9, §B.10 and §B.11.

6.6.5 Summary

The body-system we use is the same as that introduced by [5]. The 1a models are nested within their BB equivalents at each interval dependence level (0, 1, 2). Based on the simulation study in Chapter 5 we expect the 1a models to detect more events with raised treatment rates than their BB equivalents, but to have higher Type-I error rates. For lower rate differences, or small trials, we may expect that the BB models will detect very much less events with raised treatment rates than their 1a counterparts, due to the difficulty of overcoming the effect of the point-mass and the proliferation of additional parameters compared to the 1a models.

6.7 Poisson Bayesian Models: Two-Level Hierarchy

We can consider a modelling approach with a reduced number of parameters by removing the lowest part of the three-level hierarchy. This allows two possible levels of dependence only, independent intervals, and common body-system means across the intervals. For models without the point-mass this is relatively straightforward, but for models with a point-mass care needs to be taken. In the three-level hierarchical models with point-mass the parameters $\pi_{b,h}, \pi_b$ are modelled by Beta priors whose parameters are themselves random variables. In moving to a two-level model the parameters will have fixed values and we will see (§6.7.2) that choosing these values may not be straightforward.

6.7.1 $\mathbf{BB_{20}}$ Poisson Point-mass Model (Level 0)

The most general two-level model we consider includes a point-mass, independent intervals, and a body-system hierarchy as follows:

$$\begin{aligned} X_{bj,h}^{(c)} &\sim \text{Poisson}(\lambda_{bj,h}^{(c)} T_{bj,h}^{(c)}) \\ T_{bj,h}^{(c)} &= \sum_{i \in \mathcal{R}_{bj,h}^{(c)}} t_{ih} \\ \log \lambda_{bj,h}^{(c)} &= \gamma_{bj,h} + x_{(c)} \theta_{bj,h} \end{aligned} \tag{6.10}$$

$$\begin{gathered} h = 1, \ldots, H; \quad b = 1, \ldots B_h, \quad j = 1, \ldots k_{bh} \\ c = 1, 2; \quad x_{(1)} = 0; x_{(2)} = 1 \end{gathered}$$

with the following priors for the model hyperparameters:

$$\gamma_{bj,h} \sim N(\mu_{\gamma b,h}, \sigma_{\gamma b,h}^2) \qquad \theta_{bj,h} \sim \pi_{b,h} \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1 - \pi_{b,h}) N(\mu_{\theta b,h}, \sigma_{\theta b,h}^2) \tag{6.11}$$

$$\begin{aligned} \mu_{\gamma b,h} &\sim N(\mu_{\gamma 0}, \tau_{\gamma 0}^2) & \mu_{\theta b,h} &\sim N(\mu_{\theta 0}, \tau_{\theta 0}^2) \\ \sigma_{\gamma b,h}^2 &\sim \text{IG}(\alpha_\gamma, \beta_\gamma) & \sigma_{\theta b,h}^2 &\sim \text{IG}(\alpha_\theta, \beta_\theta) \\ \pi_{b,h} &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \end{aligned} \tag{6.12}$$

There is an individual hierarchy for each trial interval. Apart from α_π and β_π, the lowest level model hyperparameters, $\left\{\mu_{\gamma 0}, \tau_{\gamma 0}^2, \alpha_\gamma, \beta_\gamma, \mu_{\theta 0}, \tau_{\theta 0}^2, \alpha_\theta, \beta_\theta\right\}$, all have common values over the intervals based on the values used in [5]:

$$\begin{aligned} \mu_{\gamma 0} = 0, \tau_{\gamma 0}^2 = 10, \alpha_\gamma = 3, \beta_\gamma = 1 \\ \mu_{\theta 0} = 0, \tau_{\theta 0}^2 = 10, \alpha_\theta = 3, \beta_\theta = 1 \end{aligned} \tag{6.13}$$

As for the three-level models we write

$$\begin{aligned} X_{bj,h} &= X_{bj,h}^{(1)}, \quad T_{bj,h}^{(1)} = C_{bj,h} \\ Y_{bj,h} &= X_{bj,h}^{(2)}, \quad T_{bj,h}^{(2)} = T_{bj,h} \end{aligned}$$

and we have:

$$\begin{aligned}\log \lambda^{(1)}_{bj,h} &= \gamma_{bj,h} \\ \log \lambda^{(2)}_{bj,h} &= \gamma_{bj,h} + \theta_{bj,h}\end{aligned}$$

and $\theta_{bj,h}$ is again the log of the relative risk. The joint distribution and complete conditional distributions are given in §B.7.

6.7.2 Choice of Prior for $\pi_{b,h}$

The prior Beta(a, b) is a natural choice for a probability in a hierarchical model [126]. When a and b are both 1 this becomes the (uninformative) uniform distribution. When a and b are both less than 1 the probability may become concentrated close to 0 or 1. In particular, in the limit as a or b tend to zero, the Beta distribution approaches a point-mass at 0 or 1. In [5], where the prior for π_b is Beta(α_π, β_π), this possibility is handled by restricting α_π and β_π, which are themselves random variables, to having values greater than 1. We would ordinarily like to choose the uninformative uniform distribution as our prior for $\pi_{b,h}$. However, this does include the possibility that the probabilities may become concentrated at the edges in the complete conditional distributions (B.37). When using a Gibbs sampling MCMC approach to model fitting the values of the parameters change on each iteration of the sampler. This may or may not be a serious issue, but choosing values of $\alpha_{\pi,h}$ and $\beta_{\pi,h}$ larger than but close to 1 would allow an approximation of a uniform prior while, excluding the possibility of an edge concentrated complete conditional.

6.7.3 $\mathbf{BB_{21}}$ Poisson Point-mass Model (level 1)

We restrict the parameters in §6.7.1 to have common body-system means across the intervals as follows:

$$\begin{aligned}\mu_{\gamma b,h} &= \mu_{\gamma b}, \quad \mu_{\theta b,h} = \mu_{\theta b} \\ \sigma^2_{\gamma b,h} &= \sigma^2_{\gamma b}, \quad \sigma^2_{\theta b,h} = \sigma^2_{\theta b} \\ \pi_{b,h} &= \pi_b\end{aligned} \tag{6.14}$$

The joint distribution and complete conditionals are given in §B.7.

6.7.4 Poisson Models Without Point-mass

The models without the point-mass ($1a_{2l}$, $l = 0, 1$) may be defined by setting $\pi_{b,h} = 0$ in §6.7.1 and $\pi_b = 0$ in §6.7.3. The complete conditional distributions may be derived in a similar fashion. The joint and complete conditional distributions are given §B.5 and §B.6.

6.8 Flagging Adverse Events as Having Raised Treatment Rates

None of the models discussed above have a formal process for flagging an adverse event as having a raised treatment rate, but as they are updated at each interim analysis, they may provide an emerging picture of how the adverse events occur on each trial arm. The model parameter $\theta_{bj,h}$, the log relative risk (§6.6.1), is one approach to assessing increased treatment rates. A large posterior probability that this is greater than zero is an indication of a raised treatment rate over an interval.

The value of $\theta_{bj,h}$ may change over the course of the trial as more data accumulates. There are two important times we may need to consider when deciding to flag an event. The first is the time of the recruitment of the last patient to the trial. After a time $|I_1|$ from this point all patients will have been through the first interval, I_1, of the trial, and no more events can occur during this interval. If we choose to model the intervals independently, then any estimates from the model for this interval will not change, subject to minor differences due to the random nature of MCMC fitting. Fitting independent models to each interval leaves any unadjusted conclusions vulnerable to the type of multiple comparison issue which we are trying to avoid. For models where the intervals are not considered to be independent this will not be the case and the estimates will continue to change as more data accumulates later in the trial, so we are more limited in what we can say, especially early in the trial, particularly for the higher level of dependence (level 1).

The second time we need to consider is the time when the first recruited patients have been followed up to the end of their participation in the trial. At this point we will have information on the rates over all intervals and can possibly be more confident in what we say, particularly about earlier intervals. This also raises the issue of there being no information at early analyses for the later intervals. The straightforward approach to this is to fit the models without this information, just

fitting parameters for which data exists. Interim analyses where intervals may have only a small number of events of each particular type may occur throughout the early part of the trial. This will affect the model fitting process, particularly for the point-mass models where we have extra parameters but little data.

Assessing the reliability of early interval results is not straightforward. As more data accumulates we should be more confident that our conclusions are correct. A possible approach to assessing early model predictions of differences between treatment and control would be to examine how the parameters would vary under different assumptions about the following intervals. We could assume, for any intervals for which we do not have any information, that the rates for treatment and control are equal to a background rate, if such a rate exists. For models with related intervals this would ensure that any early conclusions regarding increases in treatment rate versus control rate would be more robust.

We will look at some of these issues and explore some of differences between the models' performance when we compare them in a demonstration analysis in the next chapter.

Chapter 7

Demonstration Interim Analyses

7.1 Introduction

We illustrate the methods from Chapter 6 using simulated data for a number of complete (Phase III) clinical trials. The trial data is simulated at the individual patient level with the adverse events experienced by the patients having both a time of occurrence and a severity grade. There are many different possible combinations of event rates and severities which could be applied to each patient in the simulation but unlike the study in Chapter 5, where we compared the different methods over a number of varying parameters with regard to event flagging and error rates, our goal here is to demonstrate the methods, so we take only a small number of possibly interesting scenarios and look to give an indication of how the methods work in practice, and how they compare to each other over this small data set.

The trial data simulation is mainly based on the hierarchical body-system approach (Figure 3.1). We choose an underlying overall adverse event rate, or possibly rates, for the trial. The adverse event rate in each particular body-systems is then a random sample from a normal distribution whose mean is an overall rate. A number of adverse events on the treatment arm have increased rates compared to the control over some intervals. The mechanism for generating the simulated adverse event data and event severities is described in Appendix E.

The major assumptions of the data simulation process are that the underlying adverse event rates for the control arm do not vary over the intervals, and that the event rates and the probability of an event of a particular severity are the same for each patient in the trial. We also assume that each trial has similar numbers of recruits on each arm and that the recruitment rates are constant. While these assumptions are open to criticism in terms of how generally applicable they may be, they are in line with the body-system approach we are taking, and are suitable

to the type of demonstration analysis we wish to perform.

In the first part of the demonstration analysis we look at a single underlying rate and three different trial scenarios. In the first scenario the adverse event rate is raised for treatment in one body-system over the whole duration of the trial. In the second, the adverse event rate is raised for treatment in one body-system early in the trial but then the rates fall-off to the same levels as for control. In the third, the adverse event rate between body-systems are the same early in the trial but become raised for one body-system later in the trial. For each trial we look at the effects of different increases in event rate for treatment arm. We consider two different types of analysis, an incidence analysis where we look at only the first occurrence of an event for a patient, and an analysis where all events are included. For each type of analysis we consider two cases, one where we include events of all severities, denoted severity 1+, and a second case where we just include severe events, grade 3 or higher, denoted severity 3+. The methods are also applied to the clinical trial GSK EGF100151 (§1.8) although here the absence of adverse event timings means we have to make a number of assumptions when applying the models.

A sensitivity analysis is then performed where we consider much lower treatment rates, a mixture of two different treatment rates, the effect of changing the thresholds used for flagging events, and the effect of missing data on the models. For trials consisting of the lower rate events we are interested in seeing how the increase in treatment rate affects the correct flagging of events and the error rates, given the already low rate of occurrence.

Although we are performing a Bayesian analysis, for consistency we use the same terms for the error rates (Type-I and Type-II) as we used in Chapter 5. All the results and parameter estimations reported are derived from a Bayesian inference based on the posterior distributions of the model parameters.

The models, described Table 6.1, are implemented in the `c212` package for `R` and fitted using Markov Chain Monte Carlo (MCMC) methods (Appendix A, [139]). Slice samplers were used for all non-standard distributions apart from the point-mass models where θ was sampled using a Metropolis-Hastings step (§A.2). All results are presented under the assumption that the models have reached (approximate) convergence (Appendix C).

7.2 Trial Structure

The simulated trials have the common structure described Table 7.1. Patients are recruited for the first 720 days of the trial (approximately 2 years), at a rate of 1.3 per day on each arm, up to a maximum of 1000 patients per arm. Each patient is followed up for a total 1800 days (approximately 5 years). The trial will be completed after a maximum of 2520 days corresponding to last possible recruitment time for a patient on the trial (720 days) plus the follow-up period (1800 days), approximately 7 years in total.

Trial Parameter	**Value**
Unit of Time	Day
Trial Start Time	0
Patient Recruitment Period	0 - 720
Patient Follow-up	1800
Trial End	2520
Control Recruitment Rate	1.3 per day
Treatment Recruitment Rate	1.3 per day

Table 7.1. Demonstration Analysis: Common trial details.

The Data Monitoring Committee (DMC) planned safety reviews are scheduled to take place every 360 days from the start of the trial until the end of follow up for the last recruited patient. They are as follows:

Time	**Analysis**
360	Initial Safety Review (1)
720	Safety Review (2)
1080	Safety Review (3)
1440	Safety Review (4)
1800	Safety Review (5)
2160	Safety Review (6)
2520	Final Safety Review (7)

Table 7.2. Demonstration Analysis: Planned safety reviews.

7.2.1 Body-Systems and Adverse Event Severity

The adverse events for each trial are grouped into 15 Body-systems with 155 adverse events in total. The total events in each body-system is given in Table 7.3.

Body System	**Number of AEs** (k_b)
Bdy-sys_1	10
Bdy-sys_2	8
Bdy-sys_3	7
Bdy-sys_4	8
Bdy-sys_5	9
Bdy-sys_6	11
Bdy-sys_7	7
Bdy-sys_8	6
Bdy-sys_9	9
Bdy-sys_10	14
Bdy-sys_11	19
Bdy-sys_12	8
Bdy-sys_13	16
Bdy-sys_14	14
Bdy-sys_15	9

Table 7.3. Demonstration Analysis: Simulation body-systems and numbers of adverse events.

Associated with each event is a severity grade ranging from 1-5 based on the NCI CTCAE (Table 1.1). The probability that any adverse event has a particular severity is assumed to be the same for treatment and control. These probabilities are given in Table 7.4. If a patient has a severity 5 adverse event then he/she no longer contributes to the trial.

Severity	Description	Probability
1	Mild	0.5
2	Moderate	0.3
3	Severe	0.1
4	Life-threatening	0.0999
5	Death	0.0001

Table 7.4. Demonstration Analysis: Adverse event severity probabilities.

7.2.2 Intervals

The approach we have taken in defining our models (§6.5.1) is to split the trial duration up into a number of intervals. The intervals need to be large enough to accumulate enough events to allow us to determine which, if any, events have increased rate on the treatment arm. We split the trial into the following 10 intervals covering the total 1800 days a patient is under observation. In this case a number of the interval end-points coincide with the DMC safety reviews (Table 7.2) although this may not always be appropriate:

$$[0, 180], (180, 360], (360, 540], (540, 720], (720, 900], (900, 1080],$$
$$(1080, 1260], (1260, 1440], (1440, 1620], (1620, 1800]$$

7.2.3 Trial Types

We simulate data for the following 3 trial types:

Trial Type	Description
I	Rate raised for one body-system across all intervals.
II	Rate raised for one body-system over first two intervals.
III	Rate raised for one body-system over final two intervals.

Table 7.5. Demonstration Analysis: Trial types.

7.3 Trial Simulation Parameters

7.3.1 Background Trial Adverse Event Rates

Each trial in the demonstration analysis has a background event rate. The following table gives the probability of an event, and the expected number of number of events per patient, over the course of the trial, assuming events occur at the background rate according to a Poisson process:

Rate[1]	Expected Number[2] of Events	Probability of[3] an Event
0.0005555	0.9999	0.6320838

Table 7.6. Demonstration Analysis: Background trial adverse event rate.

[1] In units of events per day.
[2] Expected number of events over the course of the trial (1800 days): 0.0005555 * 1800 = 0.9999.
[3] The probability an individual has one or more events over the course of the trial.

7.3.2 Increased Treatment Rates

For all trials the treatment rates are the same as the control rates apart from one body-system, Bdy-sys_3. The increased treatment rates, given as a percentage increase of the corresponding control adverse event rates, and the intervals over which they apply are given in Table 7.7. We look at a number of different treatment increases for each trial type.

Simulated[1] Trial	Body-System	Intervals[2]	Relative Increase in[3] Treatment Rate
I(a)	Body-sys_3	All	100%
I(b)	Body-sys_3	All	50%
I(c)	Body-sys_3	All	10%
II(a)	Body-sys_3	[0-180]	100%
II(a)	Body-sys_3	(180-360]	50%
II(b)	Body-sys_3	[0-180]	50%
II(b)	Body-sys_3	(180-360]	25%
II(c)	Body-sys_3	[0-180]	20%
II(c)	Body-sys_3	(180-360]	10%
III(a)	Body-sys_3	(1440-1620]	50%
III(a)	Body-sys_3	(1620-1800]	100%
III(b)	Body-sys_3	(1440-1620]	25%
III(b)	Body-sys_3	(1620-1800]	50%
III(c)	Body-sys_3	(1440-1620]	10%
III(c)	Body-sys_3	(1620-1800]	20%

Table 7.7. Demonstration Analysis: Body-systems and intervals with increased treatment rates.

[1] Unique simulation name consisting of trial type (I, II, III) and identifier ((a), (b), (c)).
[2] The intervals over which the background treatment rate is increased.
[3] For example, for Trial III(c) there is a 10% increase in the underlying treatment rate in Body-sys_3 over the interval (1440-1620]. The background underlying rates are given in Table 7.6.

7.4 Adverse Events with Raised Treatment Rates

There are 7 adverse events in Bdy-sys_3 and 155 adverse events in total. The trial duration is split into 10 intervals (§7.2.2). Table 7.8 gives the total numbers of adverse events with raised treatment rates where an adverse event is counted once in each interval in which its rate is raised.

Time	Trial Type	Number of[1] Intervals in Analysis	Adverse Events[2] with Raised Treatment Rates	Total[3] Adverse Events
360	I	2	14	310
720	I	4	28	620
1080	I	6	42	930
1440	I	8	56	1240
1800	I	10	70	1550
2160	I	10	70	1550
2520	I	10	70	1550
360	II	2	14	310
720	II	4	14	620
1080	II	6	14	930
1440	II	8	14	1240
1800	II	10	14	1550
2160	II	10	14	1550
2520	II	10	14	1550
360	III	2	0	310
720	III	4	0	620
1080	III	6	0	930
1440	III	8	0	1240
1800	III	10	14	1550
2160	III	10	14	1550
2520	III	10	14	1550

Table 7.8. Demonstration Analysis: Adverse event and interval totals per trial.

[1] The number of intervals for which we have patient data.
[2] Total number of adverse events in Bdy-sys_3 (7) × Number of intervals with raised rates (Table 7.7).
[3] Total number of adverse events (155) × Number of Intervals in Analysis.

For the first 4 analyses we do not have data for a number of intervals. For example, for the analysis at day 720 there are no patients who have been in the trial for longer than 720 days. The total number of events over all simulated trials at each interim analysis is given in Table 7.9.

Time	Number of Intervals in Analysis	Adverse Events[1] with Raised Treatment Rates	Total[2] Adverse Events
360	2	84	2790
720	4	126	5580
1080	6	168	8370
1440	8	210	11160
1800	10	294	13950
2160	10	294	13950
2520	10	294	13950

Table 7.9. Demonstration Analysis: Adverse event and interval totals, all trials combined.

[1] There are 7 adverse events in Bdy-sys_3 and 9 different trials as specified by the rates in Table 7.7. At time 360, for example, there are 2 intervals in the analysis and 6 of the 9 trials will have raised rates over these 2 intervals giving $7 \times 2 \times 6 = 84$ events with raised treatment rates.

[2] There are 155 adverse events in total and 9 different trials as specified by the rates in Table 7.7. So, for example, at time 360 there are 2 intervals in the analysis giving $155 \times 9 \times 2 = 2790$ total events.

7.5 Patient Recruitment

Patients are recruited into the trial up to day 720 or until a maximum of 1000 have been recruited on each arm. The simulated numbers in the trials at each interim analysis are as follows:

Simulated Trial	Time[1]	Control	Treatment
I(a)	360	445	490
I(a)	720 - End of trial	893	962
I(b)	360	469	454
I(b)	720 - End of trial	930	927
I(c)	360	441	460
I(c)	720 - End of trial	950	903
II(a)	360	486	426
II(a)	720 - End of trial	947	876
II(b)	360	410	474
II(b)	720 - End of trial	946	931
II(c)	360	456	493
II(c)	720 - End of trial	946	951
III(a)	360	463	453
III(a)	720 - End of trial	919	904
III(b)	360	453	494
III(b)	720 - End of trial	930	955
III(c)	360	493	443
III(c)	720 - End of trial	954	898

Table 7.10. Demonstration Analysis: Simulated trial patient enrolment totals.

[1] The number of patients recruited to the trial at this particular time.

7.6 Flagging Adverse Events as Having Raised Treatment Rates

We use the posterior probability that $\theta_{bj,h}$ (the log relative risk) is greater than zero as the method for flagging adverse events with raised treatment rates (§6.8).

This is similar to the approach taken in the simulation study in Chapter 5. We have seen in the simulation study that point-mass models perform best in larger trials with higher rate differences between treatment and control. We have also seen that lowering the threshold for flagging an adverse event from 95% to 90% does not inflate the Type-I error rate. With this in mind we use a 90% posterior probability that $\theta_{bj,h}$ is greater than zero as the cut-off for flagging events for these models. For the models without the point-mass we use a 95% posterior probability. In §7.8.1 we investigate the effect of changing the significance thresholds in some of the models.

7.7 Demonstration Analysis Results

In this section we first look in detail at the outputs for a single trial (Trial II(a)) before looking at the overall results for all trials. Unlike Trial Type I, where the background rates are fixed over the course of the trial and the choice of the interval durations may not overly influence the results, for Trial II(a) the rates for Bdy-sys_3 are raised over the first two intervals before returning to the control rates thereafter (Table 7.7). In this case the choice of interval will allow us to investigate in more detail how the methods cope with changing rates over the course of the trial.

For Trial II(a) the rate increase between treatment and control is quite large for the first interval, up to day 180, before declining over the following interval, so we expect that we should see events flagged in these intervals, and the detection rate should increase and then stabilise over the course of the trial as more events accumulate.

7.7.1 Trial II(a) Incidence Event Analysis

7.7.1.1 Cumulative Adverse Event Incidence Totals

As the trial proceeds the total incidences of events in each interval increases until there are no more patients left within that time interval. Figures 7.1, 7.2, and 7.3 below show the overall incidence counts for severity 1+ and severity 3+ events for body-system Bdy-sys_3 at the different interval time points in the trial. We can see that early in the trial (Figure 7.1), even before all patients have been recruited, that the counts are raised for treatment in Bdy-sys_3 compared to control. The pattern over the whole trial is similar for both severity 1+ and severity 3+ incidence. More events occur earlier in the trial on the treatment arm than on the control

arm, but as the trial proceeds the rates appear to coalesce for later intervals. This is consistent with an earlier higher rate of event incidence on the treatment arm. In Figure 7.3 we can also see that the counts of events are decreasing for later intervals, particularly for severity 1+ events. This is because subjects are dropping out of the risk set so there are less patients at risk for later intervals. We know from Table 7.10 that more patients have been recruited to the control arm than the treatment arm, 947 as opposed to 876, and this will have an effect on the total numbers of events for the trial. In this case, even with less patients on the treatment arm, there appear to be many more events early in the trial on this arm.

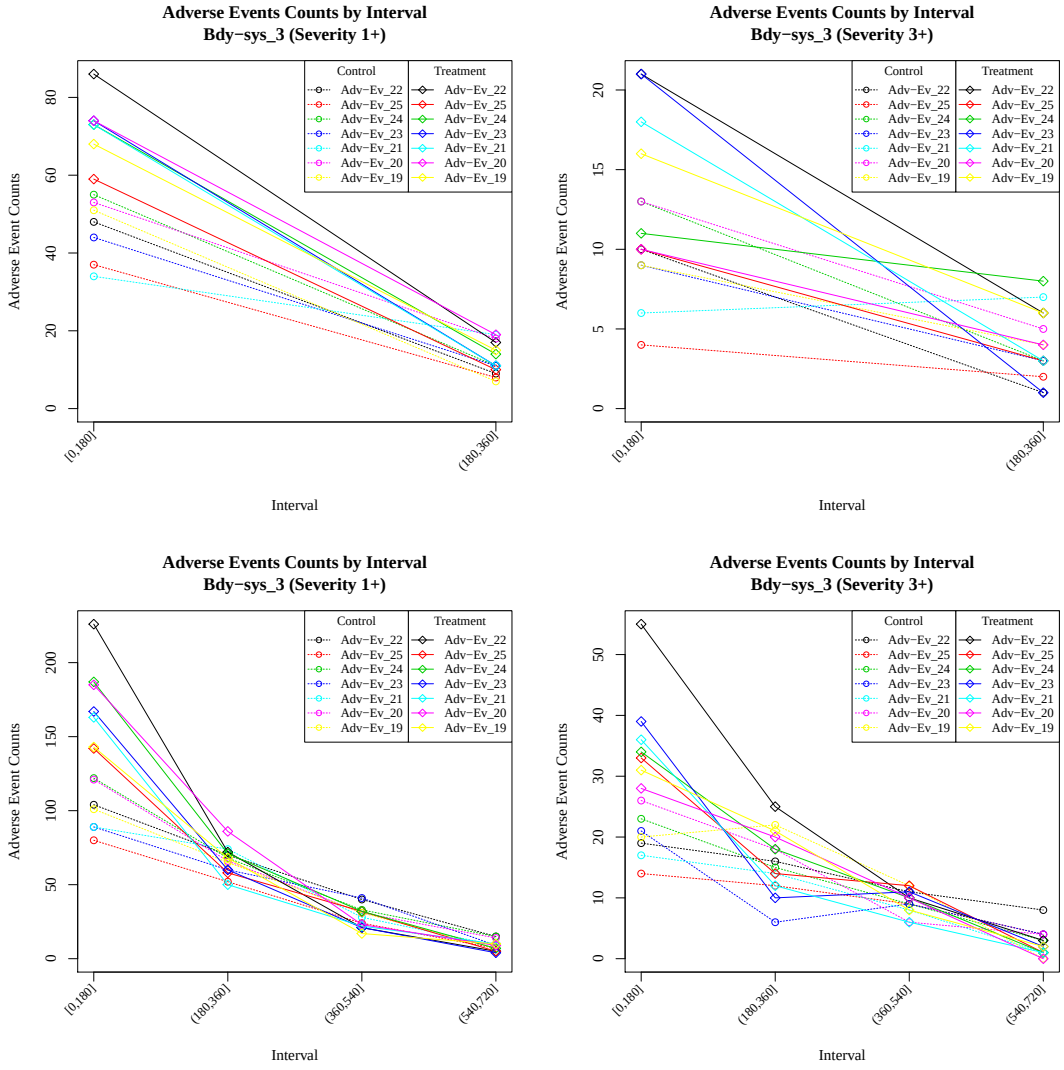


Figure 7.1. Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to day 720 of the trial. Severity 1+ events on left, severity 3+ events on the right.

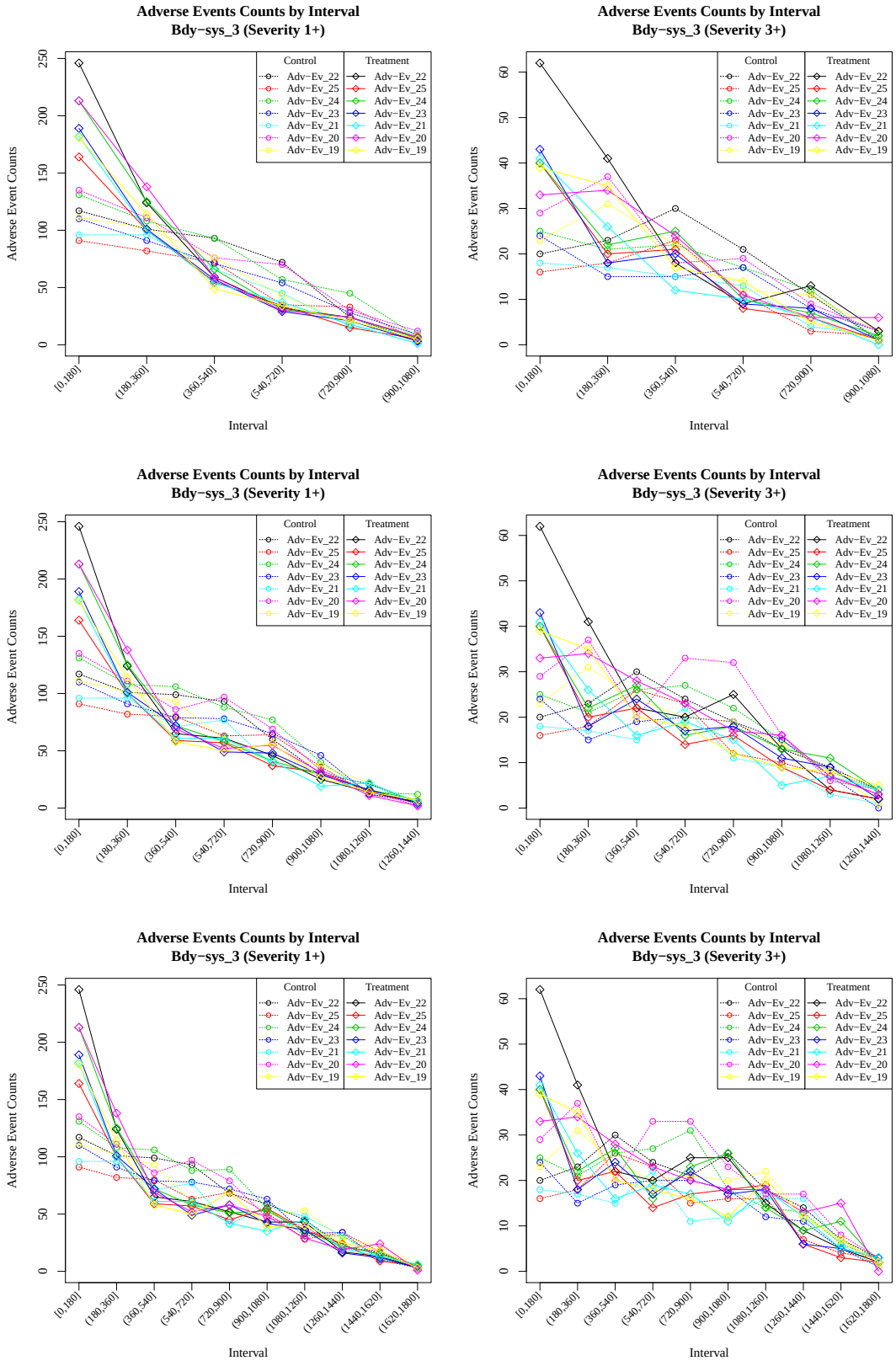


Figure 7.2. Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to day 1800 of the trial. Severity 1+ events on left, severity 3+ events on the right.

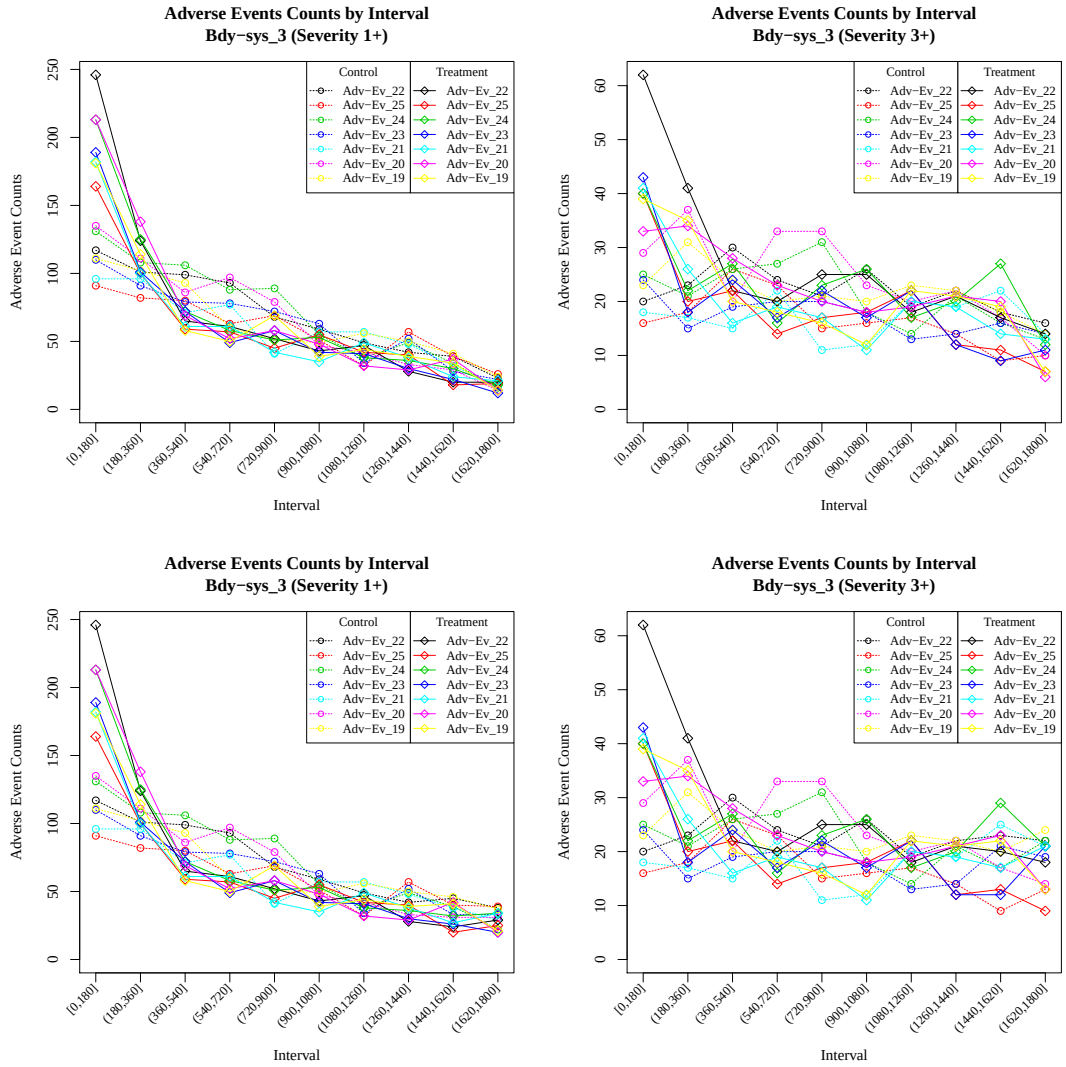


Figure 7.3. Demonstration Analysis: Adverse event cumulative incidence counts (Bdy-sys_3) by interval up to the end of trial. Severity 1+ events on left, severity 3+ events on the right.

7.7.1.2 Interim and Final Adverse Event Incidence Counts

The total incidence of adverse events that have occurred at each interim safety analysis (Table 7.2) is given for each trial arm in Table 7.11. We can see that overall there are more events on the control arm at each interim analysis. However less patients have been recruited on the treatment arm (Table 7.10). A simple analysis for all adverse events would possible submerge any potential safety signal associated with Bdy-sys_3.

Time[1]	Severity 1+		Severity 3+	
	Control	Treatment	Control	Treatment
360	7340	6661	1548	1393
720	26607	24356	5890	5522
1080	48090	44436	11449	10665
1440	65403	60565	16726	15585
1800	79324	73594	21775	20228
2160	88206	81853	25360	23602
2520	90638	84570	26464	24804

Table 7.11. Demonstration Analysis: Trial II(a) total adverse event incidence by trial arm at each interim safety analysis.

[1] The time of the safety analysis relative to the start of the trial (Table 7.2).

The incidence counts for Bdy-sys_3 only are shown in Table 7.12. We can see differences in the body-system counts for treatment and control emerging at the first interim analysis with approximately 50% more events on the treatment arm despite the lower number of patients. By the end of the trial the numbers of severity 1+ events are almost the same but for severity 3+ events a larger number have occurred on the treatment arm.

Time[1]	Severity 1+		Severity 3+	
	Control	Treatment	Control	Treatment
360	405	604	89	138
720	1465	1896	331	453
1080	2655	2996	645	767
1440	3480	3691	937	1049
1800	4114	4261	1194	1301
2160	4525	4560	1398	1500
2520	4642	4652	1458	1565

Table 7.12. Demonstration Analysis: Trial II(a) total adverse event incidence for Bdy-sys_3 by trial arm at each interim safety analysis.

[1] The time of the safety analysis relative to the start of the trial (Table 7.2).

7.7.1.3 Model Analyses: Model 1a (No Point-mass)

Running the 1a models (Table 6.1) over the data available at the interim analyses, using a 95% posterior probability that $\theta_{bj,h} > 0$ as the threshold for flagging an event (corresponding to a 5% significance level), gives the results presented in Tables 7.13, 7.14, 7.15, and 7.16.

Time[1]	**Model**[2]	**Flagged**[3]	**Correct**[4]	**Type-I**[5]	**Type-II**[6]
360	$1a_{20}$	16	9	7	5
360	$1a_{21}$	16	11	5	3
360	$1a_{30}$	16	9	7	5
360	$1a_{31}$	16	11	5	3
360	$1a_{32}$	16	9	7	5
720	$1a_{20}$	39	13	26	1
720	$1a_{21}$	30	13	17	1
720	$1a_{30}$	39	13	26	1
720	$1a_{31}$	30	13	17	1
720	$1a_{32}$	39	13	26	1
1080	$1a_{20}$	52	14	38	0
1080	$1a_{21}$	33	14	19	0
1080	$1a_{30}$	51	14	37	0
1080	$1a_{31}$	35	14	21	0
1080	$1a_{32}$	51	14	37	0

Table 7.13. Demonstration Analysis: Trial II(a) model 1a severity 1+ adverse event incidence analysis results for the first 3 safety analyses.

[1] The time of the safety analysis relative to the start of the trial (Table 7.2).
[2] The models are defined in Table 6.1.
[3] The number of adverse events flagged by the model. The total number of adverse events with raised treatment rates is give in Table 7.8.
[4] The number of flagged events which have raised treatment rates compared to control.
[5] The number of flagged events which do not have raised treatment rates compared to control.
[6] The number of events with raised treatment rates compared to control which are not flagged by the model.

Time	Model	Flagged	Correct	Type-I	Type-II
1440	$1a_{20}$	53	14	39	0
1440	$1a_{21}$	28	14	14	0
1440	$1a_{30}$	51	14	37	0
1440	$1a_{31}$	29	14	15	0
1440	$1a_{32}$	52	14	38	0
1800	$1a_{20}$	63	14	49	0
1800	$1a_{21}$	39	14	25	0
1800	$1a_{30}$	62	14	48	0
1800	$1a_{31}$	39	14	25	0
1800	$1a_{32}$	62	14	48	0
2160	$1a_{20}$	69	14	55	0
2160	$1a_{21}$	40	14	26	0
2160	$1a_{30}$	69	14	55	0
2160	$1a_{31}$	40	14	26	0
2160	$1a_{32}$	64	14	50	0
2520	$1a_{20}$	73	14	59	0
2520	$1a_{21}$	42	14	28	0
2520	$1a_{30}$	72	14	58	0
2520	$1a_{31}$	42	14	28	0
2520	$1a_{32}$	73	14	59	0

Table 7.14. Demonstration Analysis: Trial II(a) model 1a severity 1+ adverse event incidence analysis results for the final 4 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	$1a_{20}$	6	5	1	9
360	$1a_{21}$	6	6	0	8
360	$1a_{30}$	5	4	1	10
360	$1a_{31}$	6	6	0	8
360	$1a_{32}$	5	4	1	10
720	$1a_{20}$	21	7	14	7
720	$1a_{21}$	16	7	9	7
720	$1a_{30}$	21	7	14	7
720	$1a_{31}$	16	7	9	7
720	$1a_{32}$	20	7	13	7
1080	$1a_{20}$	24	8	16	6
1080	$1a_{21}$	15	7	8	7
1080	$1a_{30}$	21	7	14	7
1080	$1a_{31}$	15	7	8	7
1080	$1a_{32}$	21	7	14	7
1440	$1a_{20}$	31	8	23	6
1440	$1a_{21}$	14	7	7	7
1440	$1a_{30}$	30	7	23	7
1440	$1a_{31}$	14	7	7	7
1440	$1a_{32}$	29	7	22	7
1800	$1a_{20}$	39	8	31	6
1800	$1a_{21}$	16	7	9	7
1800	$1a_{30}$	38	7	31	7
1800	$1a_{31}$	16	7	9	7
1800	$1a_{32}$	34	7	27	7
2160	$1a_{20}$	45	8	37	6
2160	$1a_{21}$	16	7	9	7
2160	$1a_{30}$	40	7	33	7
2160	$1a_{31}$	16	7	9	7
2160	$1a_{32}$	40	7	33	7

Table 7.15. Demonstration Analysis: Trial II(a) model 1a severity 3+ adverse event incidence analysis results for the first 6 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	$1a_{20}$	43	8	35	6
2520	$1a_{21}$	15	7	8	7
2520	$1a_{30}$	41	7	34	7
2520	$1a_{31}$	15	7	8	7
2520	$1a_{32}$	40	7	33	7

Table 7.16. Demonstration Analysis: Trial II(a) model 1a severity 3+ adverse event incidence analysis results for the final safety analysis.

When taking into account both event detection and control of the Type-I error rate, the level 1 models (Table 6.1), which have a correlation between the intervals, performed best overall. There was very little to choose between the two models, although $1a_{21}$ possibly performed better than $1a_{31}$ for this data set as it has slightly fewer Type-I errors for some of the safety analyses. In terms of event detection $1a_{20}$ correctly detected most adverse events with raised treatment rates by the end of the trial but its Type-I error rate was much higher than the level 1 models.

We have seen over-estimation of the numbers of adverse events with raised treatment rates in the simulation study for the models without the point-mass, and while the number of Type-I errors made is quite high for some of the models, the $1a_{21}$ and $1a_{31}$ models control the error rates much more tightly than the other models. Overall the error rates are less than 5%. For severity 1+ events by the end of the second interim analysis period, day 720, the models had flagged the majority of events with raised treatment rates, and after day 1080 (interim safety analysis (3)) all events with raised rates had been detected. For severity 3+ events the detection rates weren't as good but again by day 1080 the models had correctly identified their maximum number of events.

From prior knowledge we know that all adverse events with raised treatment rates should occur before day 360, and that all patients have been recruited by day 720. So by day 1080 all patients have been through their first 360 days of treatment. For the independent model (1a level 0) we expect that there should no change in the number of events correctly detected from this point on, other than possibly due to sampling variation in the MCMC methods used to fit the models, and this is indeed the case. For the methods which have relationships between the intervals we can't make such statements, however in this case the numbers correctly detected remained constant after day 1080.

Figures 7.4, 7.5, 7.6, and 7.7 show the proportion of events correctly detected (out of those with raised treatment rates) and the Type-I error rates for severity 1+ and severity 3+ events respectively. We can clearly see that the error rates for the level 1 models decline after day 720, and are lower than any of the other models considered.

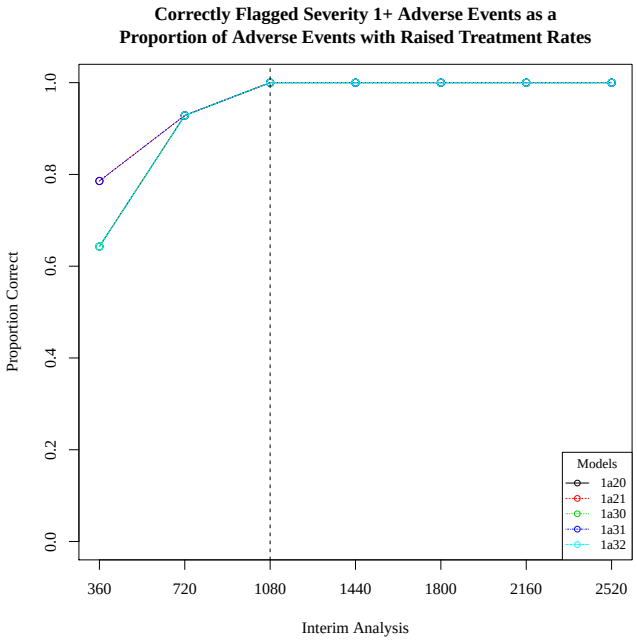


Figure 7.4. Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged.

The total number of events with raised treatment rates is given in Table 7.8.

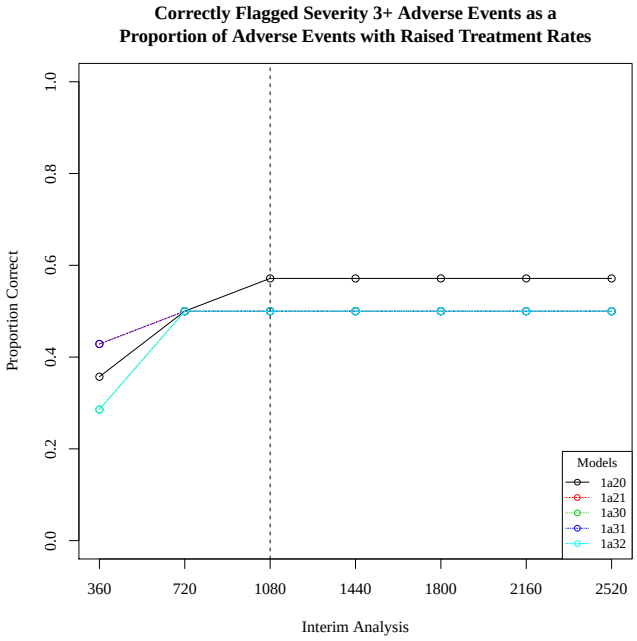


Figure 7.5. Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Proportion of severity 3+ adverse events with raised treatment rates correctly flagged.

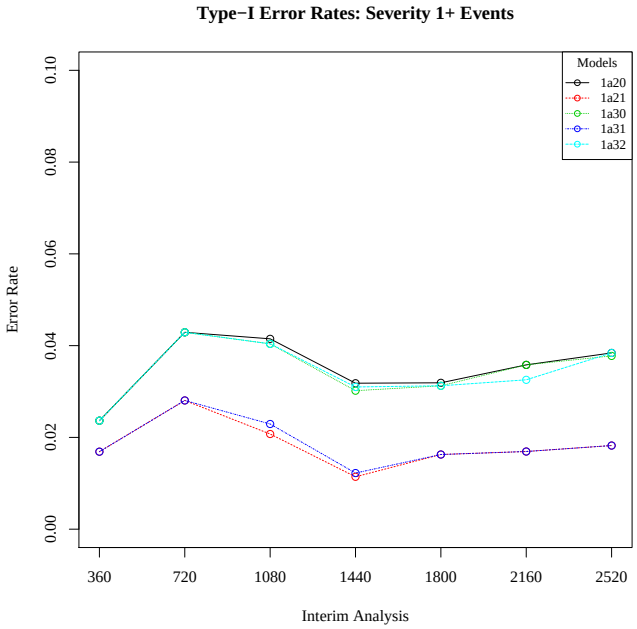


Figure 7.6. Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Type-I error rates severity 1+ events.

The total number of events is given in Table 7.8.

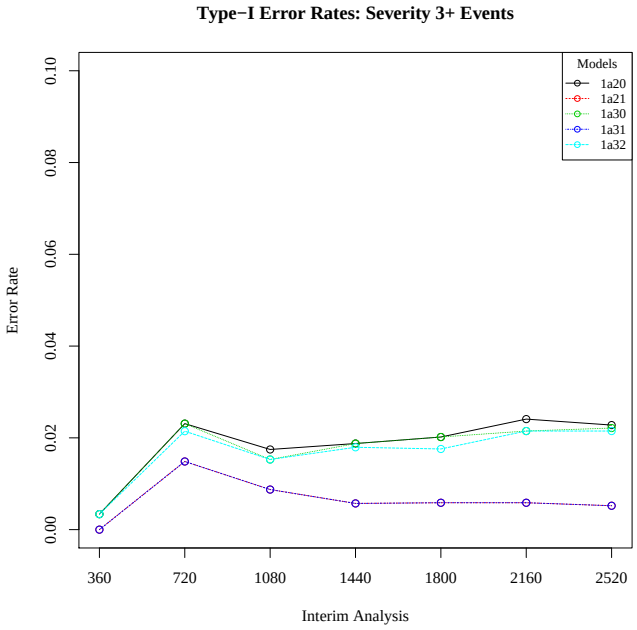


Figure 7.7. Demonstration Analysis: Trial II(a) model 1a adverse event incidence analysis. Type-I error rates severity 3+ events.

The total number of events is given in Table 7.8.

7.7.1.4 Model Analyses: Model BB (Point-mass)

Running the BB models (Tables 6.1) over the data available at the interim analyses, using a 90% posterior probability threshold for flagging an event, gives the results in Tables 7.17, 7.18, 7.19, and 7.20. The 90% threshold is lower than that used for the 1a models for the reasons given in §7.6.

Time[1]	**Model**[2]	**Flagged**[3]	**Correct**[4]	**Type-I**[5]	**Type-II**[6]
360	BB_{20}	7	7	0	7
360	BB_{21}	9	9	0	5
360	BB_{30}	7	7	0	7
360	BB_{31}	8	8	0	6
360	BB_{32}	6	6	0	8
720	BB_{20}	9	9	0	5
720	BB_{21}	10	9	1	5
720	BB_{30}	8	8	0	6
720	BB_{31}	9	9	0	5
720	BB_{32}	8	8	0	6
1080	BB_{20}	13	13	0	1
1080	BB_{21}	11	11	0	3
1080	BB_{30}	11	11	0	3
1080	BB_{31}	10	10	0	4
1080	BB_{32}	9	9	0	5

Table 7.17. Demonstration Analysis: Trial II(a) model BB severity 1+ adverse event incidence analysis results for the first 3 safety analyses.

[1] The time of the safety analysis relative to the start of the trial (Table 7.2).
[2] The models are defined in Table 6.1.
[3] The number of adverse events flagged by the model. The total number of adverse events with raised treatment rates is give in Table 7.8.
[4] The number of flagged events which have raised treatment rates compared to control.
[5] The number of flagged events which do not have raised treatment rates compared to control.
[6] The number of events with raised treatment rates compared to control which are not flagged by the model.

Time	Model	Flagged	Correct	Type-I	Type-II
1440	BB_{20}	13	13	0	1
1440	BB_{21}	11	11	0	3
1440	BB_{30}	11	11	0	3
1440	BB_{31}	10	10	0	4
1440	BB_{32}	9	9	0	5
1800	BB_{20}	13	13	0	1
1800	BB_{21}	13	11	2	3
1800	BB_{30}	11	11	0	3
1800	BB_{31}	11	10	1	4
1800	BB_{32}	9	9	0	5
2160	BB_{20}	14	13	1	1
2160	BB_{21}	10	10	0	4
2160	BB_{30}	12	11	1	3
2160	BB_{31}	10	10	0	4
2160	BB_{32}	10	9	1	5
2520	BB_{20}	13	13	0	1
2520	BB_{21}	11	10	1	4
2520	BB_{30}	11	11	0	3
2520	BB_{31}	10	10	0	4
2520	BB_{32}	9	9	0	5

Table 7.18. Demonstration Analysis: Trial II(a) model BB severity 1+ adverse event incidence analysis results for the final 4 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB_{20}	3	3	0	11
360	BB_{21}	3	3	0	11
360	BB_{30}	3	3	0	11
360	BB_{31}	2	2	0	12
360	BB_{32}	1	1	0	13
720	BB_{20}	6	6	0	8
720	BB_{21}	5	5	0	9
720	BB_{30}	4	4	0	10
720	BB_{31}	4	4	0	10
720	BB_{32}	4	4	0	10
1080	BB_{20}	6	6	0	8
1080	BB_{21}	5	5	0	9
1080	BB_{30}	5	5	0	9
1080	BB_{31}	4	4	0	10
1080	BB_{32}	4	4	0	10
1440	BB_{20}	6	6	0	8
1440	BB_{21}	5	5	0	9
1440	BB_{30}	5	5	0	9
1440	BB_{31}	4	4	0	10
1440	BB_{32}	4	4	0	10
1800	BB_{20}	6	6	0	8
1800	BB_{21}	5	5	0	9
1800	BB_{30}	6	6	0	8
1800	BB_{31}	3	3	0	11
1800	BB_{32}	4	4	0	10
2160	BB_{20}	6	6	0	8
2160	BB_{21}	4	4	0	10
2160	BB_{30}	5	5	0	9
2160	BB_{31}	3	3	0	11
2160	BB_{32}	3	3	0	11

Table 7.19. Demonstration Analysis: Trial II(a) model BB severity 3+ adverse event incidence analysis results for the first 6 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	BB_{20}	6	6	0	8
2520	BB_{21}	4	4	0	10
2520	BB_{30}	5	5	0	9
2520	BB_{31}	3	3	0	11
2520	BB_{32}	3	3	0	11

Table 7.20. Demonstration Analysis: Trial II(a) model BB severity 3+ adverse event incidence analysis results for the final safety analysis.

The two-level hierarchical model BB_{20} performed best overall taking into account both severity 1+ and severity 3+ events. There were eight Type-I errors for the severity 1+ events and none for severity 3+ events. The Type-II error rate is higher than for the corresponding 1a models. Even with the lower threshold of 90% the BB models have not performed as well as the 1a models in terms of event detection. Figures 7.8 and 7.9 show the proportions of correctly identified events.

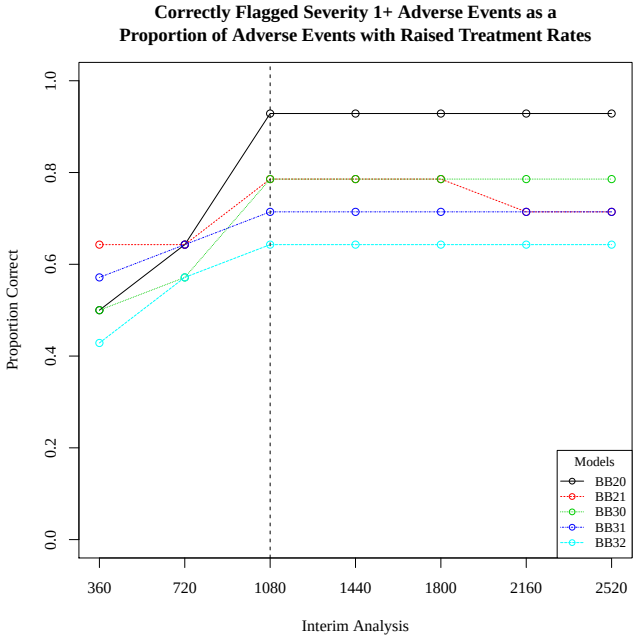


Figure 7.8. Demonstration Analysis: Trial II(a) model BB adverse event incidence analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged.

The total number of events with raised treatment rates is given in Table 7.8.

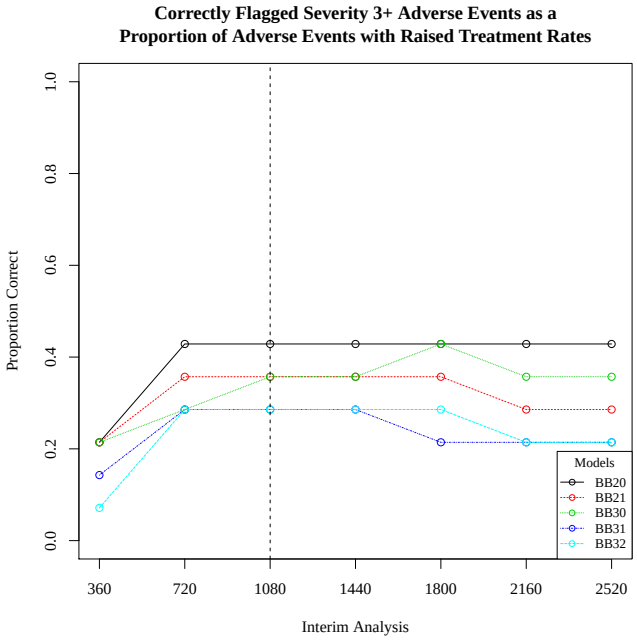


Figure 7.9. Demonstration Analysis: Trial II(a) model BB adverse event incidence analysis. Proportion of severity 3+ adverse events with raised treatment rates correctly flagged.

The total number of events with raised treatment rates is given in Table 7.8.

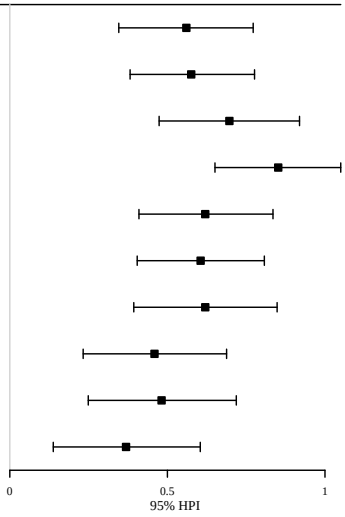
7.7.1.5 Assessing the Adverse Event Rates by Posterior Distribution

The analyses in §7.7.1.3 and §7.7.1.4 concentrated mainly on comparing the numbers of correctly detected adverse events, and the Type-I and Type-II error rates, between the different models using specific posterior probability cut-offs. In this section we look at how two of the models, $1a_{31}$ and BB_{21}, may be used to assess the adverse events for the end of trial data (day 2520).

For model $1a_{31}$ the top 10 adverse events with the highest posterior probability that $\theta_{bj,h} > 0$, together with their means and 95% credible intervals (Highest Posterior Density Intervals (HPI)), are shown in Figure 7.10. For severity 1+ events (Figure 7.10a) at the 95% cut-off all 10 events would be correctly flagged as having raised treatment rates over their corresponding intervals (Table 7.14). For severity 3+ events (Figure 7.10b) 7 out of the 10 events would be correctly flagged at the 95% level (Table 7.16). Comparing the two figures we can see that the more common severity 1+ events generally have higher posterior probabilities of being associated with treatment than the severity 3+ events.

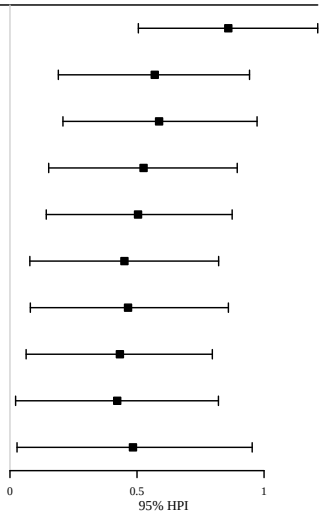
For model BB_{21} the equivalent data is shown in Figure 7.11. Here all 10 severity 1+ events would be correctly flagged at the 90% cut-off (Table 7.18). For the severity 3+ events, which are much rarer, the top 7 events all have raised treatment rates over their corresponding intervals, but in this case only 4 events would be flagged at the 90% cut-off (Table 7.20). Nevertheless there does appear to be a strong body-system effect for Bdy-sys_3. From Figure 7.11b we can see that the probability that $\theta_{bj,h} > 0$ decreases rapidly for severity 3+ events and that the credible intervals for these lower probability events are truncated at 0. This is an effect of the point-mass term in the $\theta_{bj,h}$ parameter (6.4). For adverse events with no differences in rate between treatment and control we expect that the posterior distribution of $\theta_{bj,h}$ will retain a substantial point-mass component, and that many of the values sampled during the model fit will be 0.

Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_19	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_20	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_21	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_24	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_25	1.000
180.0−360.0	Bdy−sys_3	Adv−Ev_20	1.000
180.0−360.0	Bdy−sys_3	Adv−Ev_22	1.000
180.0−360.0	Bdy−sys_3	Adv−Ev_24	0.999



(a) Severity 1+ adverse events.

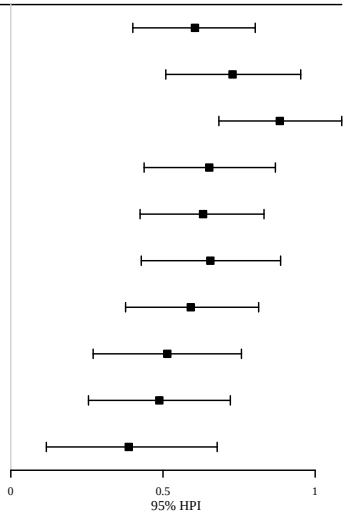
Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_21	0.999
0.0−180.0	Bdy−sys_3	Adv−Ev_25	0.999
180.0−360.0	Bdy−sys_3	Adv−Ev_22	0.997
0.0−180.0	Bdy−sys_3	Adv−Ev_23	0.997
0.0−180.0	Bdy−sys_3	Adv−Ev_19	0.992
540.0−720.0	Bdy−sys_5	Adv−Ev_35	0.990
0.0−180.0	Bdy−sys_3	Adv−Ev_24	0.990
1440.0−1620.0	Bdy−sys_3	Adv−Ev_24	0.980
1080.0−1260.0	Bdy−sys_7	Adv−Ev_57	0.980



(b) Severity 3+ adverse events.

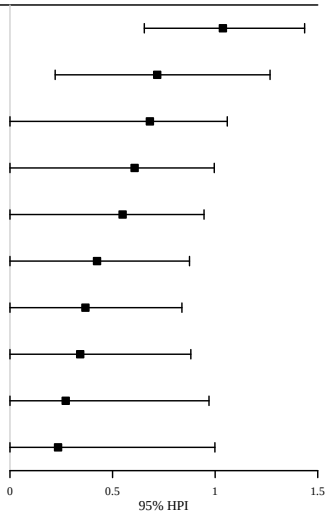
Figure 7.10. Model $1a_{31}$: Top 10 adverse events by posterior probability at the end of trial (day 2520).

Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_20	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_21	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_24	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_25	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_19	1.000
180.0−360.0	Bdy−sys_3	Adv−Ev_22	0.999
180.0−360.0	Bdy−sys_3	Adv−Ev_20	0.998
180.0−360.0	Bdy−sys_3	Adv−Ev_24	0.970



(a) Severity 1+ adverse events.

Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_25	0.965
0.0−180.0	Bdy−sys_3	Adv−Ev_21	0.959
180.0−360.0	Bdy−sys_3	Adv−Ev_22	0.918
0.0−180.0	Bdy−sys_3	Adv−Ev_23	0.879
0.0−180.0	Bdy−sys_3	Adv−Ev_19	0.748
0.0−180.0	Bdy−sys_3	Adv−Ev_24	0.687
1440.0−1620.0	Bdy−sys_3	Adv−Ev_24	0.615
540.0−720.0	Bdy−sys_5	Adv−Ev_35	0.392
1080.0−1260.0	Bdy−sys_7	Adv−Ev_57	0.340



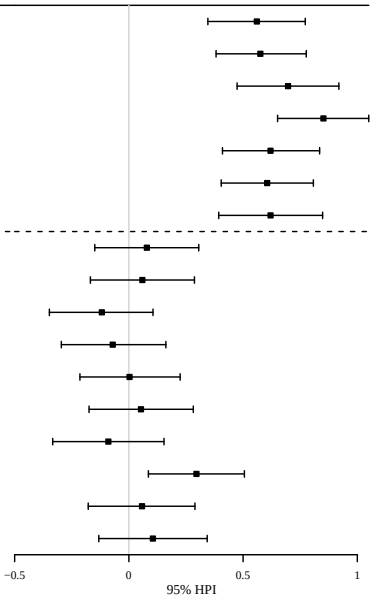
(b) Severity 3+ adverse events.

Figure 7.11. Model BB_{21}: Top 10 adverse events by posterior probability at the end of trial (day 2520).

It is also informative to compare the adverse event rates between different body-systems. Figure 7.12 shows the estimated rates from model $1a_{31}$ for Bdy-sys_3 and Bdy-sys_1 over the first trial interval. We can see (Figures 7.12a, 7.12a) that

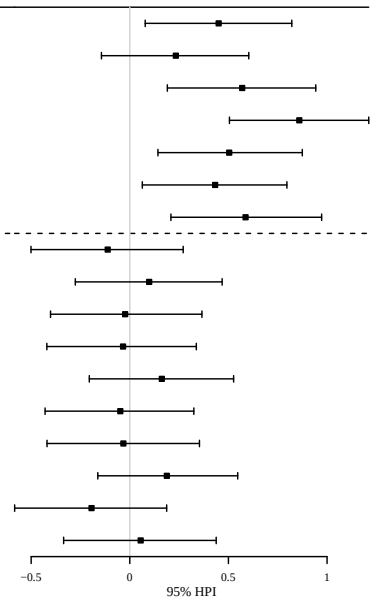
adverse events in Bdy-sys_3 have higher rates than Bdy-sys_1, even for the rarer severity 3+ events. This agrees with what we know of the underlying data model where the rates for adverse events in Bdy-sys_3 are higher for treatment than control and those for Bdy-sys_1 are the same for both treatment and control. In the equivalent plots for model BB_{21}, Figure 7.13, this difference between the rates in the different body-systems is even more clear. Here the rates for majority of the adverse events for Bdy-sys_1 remain clustered around 0, strongly indicating no treatment effect.

Interval	Body−system	Adverse Event	P(θ > 0)
0.0−180.0	Bdy−sys_3	Adv−Ev_19	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_20	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_21	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_24	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_25	1.000
0.0−180.0	Bdy−sys_1	Adv−Ev_1	0.750
0.0−180.0	Bdy−sys_1	Adv−Ev_10	0.695
0.0−180.0	Bdy−sys_1	Adv−Ev_2	0.154
0.0−180.0	Bdy−sys_1	Adv−Ev_3	0.272
0.0−180.0	Bdy−sys_1	Adv−Ev_4	0.511
0.0−180.0	Bdy−sys_1	Adv−Ev_5	0.676
0.0−180.0	Bdy−sys_1	Adv−Ev_6	0.235
0.0−180.0	Bdy−sys_1	Adv−Ev_7	0.997
0.0−180.0	Bdy−sys_1	Adv−Ev_8	0.686
0.0−180.0	Bdy−sys_1	Adv−Ev_9	0.810



(a) Severity 1+ adverse events.

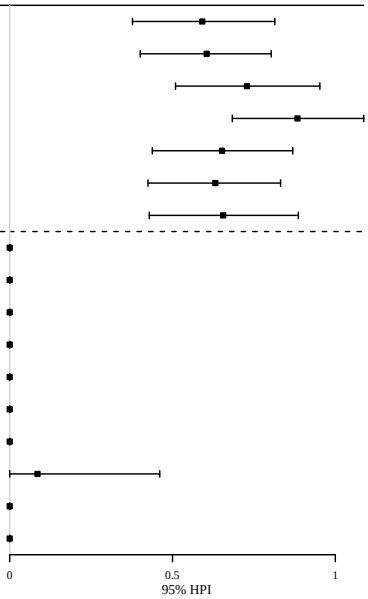
Interval	Body−system	Adverse Event	P(θ > 0)
0.0−180.0	Bdy−sys_3	Adv−Ev_19	0.992
0.0−180.0	Bdy−sys_3	Adv−Ev_20	0.892
0.0−180.0	Bdy−sys_3	Adv−Ev_21	0.999
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	0.997
0.0−180.0	Bdy−sys_3	Adv−Ev_24	0.990
0.0−180.0	Bdy−sys_3	Adv−Ev_25	0.999
0.0−180.0	Bdy−sys_1	Adv−Ev_1	0.285
0.0−180.0	Bdy−sys_1	Adv−Ev_10	0.700
0.0−180.0	Bdy−sys_1	Adv−Ev_2	0.454
0.0−180.0	Bdy−sys_1	Adv−Ev_3	0.433
0.0−180.0	Bdy−sys_1	Adv−Ev_4	0.812
0.0−180.0	Bdy−sys_1	Adv−Ev_5	0.406
0.0−180.0	Bdy−sys_1	Adv−Ev_6	0.437
0.0−180.0	Bdy−sys_1	Adv−Ev_7	0.851
0.0−180.0	Bdy−sys_1	Adv−Ev_8	0.162
0.0−180.0	Bdy−sys_1	Adv−Ev_9	0.611



(b) Severity 3+ adverse events.

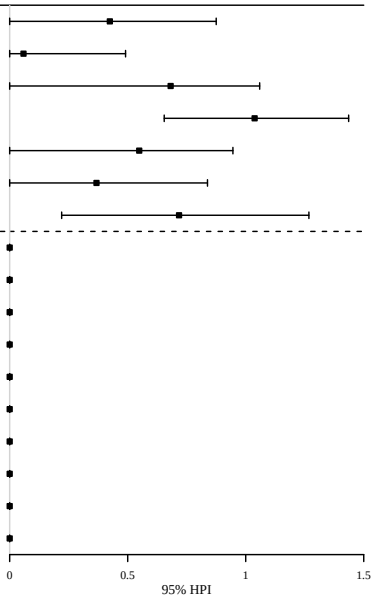
Figure 7.12. Model $1a_{31}$: Bdy-sys3 and Bdy-sys_1 end of trial rates.

Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_19	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_20	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_21	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_24	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_25	1.000
0.0−180.0	Bdy−sys_1	Adv−Ev_1	0.007
0.0−180.0	Bdy−sys_1	Adv−Ev_10	0.005
0.0−180.0	Bdy−sys_1	Adv−Ev_2	0.001
0.0−180.0	Bdy−sys_1	Adv−Ev_3	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_4	0.003
0.0−180.0	Bdy−sys_1	Adv−Ev_5	0.005
0.0−180.0	Bdy−sys_1	Adv−Ev_6	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_7	0.236
0.0−180.0	Bdy−sys_1	Adv−Ev_8	0.005
0.0−180.0	Bdy−sys_1	Adv−Ev_9	0.008



(a) Severity 1+ adverse events.

Interval	Body−system	Adverse Event	$P(\theta > 0)$
0.0−180.0	Bdy−sys_3	Adv−Ev_19	0.748
0.0−180.0	Bdy−sys_3	Adv−Ev_20	0.184
0.0−180.0	Bdy−sys_3	Adv−Ev_21	0.959
0.0−180.0	Bdy−sys_3	Adv−Ev_22	1.000
0.0−180.0	Bdy−sys_3	Adv−Ev_23	0.879
0.0−180.0	Bdy−sys_3	Adv−Ev_24	0.687
0.0−180.0	Bdy−sys_3	Adv−Ev_25	0.965
0.0−180.0	Bdy−sys_1	Adv−Ev_1	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_10	0.005
0.0−180.0	Bdy−sys_1	Adv−Ev_2	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_3	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_4	0.009
0.0−180.0	Bdy−sys_1	Adv−Ev_5	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_6	0.002
0.0−180.0	Bdy−sys_1	Adv−Ev_7	0.010
0.0−180.0	Bdy−sys_1	Adv−Ev_8	0.001
0.0−180.0	Bdy−sys_1	Adv−Ev_9	0.004



(b) Severity 3+ adverse events.

Figure 7.13. Model BB_{21}: Bdy-sys3 and Bdy-sys_1 end of trial rates.

7.7.2 Trial II(a) Total Adverse Event Analysis

Here we include all the events that have occurred over course of the trial. We expect the results to be very similar to the incidence event analysis as the underlying occurrence rates of the events are the same.

7.7.2.1 Cumulative Adverse Event Totals

Figures 7.14, 7.15, and 7.16 show the evolution of the overall counts for severity 1+ and severity 3+ events for body-system Bdy-sys_3, with higher rate of adverse events on the treatment arm early in the trial. These graphs are very similar to those in Figures 7.1-7.3, but we can see that by the end of the trial overall more events have been accumulated (Figures 7.3, 7.16).

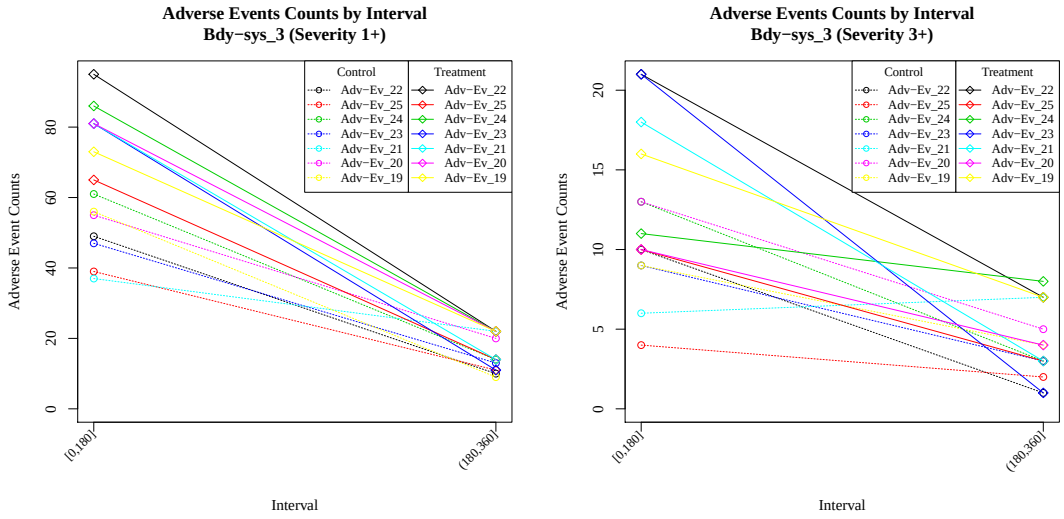


Figure 7.14. Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval up to day 360. Severity 1+ events on left, severity 3+ events on the right.

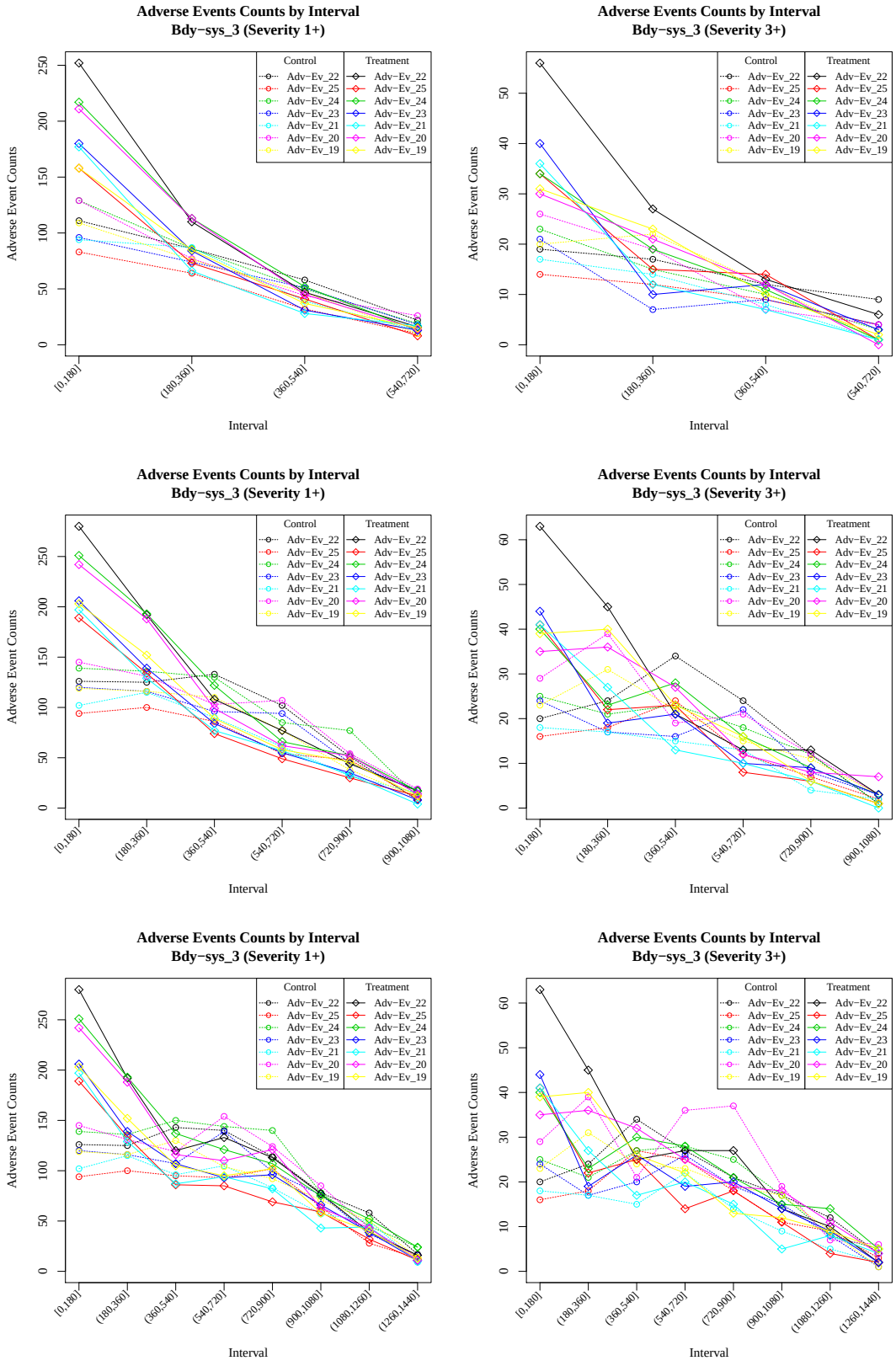


Figure 7.15. Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval up to day 1440. Severity 1+ events on left, severity 3+ events on the right.

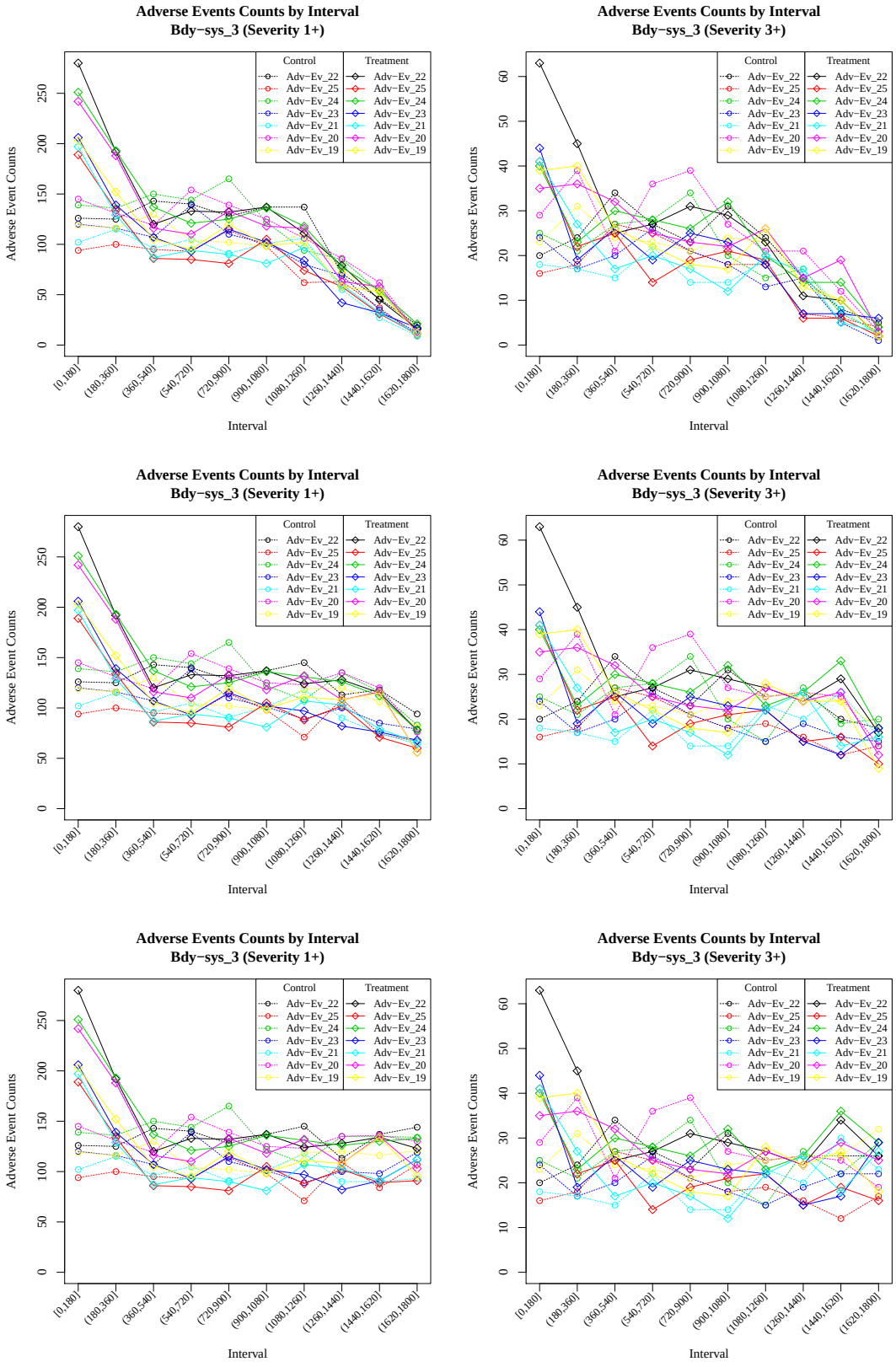


Figure 7.16. Demonstration Analysis: Total adverse event counts (Bdy-sys_3) by interval to end of trial. Severity 1+ events on left, severity 3+ events on the right.

7.7.2.2 Interim and Final Analyses Adverse Event Counts

The total counts for all events are given in Table 7.21. As for the incidence case we have more events on the control arm than the treatment arm.

Time[1]	Severity 1+		Severity 3+	
	Control	Treatment	Control	Treatment
360	7853	7176	1574	1421
720	30576	28180	6065	5700
1080	60035	55787	12002	11183
1440	89778	83195	17958	16666
1800	119051	110627	23827	22120
2160	140768	130993	28139	26117
2520	147489	138582	29481	27606

Table 7.21. Demonstration Analysis: Trial II(a) total adverse events by trial arm at each interim safety analysis.

[1] The time of the safety analysis relative to the start of the trial (Table 7.2).

Looking at the counts for Bdy-sys_3 in Table 7.22 we can again see differences in the body-system counts for treatment and control emerging quite early in the trial.

Time	Severity 1+		Severity 3+	
	Control	Treatment	Control	Treatment
360	443	689	89	140
720	1740	2374	339	481
1080	3448	4148	681	831
1440	5084	5700	1015	1162
1800	6639	7302	1317	1493
2160	7847	8451	1566	1745
2520	8253	8879	1640	1841

Table 7.22. Demonstration Analysis: Trial II(a) total adverse events by trial arm at each interim safety analysis for Bdy-sys_3.

7.7.2.3 Model Analyses: Model 1a (No Point-mass)

The results of running the models over the data available at the interim analyses gives the following results using a nominal 95% cut-off for flagging an event are given in Tables 7.23, 7.24, 7.25, and 7.26.

Time	Model	Flagged	Correct	Type-I	Type-II
360	$1a_{20}$	16	10	6	4
360	$1a_{21}$	15	10	5	4
360	$1a_{30}$	17	10	7	4
360	$1a_{31}$	16	10	6	4
360	$1a_{32}$	17	10	7	4
720	$1a_{20}$	39	13	26	1
720	$1a_{21}$	29	13	16	1
720	$1a_{30}$	39	13	26	1
720	$1a_{31}$	27	13	14	1
720	$1a_{32}$	41	13	28	1
1080	$1a_{20}$	45	14	31	0
1080	$1a_{21}$	35	13	22	1
1080	$1a_{30}$	48	13	35	1
1080	$1a_{31}$	33	13	20	1
1080	$1a_{32}$	44	13	31	1
1440	$1a_{20}$	59	14	45	0
1440	$1a_{21}$	31	13	18	1
1440	$1a_{30}$	56	13	43	1
1440	$1a_{31}$	31	13	18	1
1440	$1a_{32}$	55	13	42	1
1800	$1a_{20}$	72	14	58	0
1800	$1a_{21}$	38	13	25	1
1800	$1a_{30}$	77	14	63	0
1800	$1a_{31}$	39	13	26	1
1800	$1a_{32}$	75	13	62	1

Table 7.23. Demonstration Analysis: Trial II(a) model 1a severity 1+ total adverse event analysis results for the first 5 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2160	$1a_{20}$	75	14	61	0
2160	$1a_{21}$	44	13	31	1
2160	$1a_{30}$	76	14	62	0
2160	$1a_{31}$	43	13	30	1
2160	$1a_{32}$	74	13	61	1
2520	$1a_{20}$	82	14	68	0
2520	$1a_{21}$	46	13	33	1
2520	$1a_{30}$	82	14	68	0
2520	$1a_{31}$	47	13	34	1
2520	$1a_{32}$	82	13	69	1

Table 7.24. Demonstration Analysis: Trial II(a) model 1a severity 1+ total adverse event analysis results for the final 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	$1a_{20}$	7	5	2	9
360	$1a_{21}$	7	7	0	7
360	$1a_{30}$	5	4	1	10
360	$1a_{31}$	6	6	0	8
360	$1a_{32}$	5	4	1	10
720	$1a_{20}$	22	8	14	6
720	$1a_{21}$	17	7	10	7
720	$1a_{30}$	20	7	13	7
720	$1a_{31}$	17	7	10	7
720	$1a_{32}$	19	7	12	7
1080	$1a_{20}$	27	10	17	4
1080	$1a_{21}$	13	8	5	6
1080	$1a_{30}$	23	9	14	5
1080	$1a_{31}$	13	8	5	6
1080	$1a_{32}$	20	7	13	7

Table 7.25. Demonstration Analysis: Trial II(a) model 1a severity 3+ total adverse event analysis results for the first 3 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1440	$1a_{20}$	29	10	19	4
1440	$1a_{21}$	14	8	6	6
1440	$1a_{30}$	29	9	20	5
1440	$1a_{31}$	14	8	6	6
1440	$1a_{32}$	26	7	19	7
1800	$1a_{20}$	39	10	29	4
1800	$1a_{21}$	18	8	10	6
1800	$1a_{30}$	35	8	27	6
1800	$1a_{31}$	18	8	10	6
1800	$1a_{32}$	33	7	26	7
2160	$1a_{20}$	39	9	30	5
2160	$1a_{21}$	21	8	13	6
2160	$1a_{30}$	41	10	31	4
2160	$1a_{31}$	21	8	13	6
2160	$1a_{32}$	34	7	27	7
2520	$1a_{20}$	41	10	31	4
2520	$1a_{21}$	20	8	12	6
2520	$1a_{30}$	39	8	31	6
2520	$1a_{31}$	20	8	12	6
2520	$1a_{32}$	34	7	27	7

Table 7.26. Demonstration Analysis: Trial II(a) model 1a severity 3+ total adverse event analysis results for the final 4 safety analyses.

The results here are very similar to the incidence data results. The level 1 models with the stronger dependence ($1a_{21}$, $1a_{31}$) performed the best overall in all cases. Model $1a_{20}$ detected all severity 1+ events and the most severity 3+ which had raised treatment rates, but at the expense of much higher Type-I errors. The extra dependence in the level 1 models has had the effect of controlling both the Type-I and Type-II error rates compared to the other models. More adverse events with raised treatment rates were detected overall than in the event incidence case. This is expected as we have more event data. For severity 1+ events we have detected nearly all events with raised treatment rates by day 720, even though these events were still occurring. Figures 7.17 and 7.18 show the proportion of correctly flagged adverse events and Type-I errors respectively.

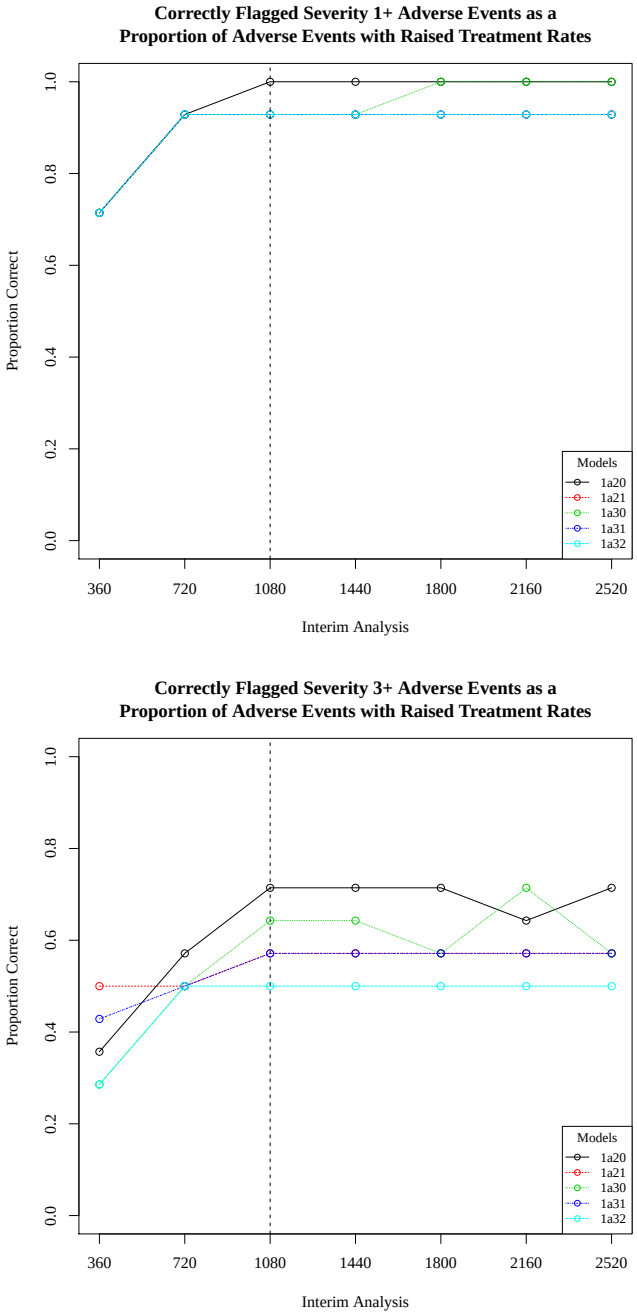


Figure 7.17. Demonstration Analysis: Trial II(a) model 1a total adverse event analysis. Proportion of adverse events with raised treatment rates correctly flagged (severity 1+ on the top, severity 3+ on the bottom).

The total number of events with raised treatment rates is given in Table 7.8.

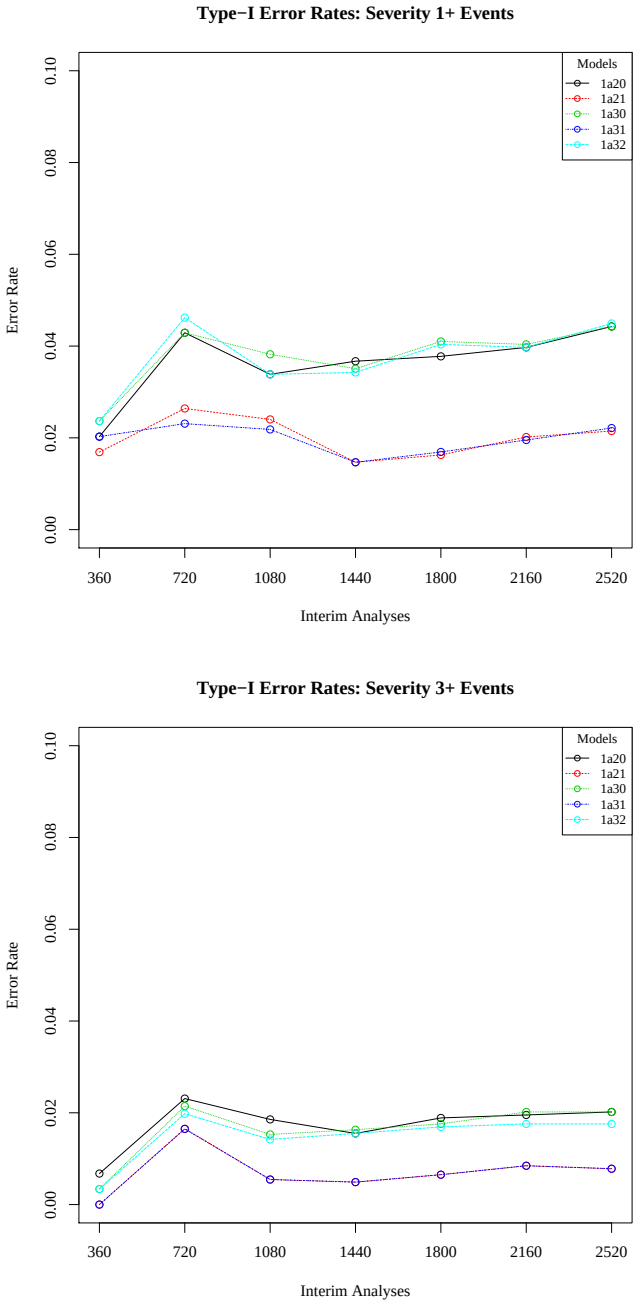


Figure 7.18. Demonstration Analysis: Trial II(a) model 1a total adverse events analysis. Type-I error rates (severity 1+ on the top, severity 3+ on the bottom)

The total number of events with raised treatment rates is given in Table 7.8.

7.7.2.4 Model Analyses: Model BB (Point-mass)

Running the models over the data available at the interim analyses gives the results (using a nominal 90% threshold for flagging an event) in Tables 7.27, 7.28, 7.29, and 7.30.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB_{20}	7	7	0	7
360	BB_{21}	10	10	0	4
360	BB_{30}	6	6	0	8
360	BB_{31}	9	9	0	5
360	BB_{32}	6	6	0	8
720	BB_{20}	10	10	0	4
720	BB_{21}	10	10	0	4
720	BB_{30}	9	9	0	5
720	BB_{31}	10	10	0	4
720	BB_{32}	8	8	0	6
1080	BB_{20}	12	12	0	2
1080	BB_{21}	12	12	0	2
1080	BB_{30}	12	12	0	2
1080	BB_{31}	11	11	0	3
1080	BB_{32}	10	10	0	4
1440	BB_{20}	13	13	0	1
1440	BB_{21}	11	11	0	3
1440	BB_{30}	12	12	0	2
1440	BB_{31}	10	10	0	4
1440	BB_{32}	10	10	0	4
1800	BB_{20}	12	12	0	2
1800	BB_{21}	11	11	0	3
1800	BB_{30}	12	12	0	2
1800	BB_{31}	10	10	0	4
1800	BB_{32}	10	10	0	4

Table 7.27. Demonstration Analysis: Trial II(a) model BB severity 1+ total adverse event analysis results for the first 5 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2160	BB_{20}	12	12	0	2
2160	BB_{21}	10	10	0	4
2160	BB_{30}	12	12	0	2
2160	BB_{31}	10	10	0	4
2160	BB_{32}	10	10	0	4
2520	BB_{20}	12	12	0	2
2520	BB_{21}	10	10	0	4
2520	BB_{30}	12	12	0	2
2520	BB_{31}	10	10	0	4
2520	BB_{32}	10	10	0	4

Table 7.28. Demonstration Analysis: Trial II(a) model BB severity 1+ total adverse event analysis results for the final 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB_{20}	3	3	0	11
360	BB_{21}	3	3	0	11
360	BB_{30}	2	2	0	12
360	BB_{31}	2	2	0	12
360	BB_{32}	0	0	0	14
720	BB_{20}	6	6	0	8
720	BB_{21}	5	5	0	9
720	BB_{30}	4	4	0	10
720	BB_{31}	4	4	0	10
720	BB_{32}	4	4	0	10
1080	BB_{20}	6	6	0	8
1080	BB_{21}	5	5	0	9
1080	BB_{30}	6	6	0	8
1080	BB_{31}	5	5	0	9
1080	BB_{32}	4	4	0	10

Table 7.29. Demonstration Analysis: Trial II(a) model BB severity 3+ total adverse event analysis results for the first 3 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1440	BB_{20}	6	6	0	8
1440	BB_{21}	5	5	0	9
1440	BB_{30}	5	5	0	9
1440	BB_{31}	4	4	0	10
1440	BB_{32}	3	3	0	11
1800	BB_{20}	6	6	0	8
1800	BB_{21}	5	5	0	9
1800	BB_{30}	5	5	0	9
1800	BB_{31}	4	4	0	10
1800	BB_{32}	4	4	0	10
2160	BB_{20}	6	6	0	8
2160	BB_{21}	5	5	0	9
2160	BB_{30}	5	5	0	9
2160	BB_{31}	4	4	0	10
2160	BB_{32}	4	4	0	10
2520	BB_{20}	6	6	0	8
2520	BB_{21}	5	5	0	9
2520	BB_{30}	6	6	0	8
2520	BB_{31}	4	4	0	10
2520	BB_{32}	3	3	0	11

Table 7.30. Demonstration Analysis: Trial II(a) model BB severity 3+ total adverse event analysis results for the final 4 safety analyses.

The extra events included in the data have generally allowed the models to perform better than the corresponding incidence models and, in terms of event detection for severity 1+ events, they perform well compared to their 1a equivalents. Their overall Type-I error control is much better than for 1a models as expected. This is shown in Figure 7.19.

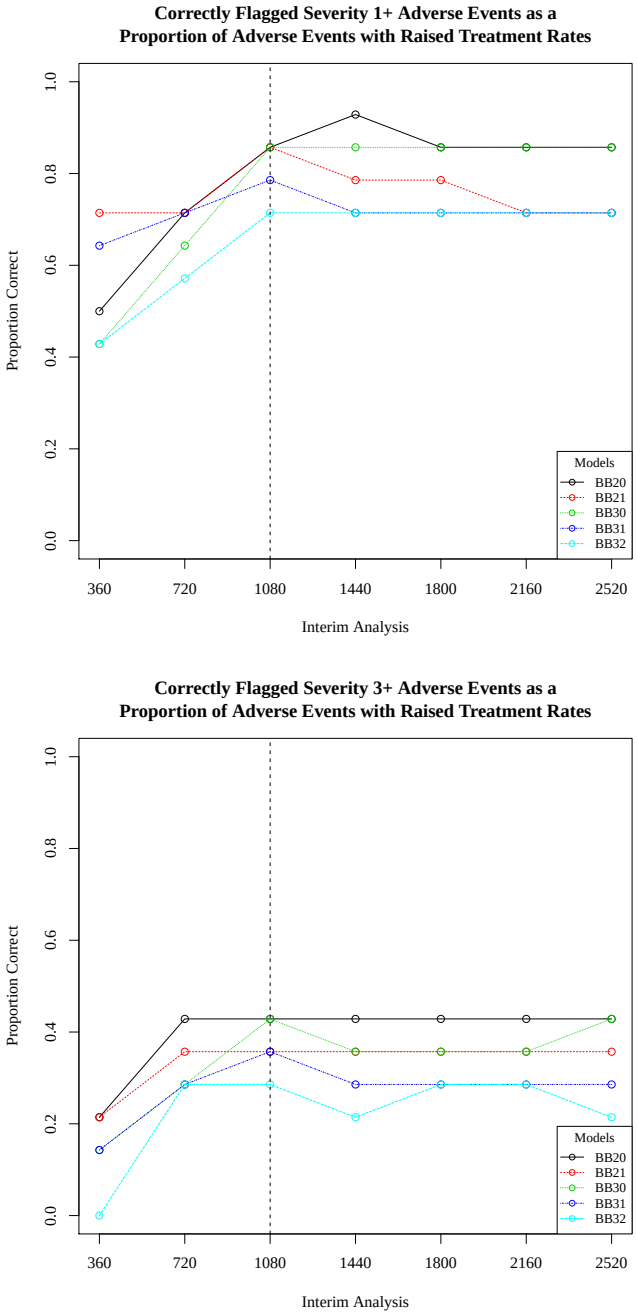


Figure 7.19. Demonstration Analysis: Trial II(a) model BB total adverse event analysis. Proportion of severity 1+ adverse events with raised treatment rates correctly flagged (severity 1+ on the top, severity 3+ on the bottom).

7.7.2.5 All Events versus Event Incidence Analysis

The results of the analysis when including all events are very similar to including event incidence only, particularly for the 1a (no point-mass) models. For the BB (point-mass) models the extra events have improved the performance of some of

the models, particularly for severity 1+ events. We have seem similar behaviour for point-mass models in the simulations study in Chapter 5, where the models perform best for larger trials and higher event rates. The adverse events may also be assessed as in §7.7.1.5 for the incidence data. In this case the outputs are very similar to those shown in Figures 7.10 - 7.13.

7.7.2.6 Model Parameter Estimation

The posterior distributions of the γ and θ parameters may be used to find point-estimates of the underlying model parameters and the underlying rate function. We look at the $1a_{21}$ and BB_{21} models and, as in §5.5.1, we use the posterior mean to estimate the underlying parameters. Figures 7.20 and 7.21 show the estimated underlying control and treatment rates for Bdy-sys_3 at the final analysis (day 2520) for the models. The overall trial background rates are given in Table 7.6 for the control arm. Table 7.7 gives the increase in rate for treatment.

Figure 7.20 shows the parameter estimates for model $1a_{21}$. The point estimates are able to capture the underlying rates quite well, even for severity 3+ events, which have much lower rates of occurrence.

For BB_{21} the estimates are shown in Figure 7.21. Here we can see that for the severity 1+ events the parameter estimates are quite good, particularly for the first interval. For the second interval, 180.0-360.0, there is not as clear a differentiation between treatment and control as there is in Fig 7.20. For severity 3+ events, which are rarer, the model is less able to distinguish differences between treatment and control, particularly for the second interval where the treatment rates are generally underestimated. We have seen similar behaviour for models with a point-mass in §5.5.1.4.

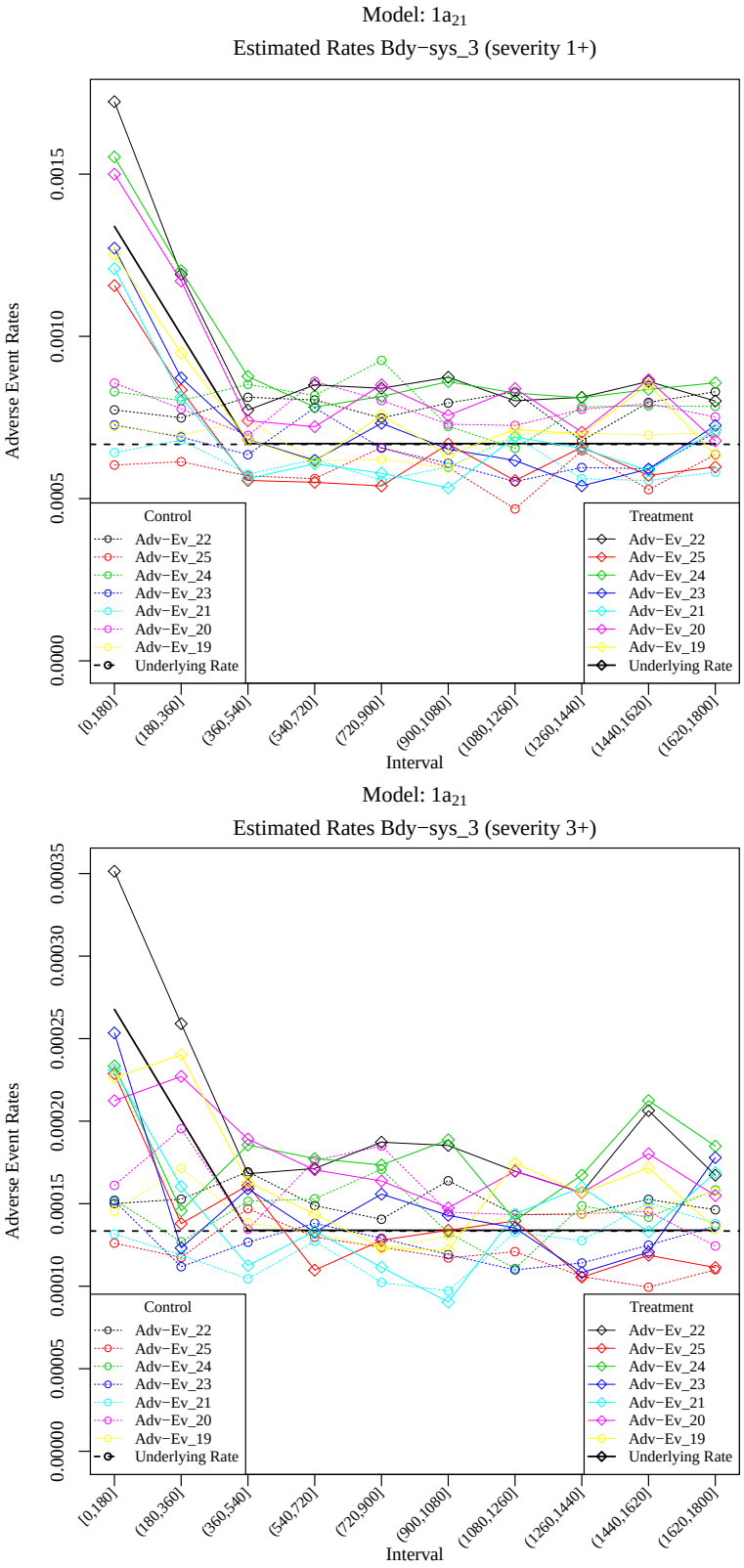


Figure 7.20. Demonstration Analysis: Day 2520 underlying parameter estimates model $1a_{21}$ (Bdy-sys_3) (severity 1+ on top, severity 3+ on bottom).

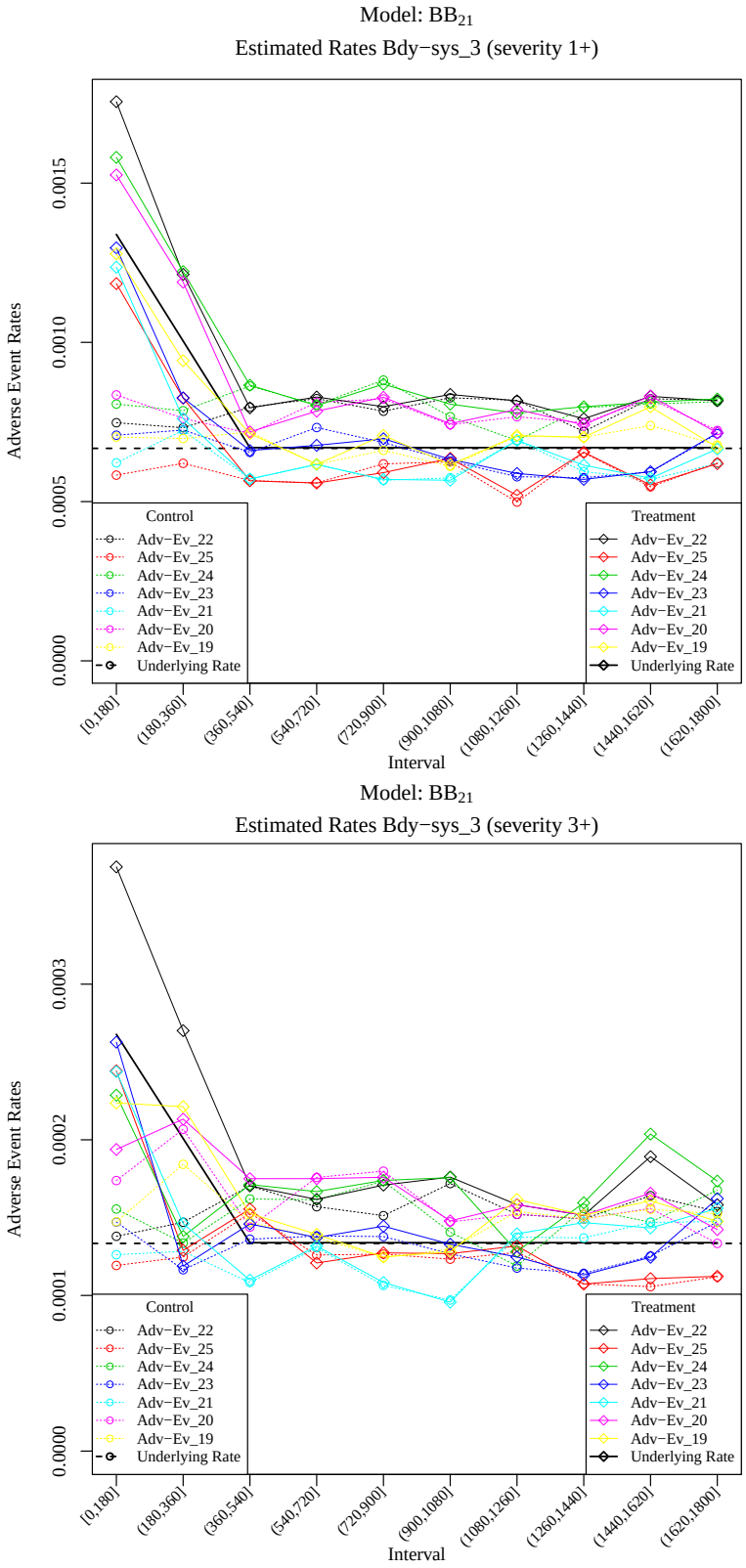


Figure 7.21. Demonstration Analysis: Day 2520 underlying parameter estimates model BB_{21} (Bdy-sys_3) (severity 1+ on top, severity 3+ on bottom).

7.7.3 Overall Results

The results for the other trial types (I, III) are in line with for Trial II(a) discussed above. Rather than look at each individual trial, in this section we present the results for all trials combined. We note that while the rates differences between treatment and control for Trial II(a) were relatively high, for smaller rate increases the models' performances will not be as good. Further, for Trial III, where the treatment rates increase over the last two intervals of the trial only, we should expect to see poorer performance for incidence data compared to Trial II, even when the rates are similar. This is because patients will leave the risk sets for both treatment and control at fixed rates (based on the background trial rate (Table 7.6)) up until the final two intervals, so there will be less patients to experience the raised rates as the trial progresses.

7.7.3.1 Incidence Event Analysis

The outputs of the 1a models over all the trials in terms of event detection and error rates, using a 95% posterior probability cut-off for flagging an event, are given in Tables 7.31, 7.32, 7.33, and 7.34.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	99	39	60	45
360	1a21	95	42	53	42
360	1a30	98	37	61	47
360	1a31	93	40	53	44
360	1a32	95	37	58	47
720	1a20	217	67	150	59
720	1a21	167	75	92	51
720	1a30	213	64	149	62
720	1a31	168	75	93	51
720	1a32	208	64	144	62

Table 7.31. Demonstration Analysis: All trials model 1a severity 1+ adverse event incidence results for the first 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1080	1a20	334	93	241	75
1080	1a21	237	109	128	59
1080	1a30	322	91	231	77
1080	1a31	236	107	129	61
1080	1a32	313	87	226	81
1440	1a20	413	127	286	83
1440	1a21	276	138	138	72
1440	1a30	405	121	284	89
1440	1a31	276	137	139	73
1440	1a32	386	116	270	94
1800	1a20	494	151	343	143
1800	1a21	330	169	161	125
1800	1a30	483	145	338	149
1800	1a31	327	168	159	126
1800	1a32	474	141	333	153
2160	1a20	578	180	398	114
2160	1a21	364	188	176	106
2160	1a30	579	177	402	117
2160	1a31	363	188	175	106
2160	1a32	562	173	389	121
2520	1a20	589	181	408	113
2520	1a21	369	191	178	103
2520	1a30	589	181	408	113
2520	1a31	374	191	183	103
2520	1a32	575	176	399	118

Table 7.32. Demonstration Analysis: All trials model 1a severity 1+ adverse event incidence results for the final 5 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	42	13	29	71
360	1a21	41	19	22	65
360	1a30	35	11	24	73
360	1a31	35	15	20	69
360	1a32	32	10	22	74
720	1a20	131	38	93	88
720	1a21	92	44	48	82
720	1a30	115	34	81	92
720	1a31	91	42	49	84
720	1a32	109	32	77	94
1080	1a20	207	63	144	105
1080	1a21	140	67	73	101
1080	1a30	191	57	134	111
1080	1a31	138	66	72	102
1080	1a32	169	46	123	122
1440	1a20	261	73	188	137
1440	1a21	164	92	72	118
1440	1a30	241	66	175	144
1440	1a31	160	88	72	122
1440	1a32	225	59	166	151
1800	1a20	326	98	228	196
1800	1a21	192	113	79	181
1800	1a30	312	84	228	210
1800	1a31	187	111	76	183
1800	1a32	277	72	205	222
2160	1a20	382	117	265	177
2160	1a21	213	136	77	158
2160	1a30	368	108	260	186
2160	1a31	215	136	79	158
2160	1a32	333	90	243	204

Table 7.33. Demonstration Analysis: All trials model 1a severity 3+ adverse event incidence results for the first 6 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	404	128	276	166
2520	1a21	228	144	84	150
2520	1a30	397	120	277	174
2520	1a31	226	143	83	151
2520	1a32	362	103	259	191

Table 7.34. Demonstration Analysis: All trials model 1a severity 3+ adverse event incidence results for the final safety analysis.

The models with the stronger dependence ($1a_{21}$, $1a_{31}$) performed the best overall. The extra dependence in these models has had the effect of controlling both the Type-I and Type-II error rates compared to the other models, which have either no dependence ($1a_{20}$, $1a_{30}$), or weaker dependence ($1a_{32}$). There is not much difference in performance between the two-level and three-level hierarchies with $1a_{21}$ possibly performing slightly better for the data considered.

The BB models were run using a 90% cut-off and the results are given in Tables 7.35, 7.36, 7.37, and 7.38.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB20	21	21	0	63
360	BB21	24	24	0	60
360	BB30	14	14	0	70
360	BB31	20	20	0	64
360	BB32	13	13	0	71
720	BB20	46	45	1	81
720	BB21	64	63	1	63
720	BB30	42	42	0	84
720	BB31	50	50	0	76
720	BB32	40	40	0	86

Table 7.35. Demonstration Analysis: All trials model BB severity 1+ adverse event incidence results for the first 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1080	BB20	71	68	3	100
1080	BB21	94	94	0	74
1080	BB30	55	52	3	116
1080	BB31	76	76	0	92
1080	BB32	45	44	1	124
1440	BB20	91	90	1	120
1440	BB21	121	121	0	89
1440	BB30	78	77	1	133
1440	BB31	108	108	0	102
1440	BB32	70	69	1	141
1800	BB20	104	103	1	191
1800	BB21	151	149	2	145
1800	BB30	85	84	1	210
1800	BB31	134	133	1	161
1800	BB32	77	76	1	218
2160	BB20	137	135	2	159
2160	BB21	156	156	0	138
2160	BB30	104	102	2	192
2160	BB31	142	142	0	152
2160	BB32	89	87	2	207
2520	BB20	135	134	1	160
2520	BB21	153	152	1	142
2520	BB30	107	106	1	188
2520	BB31	143	143	0	151
2520	BB32	96	95	1	199

Table 7.36. Demonstration Analysis: All trials model BB severity 1+ adverse event incidence results for the final 5 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB20	8	6	2	78
360	BB21	9	8	1	76
360	BB30	6	5	1	79
360	BB31	5	5	0	79
360	BB32	1	1	0	83
720	BB20	20	20	0	106
720	BB21	27	27	0	99
720	BB30	10	10	0	116
720	BB31	15	15	0	111
720	BB32	8	8	0	118
1080	BB20	35	35	0	133
1080	BB21	52	52	0	116
1080	BB30	21	21	0	147
1080	BB31	41	41	0	127
1080	BB32	16	16	0	152
1440	BB20	52	50	2	160
1440	BB21	75	75	0	135
1440	BB30	33	32	1	178
1440	BB31	55	55	0	155
1440	BB32	23	22	1	188
1800	BB20	56	54	2	240
1800	BB21	96	96	0	198
1800	BB30	34	34	0	260
1800	BB31	73	73	0	221
1800	BB32	23	23	0	271
2160	BB20	66	64	2	230
2160	BB21	109	109	0	185
2160	BB30	42	41	1	253
2160	BB31	82	82	0	212
2160	BB32	26	26	0	268

Table 7.37. Demonstration Analysis: All trials model BB severity 3+ adverse event incidence results for the first 6 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	BB20	78	76	2	218
2520	BB21	111	111	0	183
2520	BB30	50	49	1	245
2520	BB31	86	86	0	208
2520	BB32	29	29	0	265

Table 7.38. Demonstration Analysis: All trials model BB severity 3+ adverse event incidence results for the final safety analysis.

Here the models with the level 1 dependence (BB_{21}, BB_{31}) performed best overall with the two-level hierarchy model, BB_{21}, performing better than the three-level model, particularly for severity 3+ events. The reduced number of parameters in BB_{21}, compared to BB_{31}, has enabled this better performance.

7.7.3.2 Total Event Analysis

We again expect an analysis including all events to give similar results to the incidence data. The outputs for models 1a, with 95% cut-off, are given in Tables 7.39-7.42, and for BB, with 90% cut-off, in Tables 7.43-7.46.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	97	41	56	43
360	1a21	86	39	47	45
360	1a30	102	40	62	44
360	1a31	89	39	50	45
360	1a32	100	39	61	45
720	1a20	230	71	159	55
720	1a21	181	78	103	48
720	1a30	228	68	160	58
720	1a31	178	78	100	48
720	1a32	224	68	156	58

Table 7.39. Demonstration Analysis: All trials model 1a severity 1+ total adverse event results for the first 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1080	1a20	352	105	247	63
1080	1a21	251	111	140	57
1080	1a30	350	103	247	65
1080	1a31	250	111	139	57
1080	1a32	333	96	237	72
1440	1a20	495	133	362	77
1440	1a21	326	140	186	70
1440	1a30	487	130	357	80
1440	1a31	323	139	184	71
1440	1a32	476	128	348	82
1800	1a20	625	177	448	117
1800	1a21	384	185	199	109
1800	1a30	624	171	453	123
1800	1a31	385	185	200	109
1800	1a32	612	162	450	132
2160	1a20	686	192	494	102
2160	1a21	447	202	245	92
2160	1a30	684	190	494	104
2160	1a31	449	202	247	92
2160	1a32	670	186	484	108
2520	1a20	705	201	504	93
2520	1a21	465	208	257	86
2520	1a30	707	200	507	94
2520	1a31	463	207	256	87
2520	1a32	705	196	509	98

Table 7.40. Demonstration Analysis: All trials model 1a severity 1+ total adverse event results for the final 5 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	44	13	31	71
360	1a21	44	21	23	63
360	1a30	35	11	24	73
360	1a31	36	16	20	68
360	1a32	33	11	22	73
720	1a20	140	39	101	87
720	1a21	93	44	49	82
720	1a30	119	34	85	92
720	1a31	92	42	50	84
720	1a32	108	32	76	94
1080	1a20	203	64	139	104
1080	1a21	129	68	61	100
1080	1a30	190	56	134	112
1080	1a31	129	68	61	100
1080	1a32	172	47	125	121
1440	1a20	256	76	180	134
1440	1a21	160	95	65	115
1440	1a30	242	67	175	143
1440	1a31	155	91	64	119
1440	1a32	222	61	161	149
1800	1a20	333	105	228	189
1800	1a21	189	112	77	182
1800	1a30	318	88	230	206
1800	1a31	190	112	78	182
1800	1a32	287	71	216	223
2160	1a20	391	123	268	171
2160	1a21	220	134	86	160
2160	1a30	384	115	269	179
2160	1a31	218	134	84	160
2160	1a32	344	99	245	195

Table 7.41. Demonstration Analysis: All trials model 1a severity 3+ total adverse event results for the first 6 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	413	130	283	164
2520	1a21	238	143	95	151
2520	1a30	406	119	287	175
2520	1a31	236	140	96	154
2520	1a32	368	104	264	190

Table 7.42. Demonstration Analysis: All trials model 1a severity 3+ total adverse event results for the final safety analysis.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB20	25	24	1	60
360	BB21	29	29	0	55
360	BB30	14	14	0	70
360	BB31	21	21	0	63
360	BB32	14	14	0	70
720	BB20	47	46	1	80
720	BB21	64	64	0	62
720	BB30	44	44	0	82
720	BB31	52	52	0	74
720	BB32	40	40	0	86
1080	BB20	71	70	1	98
1080	BB21	95	95	0	73
1080	BB30	61	60	1	108
1080	BB31	83	83	0	85
1080	BB32	57	56	1	112
1440	BB20	101	100	1	110
1440	BB21	123	122	1	88
1440	BB30	86	85	1	125
1440	BB31	113	112	1	98
1440	BB32	77	76	1	134

Table 7.43. Demonstration Analysis: All trials model BB severity 1+ total adverse event results for the first 4 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1800	BB20	131	129	2	165
1800	BB21	159	159	0	135
1800	BB30	101	100	1	194
1800	BB31	145	145	0	149
1800	BB32	89	89	0	205
2160	BB20	149	148	1	146
2160	BB21	171	171	0	123
2160	BB30	126	126	0	168
2160	BB31	160	160	0	134
2160	BB32	110	110	0	184
2520	BB20	151	151	0	143
2520	BB21	176	176	0	118
2520	BB30	130	130	0	164
2520	BB31	165	165	0	129
2520	BB32	119	119	0	175

Table 7.44. Demonstration Analysis: All trials model BB severity 1+ total adverse event results for the final 3 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB20	7	6	1	78
360	BB21	9	8	1	76
360	BB30	3	3	0	81
360	BB31	4	4	0	80
360	BB32	1	1	0	83
720	BB20	22	22	0	104
720	BB21	28	28	0	98
720	BB30	13	13	0	113
720	BB31	17	17	0	109
720	BB32	9	9	0	117

Table 7.45. Demonstration Analysis: All trials model BB severity 3+ total adverse event results for the first 2 safety analyses.

Time	Model	Flagged	Correct	Type-I	Type-II
1080	BB20	38	38	0	130
1080	BB21	49	49	0	119
1080	BB30	23	23	0	145
1080	BB31	40	40	0	128
1080	BB32	18	18	0	150
1440	BB20	47	46	1	164
1440	BB21	75	75	0	135
1440	BB30	30	29	1	181
1440	BB31	54	54	0	156
1440	BB32	23	23	0	187
1800	BB20	56	55	1	239
1800	BB21	91	90	1	204
1800	BB30	36	35	1	259
1800	BB31	74	74	0	220
1800	BB32	28	28	0	266
2160	BB20	71	70	1	224
2160	BB21	107	107	0	187
2160	BB30	43	42	1	252
2160	BB31	83	83	0	211
2160	BB32	33	33	0	261
2520	BB20	82	81	1	213
2520	BB21	116	116	0	178
2520	BB30	60	59	1	235
2520	BB31	89	89	0	205
2520	BB32	43	43	0	251

Table 7.46. Demonstration Analysis: All trials model BB severity 3+ total adverse event results for the final 5 safety analyses.

As in the case for the incidence data, the models with the stronger level dependence ($1a_{21}$, $1a_{31}$, BB_{21}, BB_{31}) performed the best overall. We look at the relative performances of the models with respect to event detection and error control in the next section.

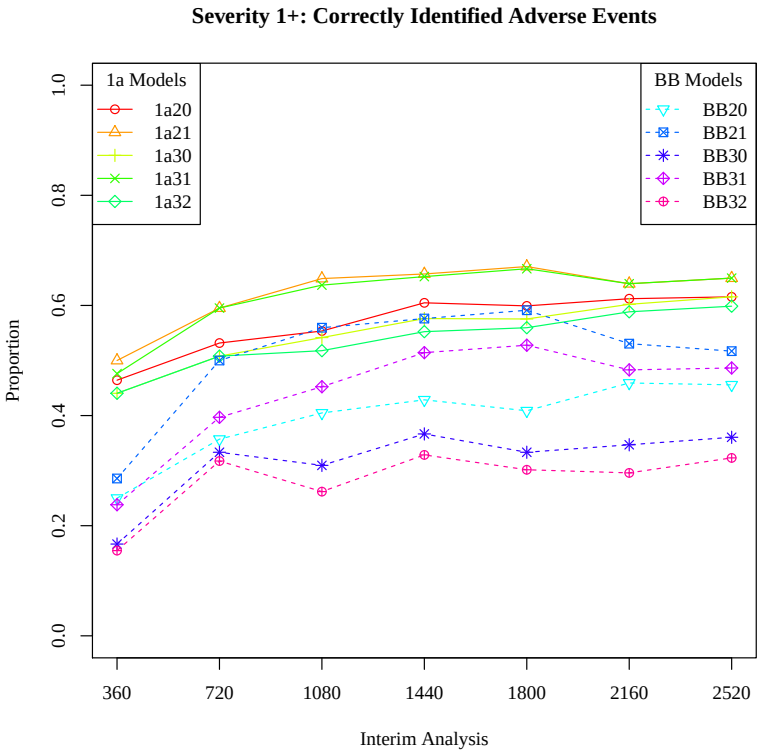
7.7.3.3 Model Comparison

We compare the models using the numbers of flagged events which were correctly identified as having raised treatment rates, and the overall Type-I error control.

For both severity 1+ and severity 3+ events we can see for the incidence data in Figures 7.22 and 7.23 that overall $1a_{21}$ and $1a_{31}$ correctly detect the most adverse events with raised treatment rates. The best performing point-mass model is BB_{21} which is comparable to some of the 1a models in terms of event detection.

All the models controlled the Type-I error rate at less than 5% with the BB models having very low rates. The results including all events, Figure 7.24 and 7.25, are similar to those for the incidence data. The results are consistent with what we saw in the simulation study in Chapter 5. The point-mass controls the Type-I error rate but reduces the number of correctly flagged adverse events.

For the data analysed, $1a_{21}$ and BB_{21} are arguably the best performing models without point-mass and with point-mass respectively.



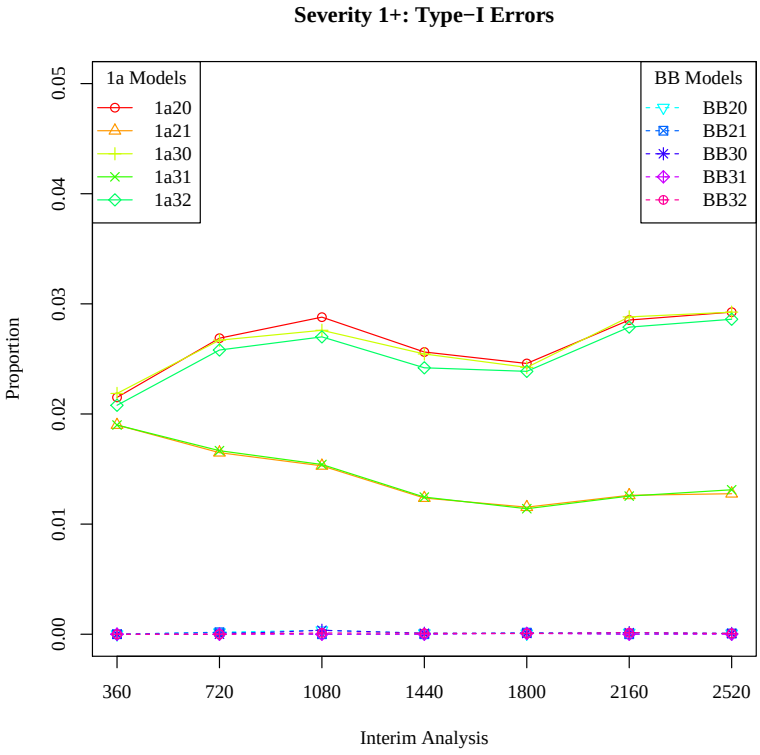


Figure 7.22. Demonstration Analysis: All trials adverse event incidence data severity 1+ events. Proportion correct and Type-I error rates.

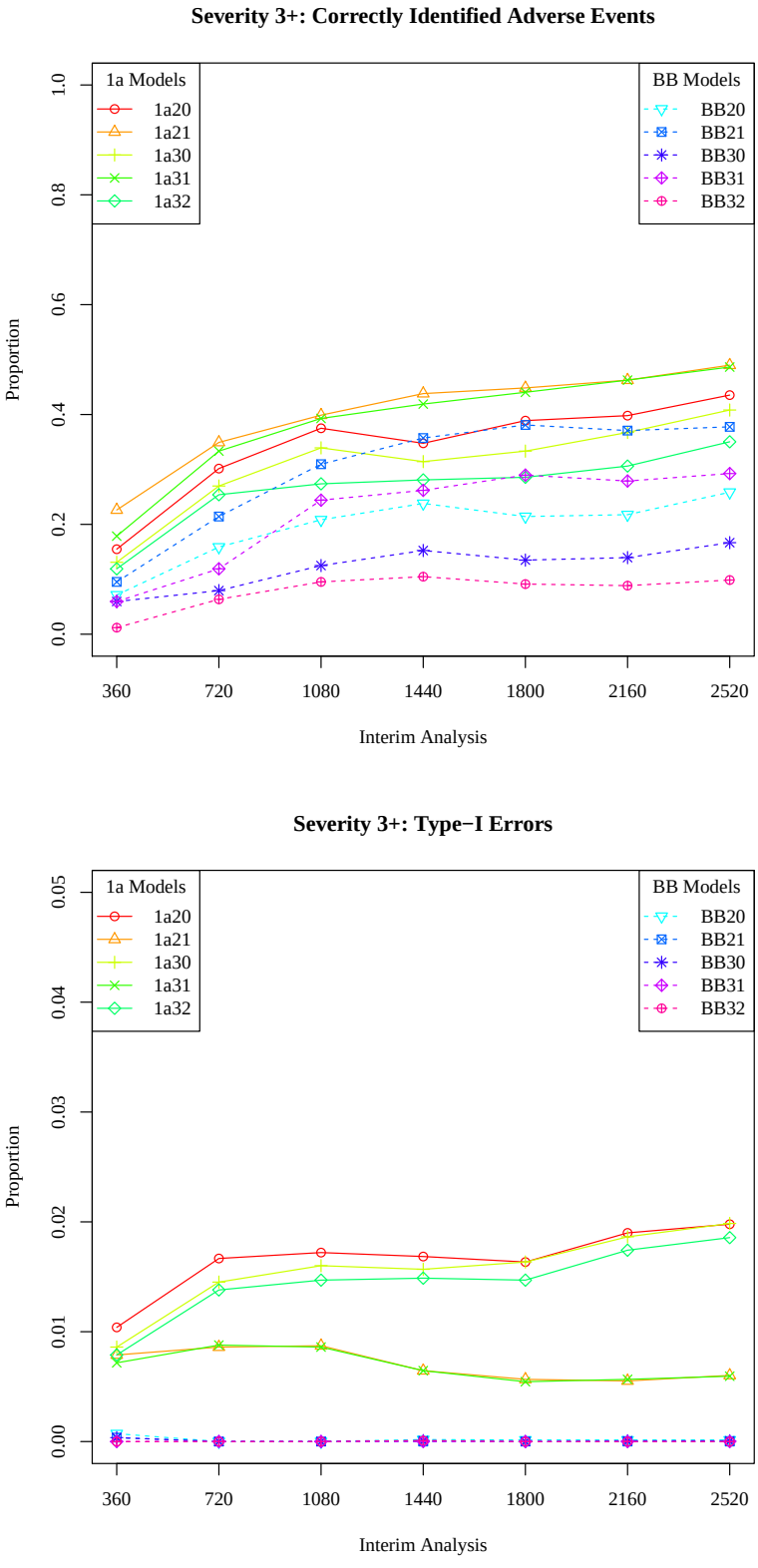


Figure 7.23. Demonstration Analysis: All trials adverse event incidence data severity 3+ events. Proportion correct and Type-I error rates.

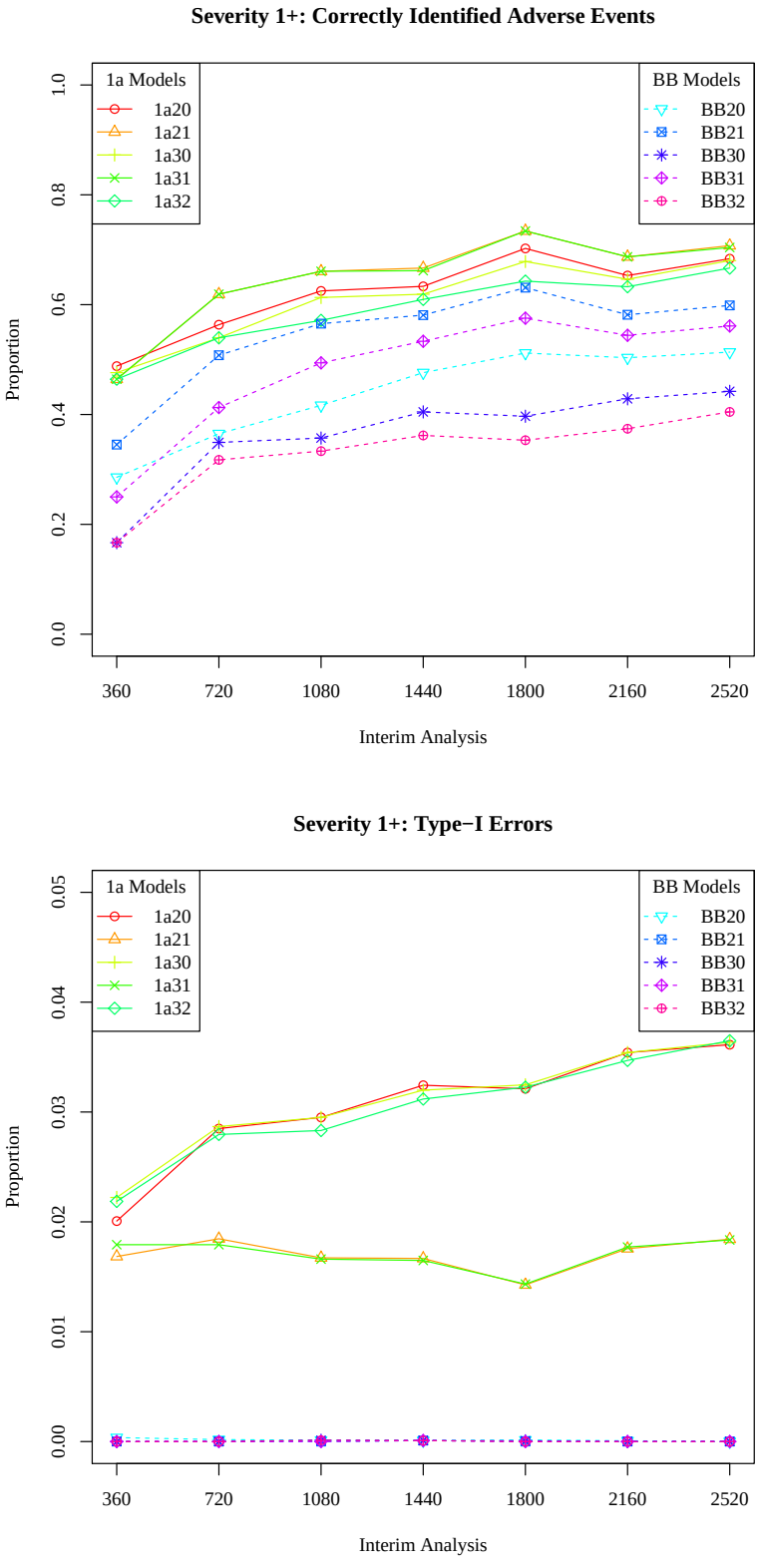


Figure 7.24. Demonstration Analysis: All trials total adverse event data severity 1+ events. Proportion correct and Type-I error rates.

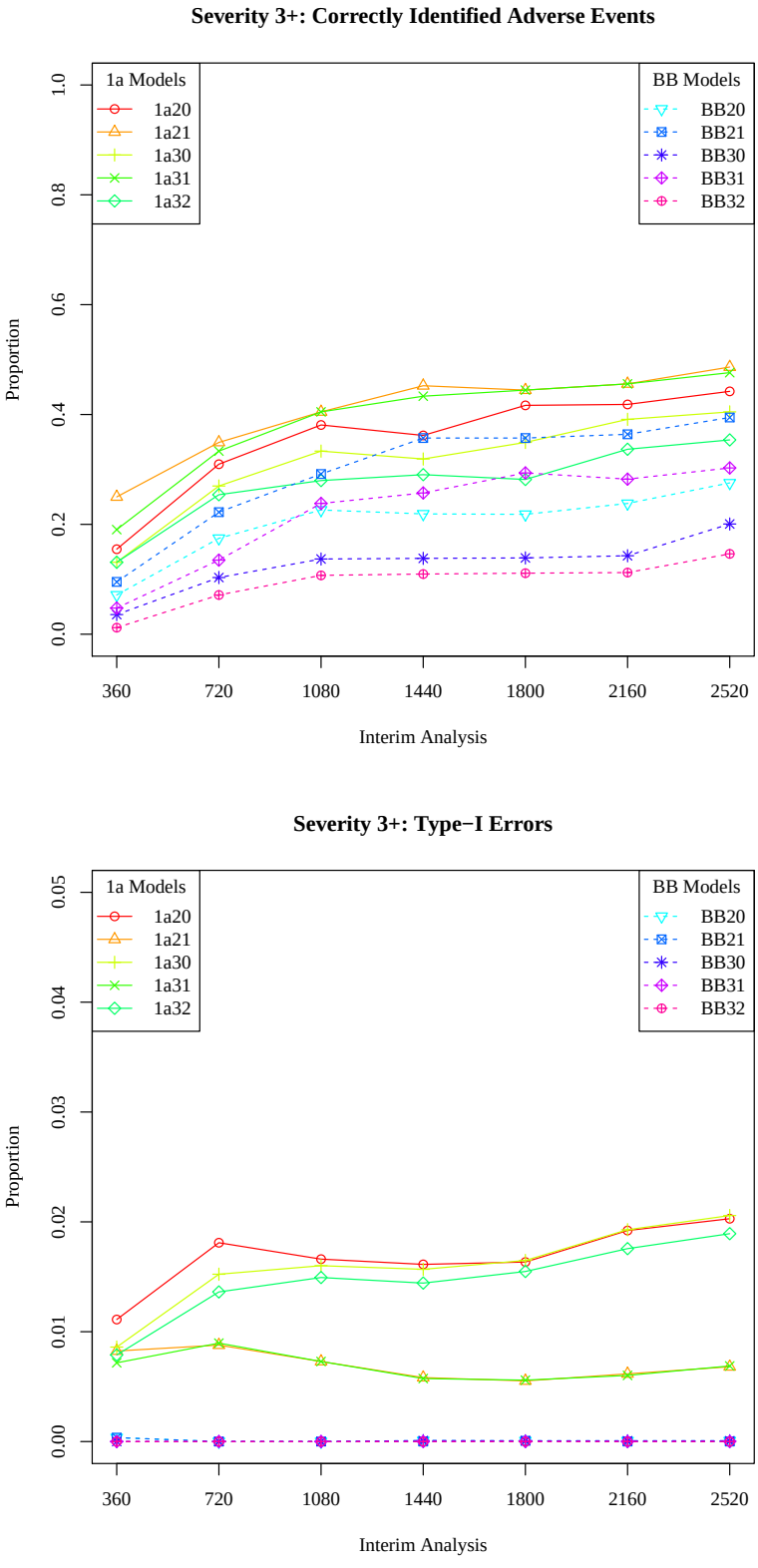


Figure 7.25. Demonstration Analysis: All trials total adverse event data severity 3+ events. Proportion correct and Type-I error rates.

7.8 Sensitivity Analysis

We are also interested in the behaviour of the models under a number of different circumstances. Here we consider how changing the thresholds for adverse event detection will change the model outputs and how well the models will perform with regard to missing patient data at the interim safety analyses. We also look at how the models perform when the background event rates are very much lower than in the demonstration analysis, and at a mixture of two different background rates, when one is considerably larger than the other.

7.8.1 Changing the Flagging Threshold

In §7.7.3.3 we tentatively identified models $1a_{21}$ and BB_{21} as the best performing models without and with a point-mass. In this section we re-analyse the outputs of these models for the trial incidence data using cut-offs for flagging events of 97.5% and 80% respectively. Here the cut-offs values are chosen based on our knowledge of the demonstration analyses, with the aim of seeing how the number of correctly flagged adverse events and Type-I error rates vary compared to the original analysis in §7.7.3.1.

7.8.1.1 Incidence Data (All Trials)

The results for $1a_{21}$ at the 97.5% cut-off are given in Tables 7.47 and 7.48, and for BB_{21} at the 80% cut-off in Tables 7.49 and 7.50.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	47	33	14	51
720	1a21	101	66	35	60
1080	1a21	132	93	39	75
1440	1a21	180	127	53	83
1800	1a21	207	154	53	140
2160	1a21	234	172	62	122
2520	1a21	242	177	65	117

Table 7.47. Sensitivity Analysis: Changed threshold model $1a_{21}$ results (all trials), severity 1+ events.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	20	12	8	72
720	1a21	44	35	9	91
1080	1a21	75	55	20	113
1440	1a21	103	78	25	132
1800	1a21	117	94	23	200
2160	1a21	135	109	26	185
2520	1a21	146	116	30	178

Table 7.48. Sensitivity Analysis: Changed threshold model $1a_{21}$ results (all trials), severity 3+ events.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	33	33	0	51
720	BB21	73	71	2	55
1080	BB21	102	102	0	66
1440	BB21	129	129	0	81
1800	BB21	162	160	2	134
2160	BB21	169	167	2	127
2520	BB21	166	163	3	131

Table 7.49. Sensitivity Analysis: Changed threshold model BB_{21} results (all trials), severity 1+ events.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	15	13	2	71
720	BB21	37	37	0	89
1080	BB21	63	63	0	105
1440	BB21	94	94	0	116
1800	BB21	120	119	1	175
2160	BB21	131	131	0	163
2520	BB21	133	133	0	161

Table 7.50. Sensitivity Analysis: Changed threshold model BB_{21} results (all trials), severity 3+ events.

For model $1a_{21}$, comparing for example Table 7.47 with Tables 7.31-7.32, we can

see that the numbers of detected adverse events have necessarily fallen but that the equivalent drop in the Type-I error rate is much more dramatic. On the other hand, for model BB_{21}, comparing Table 7.49 with Tables 7.35-7.36, we can see that the number of Type-I errors with the new threshold is still very low, but there is increased adverse event detection. For the severity 3+ events we can see from Tables 7.49 and 7.50 that the BB detection rate is better than 1a for the changed thresholds, but not quite as good as 1a under the original analysis (Tables 7.33-7.34). We plot the new results for $1a_{21}$ at the 95% cut-off and BB_{21} at the 90% cut-off in Figures 7.26 and 7.27.

In Figures 7.26 and 7.27 we can see that effect of lowering the threshold for BB_{21} is to improve the model's ability to flag adverse events correctly, and that for severity 1+ events the model is second only to $1a_{21}$ at the 95% threshold in terms of correctly detecting adverse events. For severity 3+ events BB_{21} with 80% threshold detects similar numbers of events as $1a_{21}$ at 95%. We can say that that for $1a_{21}$ increasing the threshold has the effect of reducing the numbers of Type-I errors and the number of events detected, which is now comparable to BB_{21} at the 90% cut-off. For BB_{21} the lower threshold has increased the numbers correctly detected, but the Type-I error rate remains very low compared to the other methods.

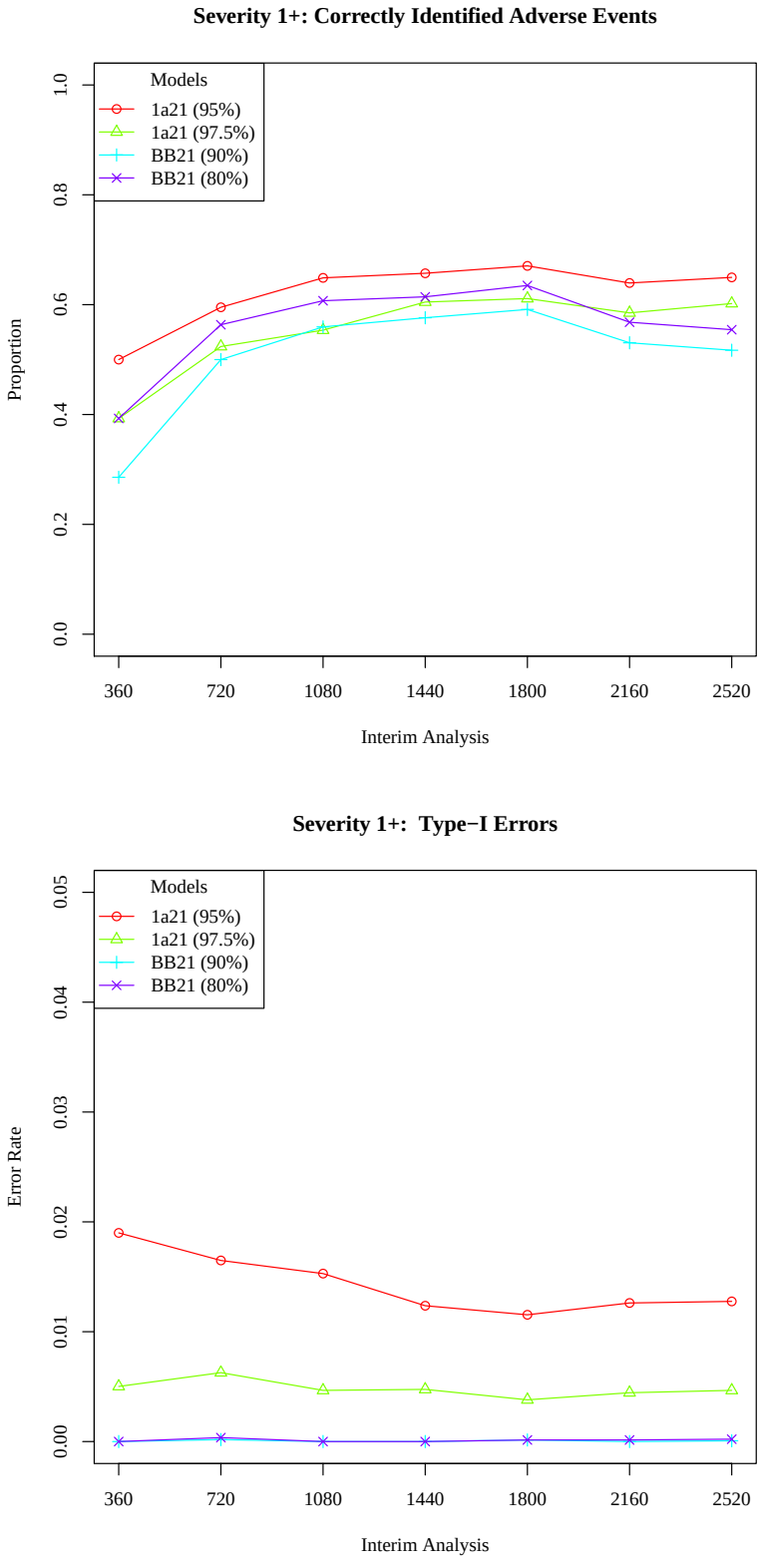


Figure 7.26. Sensitivity Analysis: Changed threshold results severity 1+ adverse event incidence data (all trials).

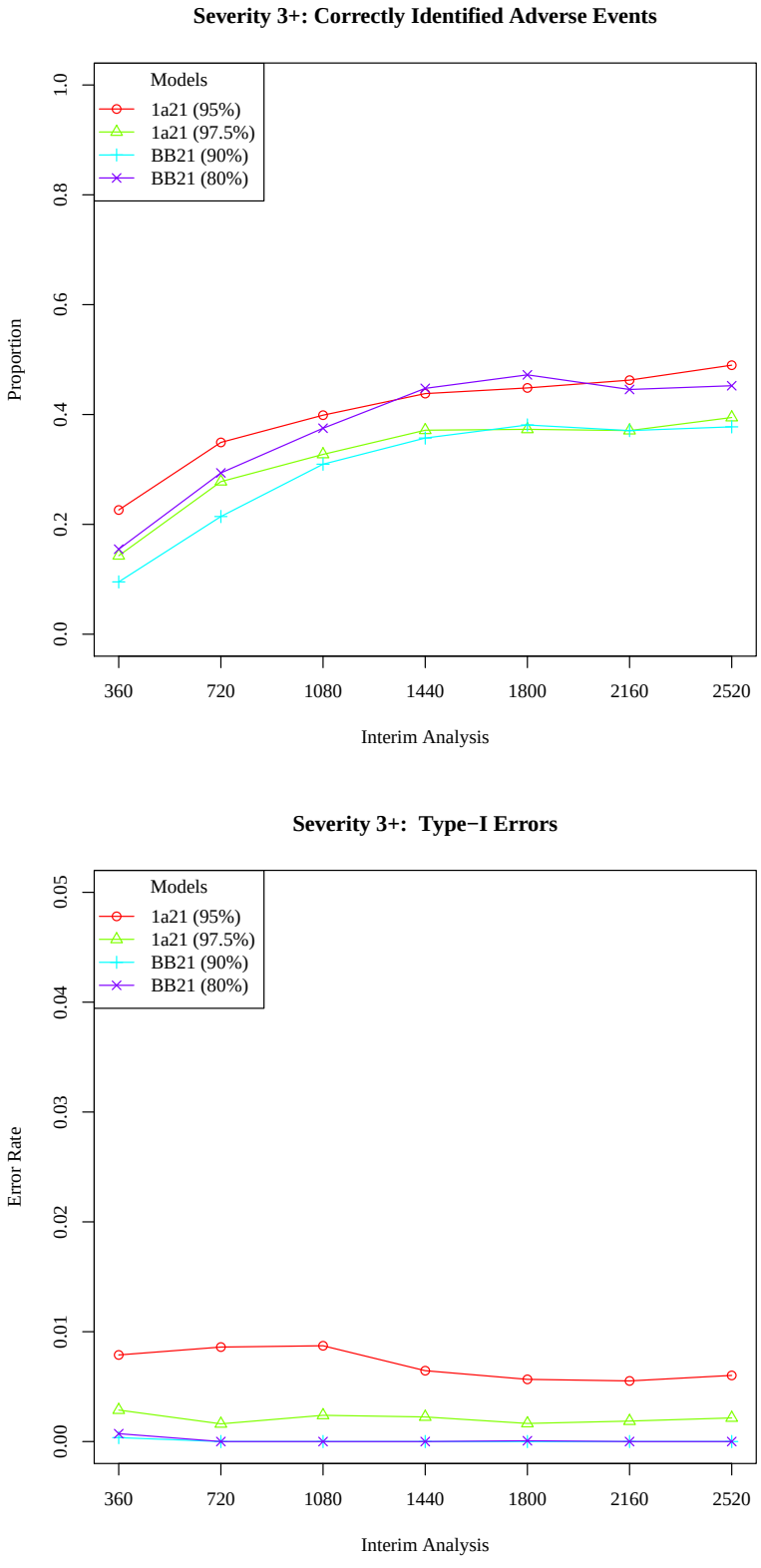


Figure 7.27. Sensitivity Analysis: Changed threshold results severity 3+ adverse event incidence data (all trials).

7.8.2 Missing Data

The results in §7.7 use all available simulated data. In this section we investigate the performance of the methods when patient data is missing. We do this by assuming that a patient who suffers a serious adverse event, severity 3 or higher, within a number of days of an upcoming interim analysis, is excluded from that interim analysis. The idea behind this approach is if events occur at a higher rate on the treatment arm then we may expect more treatment patients to be excluded from the analysis. The patients are not excluded from the trial in general. We wish to investigate how this reduction in events affects a model's ability to flag correctly adverse events with raised treatment rates at the interim safety analyses of interest. We confine the analysis to event incidence data for models $1a_{21}$ and BB_{21}. When flagging adverse events we use a 95% posterior probability threshold for the 1a models and 90% for the BB model. This allows for comparison with the results in §7.7.3.1. The four different cases we look at are:

Missing Data Scenario	**Exclusion Criterion**
Missing Data Case 1	Serious event within 2 days of an upcoming interim safety analysis.
Missing Data Case 2	Serious event within 5 days of an upcoming interim safety analysis.
Missing Data Case 3	Serious event within 10 days of an upcoming interim safety analysis.
Missing Data Case 4	Serious event within 20 days of an upcoming interim safety analysis.

Table 7.51. Sensitivity Analysis: Missing data scenarios.

7.8.2.1 Missing Data Case 1

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	88	38	50	46
720	1a21	171	74	97	52
1080	1a21	245	107	138	61
1440	1a21	285	141	144	69
1800	1a21	341	168	173	126
2160	1a21	381	186	195	108
2520	1a21	389	191	198	103

Table 7.52. Sensitivity Analysis: Missing data case 1, model $1a_{21}$ results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	41	20	21	64
720	1a21	93	40	53	86
1080	1a21	133	67	66	101
1440	1a21	161	94	67	116
1800	1a21	189	113	76	181
2160	1a21	214	131	83	163
2520	1a21	228	141	87	153

Table 7.53. Sensitivity Analysis: Missing data case 1, model $1a_{21}$ results (all trials), severity 3+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	25	25	0	59
720	BB21	61	61	0	65
1080	BB21	93	93	0	75
1440	BB21	123	123	0	87
1800	BB21	152	150	2	144
2160	BB21	153	153	0	141
2520	BB21	155	155	0	139

Table 7.54. Sensitivity Analysis: Missing data case 1, model BB_{21} results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	9	8	1	76
720	BB21	27	27	0	99
1080	BB21	51	51	0	117
1440	BB21	75	75	0	135
1800	BB21	93	93	0	201
2160	BB21	103	103	0	191
2520	BB21	107	107	0	187

Table 7.55. Sensitivity Analysis: Missing data case 1, model BB_{21} results (all trials), severity 3+.

7.8.2.2 Missing Data Case 2

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	93	40	53	44
720	1a21	163	73	90	53
1080	1a21	244	107	137	61
1440	1a21	282	138	144	72
1800	1a21	322	167	155	127
2160	1a21	368	186	182	108
2520	1a21	374	192	182	102

Table 7.56. Sensitivity Analysis: Missing data case 2, model $1a_{21}$ results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	38	18	20	66
720	1a21	100	41	59	85
1080	1a21	136	69	67	99
1440	1a21	163	90	73	120
1800	1a21	193	112	81	182
2160	1a21	215	127	88	167
2520	1a21	232	138	94	156

Table 7.57. Sensitivity Analysis: Missing data case 2, model $1a_{21}$ results (all trials), severity 3+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	24	24	0	60
720	BB21	65	64	1	62
1080	BB21	92	92	0	76
1440	BB21	120	120	0	90
1800	BB21	148	148	0	146
2160	BB21	149	149	0	145
2520	BB21	153	153	0	141

Table 7.58. Sensitivity Analysis: Missing data case 2, model BB_{21} results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	11	10	1	74
720	BB21	26	26	0	100
1080	BB21	50	50	0	118
1440	BB21	71	71	0	139
1800	BB21	92	92	0	202
2160	BB21	103	103	0	191
2520	BB21	109	109	0	185

Table 7.59. Sensitivity Analysis: Missing data case 2, model BB_{21} results (all trials), severity 3+.

7.8.2.3 Missing Data Case 3

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	77	36	41	48
720	1a21	156	65	91	61
1080	1a21	261	106	155	62
1440	1a21	298	139	159	71
1800	1a21	340	175	165	119
2160	1a21	381	194	187	100
2520	1a21	390	197	193	97

Table 7.60. Sensitivity Analysis: Missing data case 3, model $1a_{21}$ results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	39	14	25	70
720	1a21	83	36	47	90
1080	1a21	123	64	59	104
1440	1a21	142	83	59	127
1800	1a21	165	103	62	191
2160	1a21	196	122	74	172
2520	1a21	204	129	75	165

Table 7.61. Sensitivity Analysis: Missing data case 3, model $1a_{21}$ results (all trials), severity 3+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	24	24	0	60
720	BB21	58	58	0	68
1080	BB21	91	91	0	77
1440	BB21	118	118	0	92
1800	BB21	148	148	0	146
2160	BB21	157	157	0	137
2520	BB21	159	159	0	135

Table 7.62. Sensitivity Analysis: Missing data case 3, model BB_{21} results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	8	8	0	76
720	BB21	23	23	0	103
1080	BB21	48	48	0	120
1440	BB21	64	64	0	146
1800	BB21	87	87	0	207
2160	BB21	96	96	0	198
2520	BB21	103	103	0	191

Table 7.63. Sensitivity Analysis: Missing data case 3, model BB_{21} results (all trials), severity 3+.

7.8.2.4 Missing Data Case 4

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	61	32	29	52
720	1a21	134	68	66	58
1080	1a21	224	102	122	66
1440	1a21	280	136	144	74
1800	1a21	319	168	151	126
2160	1a21	363	194	169	100
2520	1a21	375	193	182	101

Table 7.64. Sensitivity Analysis: Missing data case 4, model $1a_{21}$ results (all trials), severity 1+

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a21	29	15	14	69
720	1a21	60	31	29	95
1080	1a21	103	61	42	107
1440	1a21	130	81	49	129
1800	1a21	151	102	49	192
2160	1a21	174	116	58	178
2520	1a21	190	125	65	169

Table 7.65. Sensitivity Analysis: Missing data case 4, model $1a_{21}$ results (all trials), severity 3+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	20	20	0	64
720	BB21	51	51	0	75
1080	BB21	80	80	0	88
1440	BB21	111	111	0	99
1800	BB21	138	138	0	156
2160	BB21	150	150	0	144
2520	BB21	150	149	1	145

Table 7.66. Sensitivity Analysis: Missing data case 4, model BB_{21} results (all trials), severity 1+.

Time	Model	Flagged	Correct	Type-I	Type-II
360	BB21	7	7	0	77
720	BB21	19	19	0	107
1080	BB21	45	45	0	123
1440	BB21	66	66	0	144
1800	BB21	84	84	0	210
2160	BB21	96	95	1	199
2520	BB21	100	100	0	194

Table 7.67. Sensitivity Analysis: Missing data case 4, model BB_{21} results (all trials), severity 3+.

7.8.2.5 Discussion

Comparing the results above to Tables 7.31-7.38 we can see that the missing data does not generally affect the detection of severity 1+ events but, for severity 3+ events, the detection rates appear to decline as the range for excluding events increases towards 20 days (Figure 7.28). So for model $1a_{21}$ we correctly detect 191 severity 1+ by the end of the trial with no missing data but, for the case where patients are removed from the trial if they have a severe event within 20 days of an interim analysis, the number detected is actually 193. For BB_{21} the equivalent numbers are 152 and 149. However, for severity 3+ events, the numbers are 144 versus 125 detected for $1a_{21}$, and 111 versus 100 for BB_{21}. The event detection fall-off is not every dramatic, and the methods look to be quite robust to this type of patient censoring. However, it should be stressed that for rare events with low rates of occurrences there may not be much differences between the numbers excluded on each arm of the trial, and hence the differences between the detection of adverse events from the full data set and reduced data may not differ by a large amount, as we see here.

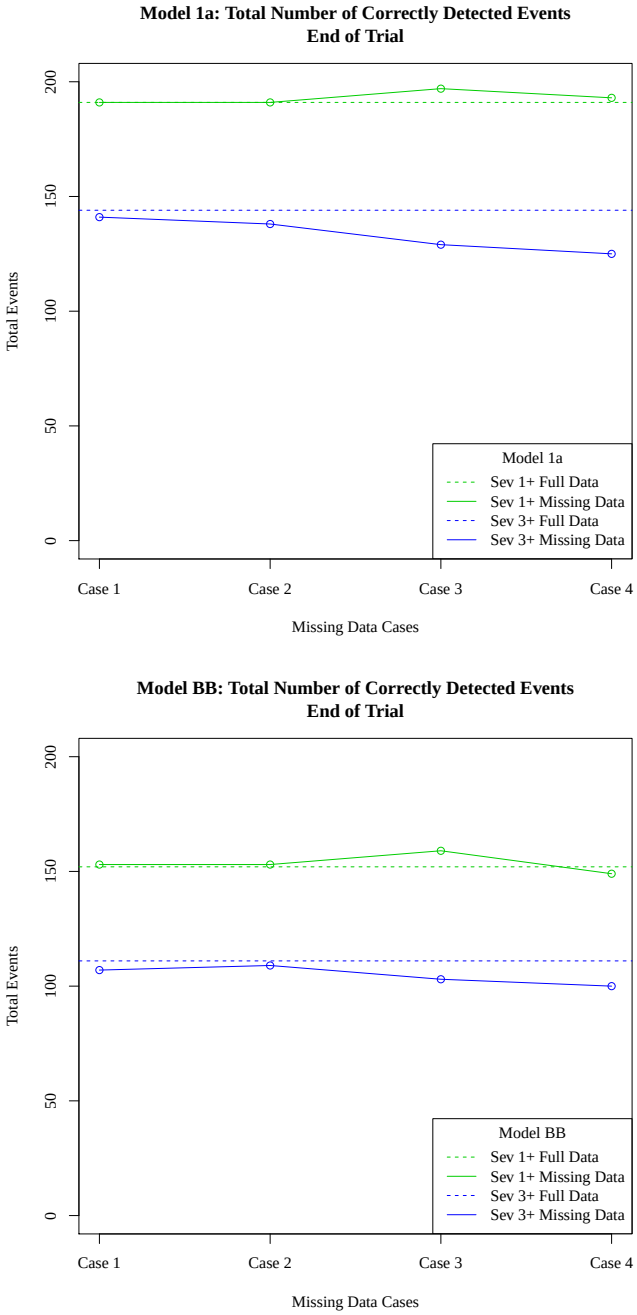


Figure 7.28. Sensitivity Analysis: End of trial detection totals for missing data by case (model 1a on top, model BB on bottom).

7.8.3 Lower Background Trial Adverse Event Rates

Here we look at how the different models perform for lower background rates. The rate we use, specified in Table 7.68, is one tenth that used in the demonstration analyses above (Table 7.6).

Rate	Expected Number of Events	Probability of an Event
0.00005555	0.09999	0.09515353

Table 7.68. Sensitivity Analysis: Low background rate adverse event rate.

As in §7.5 patients are recruited into the trial up to day 720 or a maximum of 1000 on each arm. The simulated numbers in the trials at each interim analysis are as follows:

Simulated[1] Trial	Time	Control	Treatment
I(d)	360	453	479
I(d)	720 -End of Trial	916	950
II(d)	360	434	457
II(d)	720 -End of Trial	891	879
III(d)	360	439	441
III(d)	720 -End of Trial	914	952

Table 7.69. Sensitivity Analysis: Low background rate trial patient enrolment totals.

[1] Unique simulation name consisting of trial type and identifier (d).

The increased treatment rates, relative to the corresponding control adverse event rates, and the intervals over which they apply are given in Table 7.70.

Simulated Trial	Body-System	Intervals	Relative Increase in Treatment Rate
I(d)	Body-sys_3	All	100%
II(d)	Body-sys_3	[0-180]	100%
II(d)	Body-sys_3	(180-360]	25%
III(d)	Body-sys_3	(1440-1620]	25%
III(d)	Body-sys_3	(1620-1800]	100%

Table 7.70. Sensitivity Analysis: Low rate events, body-systems and intervals with increased treatment rates.

7.8.3.1 Results

We look at the results for event incidence, the results for including all events in the analysis are very similar. The severity 1+ results for adverse event incidence early in the trial are given in Table 7.71, and for the end of trial in Table 7.72.

We can see from Table 7.71 that for severity 1+ events early in the trial model $1a_{21}$ performs best overall, with highest detection rate and a low number of Type-I errors. The only comparable BB model is BB_{21} whose event detection rate rises sharply in the second interval. By the end of the trial (Table 7.72) the level 1 models have all performed the best in terms of both detection and error control.

For severity 3+ events all models struggle to detect events in the first interval (Table 7.73) but, by the second interval, $1a_{21}$ is the best performing model. None of the BB models detect any events until the third interval. By the end of trial (Table 7.74) model $1a_{21}$ has detected most events, with a low overall number of Type-I errors. Of the BB models only BB_{21} has a comparable performance.

Overall, for severity 1+ and severity 3+ events, the model which performs best is $1a_{21}$. Early in the trial it detects events more quickly than the other methods, and, by the end of the trial, it has correctly detected most events, while keeping a low overall number of Type-I errors. As in the simulation study (Chapter 5) and the demonstrations analysis (§7.7), the BB models keep very tight control of the Type-I errors at the expense of increased Type-II errors.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	6	5	1	23
360	1a21	12	10	2	18
360	1a30	3	2	1	26
360	1a31	4	3	1	25
360	1a32	3	2	1	26
360	BB20	1	1	0	27
360	BB21	2	2	0	26
360	BB30	1	1	0	27
360	BB31	1	1	0	27
360	BB32	1	1	0	27
720	1a20	27	17	10	25
720	1a21	30	24	6	18
720	1a30	23	15	8	27
720	1a31	26	21	5	21
720	1a32	24	15	9	27
720	BB20	9	9	0	33
720	BB21	14	14	0	28
720	BB30	4	4	0	38
720	BB31	6	6	0	36
720	BB32	4	4	0	38
1080	1a20	56	31	25	25
1080	1a21	43	40	3	16
1080	1a30	46	29	17	27
1080	1a31	43	39	4	17
1080	1a32	42	28	14	28
1080	BB20	16	16	0	40
1080	BB21	36	36	0	20
1080	BB30	9	9	0	47
1080	BB31	20	20	0	36
1080	BB32	4	4	0	52

Table 7.71. Sensitivity Analysis: Low rate severity 1+ adverse event incidence early trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	123	62	61	36
2520	1a21	86	75	11	23
2520	1a30	109	56	53	42
2520	1a31	86	75	11	23
2520	1a32	94	50	44	48
2520	BB20	29	29	0	69
2520	BB21	70	70	0	28
2520	BB30	15	15	0	83
2520	BB31	55	55	0	43
2520	BB32	5	5	0	93

Table 7.72. Sensitivity Analysis: Low rate severity 1+ adverse event incidence end of trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	3	2	1	26
360	1a21	3	2	1	26
360	1a30	3	2	1	26
360	1a31	2	1	1	27
360	1a32	1	0	1	28
360	BB20	0	0	0	28
360	BB21	0	0	0	28
360	BB30	0	0	0	28
360	BB31	0	0	0	28
360	BB32	0	0	0	28
720	1a20	9	7	2	35
720	1a21	12	10	2	32
720	1a30	5	4	1	38
720	1a31	5	5	0	37
720	1a32	2	2	0	40
720	BB20	0	0	0	42
720	BB21	0	0	0	42
720	BB30	0	0	0	42
720	BB31	0	0	0	42
720	BB32	0	0	0	42
1080	1a20	31	15	16	41
1080	1a21	33	27	6	29
1080	1a30	13	10	3	46
1080	1a31	26	22	4	34
1080	1a32	11	7	4	49
1080	BB20	3	3	0	53
1080	BB21	13	13	0	43
1080	BB30	0	0	0	56
1080	BB31	0	0	0	56
1080	BB32	0	0	0	56

Table 7.73. Sensitivity Analysis: Low rate severity 3+ adverse event incidence early trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	60	30	30	68
2520	1a21	75	67	8	31
2520	1a30	35	20	15	78
2520	1a31	73	64	9	34
2520	1a32	24	14	10	84
2520	BB20	4	3	1	95
2520	BB21	55	55	0	43
2520	BB30	0	0	0	98
2520	BB31	23	23	0	75
2520	BB32	0	0	0	98

Table 7.74. Sensitivity Analysis: Low rate severity 3+ adverse event incidence end of trial results.

7.8.4 Mixed Adverse Event Background Rates

In a mixed background event rate trial simulation we allow the events to occur at either of the background rates from Tables 7.6 or Table 7.68. The adverse events which occur at the higher background rate (Table 7.6) are given in Table 7.75. All other events occur at the lower frequency.

Body-system	Adverse Event
Bdy-sys_1	Adv-Ev_1
Bdy-sys_2	Adv-Ev_12
Bdy-sys_3	Adv-Ev_19
Bdy-sys_3	Adv-Ev_21
Bdy-sys_3	Adv-Ev_23
Bdy-sys_4	Adv-Ev_28
Bdy-sys_5	Adv-Ev_41
Bdy-sys_6	Adv-Ev_53
Bdy-sys_7	Adv-Ev_54

Table 7.75. Sensitivity Analysis: Mixed event rates, adverse events with higher background rates.

The simulated patient recruitment is give in Table 7.76, and the increased treatment rates and the intervals over which they apply are given in Table 7.77.

Simulated Trial	Time	Control	Treatment
I(e)	360	442	479
I(e)	720 -End of Trial	887	926
II(e)	360	478	472
II(e)	720 - End of Trial	969	952
III(e)	360	494	427
III(e)	720 -End of Trial	985	876

Table 7.76. Sensitivity Analysis: Mixed event rates, patient enrolment totals.

Simulated Trial	Body-System	Intervals	Relative Increase in Treatment Rate
I(e)	Body-sys_3	All	100%
II(e)	Body-sys_3	[0-180]	100%
II(e)	Body-sys_3	(180-360]	25%
III(e)	Body-sys_3	(1440-1620]	25%
III(e)	Body-sys_3	(1620-1800]	100%

Table 7.77. Sensitivity Analysis: Mixed event rates, body-systems and intervals with increased treatment rates.

7.8.4.1 Results

The mixed background rate trial results for event incidence are given below. For severity 1+ event we can see from Tables 7.78 and 7.79 that $1a_{21}$ and BB_{21} perform the best overall. However, for severity 3+ events (Tables 7.80 and 7.81), only the 1a models are capable of detecting events early in the trial, and even by day 1080 the best performing BB model, BB_{21}, has only correctly detected just over half the number that $1a_{21}$ has, albeit with a lower number of Type-I errors. As is the case for the lower frequency events, model $1a_{21}$ could be considered the best performing of all the models. As in the other trial types we analysed, the results including all events were similar.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	7	2	5	26
360	1a21	5	2	3	26
360	1a30	6	2	4	26
360	1a31	6	2	4	26
360	1a32	7	2	5	26
360	BB20	0	0	0	28
360	BB21	2	2	0	26
360	BB30	0	0	0	28
360	BB31	0	0	0	28
360	BB32	0	0	0	28
720	1a20	39	22	17	20
720	1a21	32	26	6	16
720	1a30	33	22	11	20
720	1a31	31	25	6	17
720	1a32	32	21	11	21
720	BB20	13	11	2	31
720	BB21	23	22	1	20
720	BB30	11	9	2	33
720	BB31	11	11	0	31
720	BB32	8	7	1	35
1080	1a20	59	33	26	23
1080	1a21	46	34	12	22
1080	1a30	60	34	26	22
1080	1a31	47	34	13	22
1080	1a32	57	33	24	23
1080	BB20	17	17	0	39
1080	BB21	31	31	0	25
1080	BB30	13	13	0	43
1080	BB31	19	19	0	37
1080	BB32	11	11	0	45

Table 7.78. Sensitivity Analysis: Mixed rate severity 1+ adverse event incidence early trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	121	65	56	33
2520	1a21	93	73	20	25
2520	1a30	117	64	53	34
2520	1a31	94	73	21	25
2520	1a32	106	61	45	37
2520	BB20	40	39	1	59
2520	BB21	67	65	2	33
2520	BB30	29	28	1	70
2520	BB31	51	49	2	49
2520	BB32	23	22	1	76

Table 7.79. Sensitivity Analysis: Mixed rate severity 1+ adverse event incidence end of trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
360	1a20	6	2	4	26
360	1a21	3	2	1	26
360	1a30	3	1	2	27
360	1a31	3	2	1	26
360	1a32	3	1	2	27
360	BB20	0	0	0	28
360	BB21	0	0	0	28
360	BB30	0	0	0	28
360	BB31	0	0	0	28
360	BB32	0	0	0	28
720	1a20	24	7	17	35
720	1a21	19	11	8	31
720	1a30	18	8	10	34
720	1a31	17	10	7	32
720	1a32	14	6	8	36
720	BB20	4	0	4	42
720	BB21	4	3	1	39
720	BB30	2	0	2	42
720	BB31	0	0	0	42
720	BB32	2	0	2	42
1080	1a20	34	13	21	43
1080	1a21	22	15	7	41
1080	1a30	31	14	17	42
1080	1a31	21	15	6	41
1080	1a32	27	12	15	44
1080	BB20	5	4	1	52
1080	BB21	10	8	2	48
1080	BB30	5	3	2	53
1080	BB31	5	4	1	52
1080	BB32	3	2	1	54

Table 7.80. Sensitivity Analysis: Mixed rate severity 3+ adverse event incidence early trial results.

Time	Model	Flagged	Correct	Type-I	Type-II
2520	1a20	65	27	38	71
2520	1a21	47	42	5	56
2520	1a30	60	31	29	67
2520	1a31	45	40	5	58
2520	1a32	44	22	22	76
2520	BB20	8	7	1	91
2520	BB21	33	32	1	66
2520	BB30	6	5	1	93
2520	BB31	14	14	0	84
2520	BB32	2	1	1	97

Table 7.81. Sensitivity Analysis: Mixed rate severity 3+ adverse event incidence end of trial results.

7.9 Discussion

The demonstration and sensitivity analyses above are based on a very small number of the total possible trial scenarios which may exist. They were chosen to cover a number of specific types of treatment adverse event rate increases, namely increased rates over the whole course of the trial, increased rates early in the trial, and increased rates later in the trial. The fact that most events with increased rates will be from Trial Type I (increased rates over the whole course of the trial) has the potential to bias any conclusions that we may draw from the absolute numbers of flagged events in the analyses above, and care must be taken when drawing conclusions.

From the simulation study in Chapter 5 we may expect that the models without the point-mass will detect the most number of adverse events with raised treatment rates but have higher Type-I error rate, and we see this in the demonstration and sensitivity analyses. However, the BB_{21} model performed quite well in comparison to a number of the 1a models, with much tighter error control, and lowering the flagging threshold from 90% to 80% did increase the number of flagged adverse events without inflating the Type-I error rates.

Overall we can say that:

- The $1a_{21}$, $1a_{31}$ models perform the best over all the demonstration analyses

in terms of correctly identifying adverse events with raised rates. On the occasions where the $1a_{20}$ or $1a_{30}$ models performed slightly better in terms of detecting adverse events with raised treatment rates, they have higher error rates. However, models $1a_{20}$ and $1a_{30}$ essentially assume that each interval is independent, so there is no direct or indirect control for multiplicities over the intervals. If we did wish to apply some control mechanism then the number of flagged adverse events would fall in comparison to $1a_{21}$ or $1a_{31}$. The level 1 (and 2) models already take a dependency into account and do not need the addition of an extra error controlling step.

- As the difference in rates becomes smaller the power to detect the differences drops off rapidly for all methods, but $1a_{21}$ and $1a_{31}$ control the Type-1 error rate more tightly than the other 1a methods and correctly detect more events. The BB_{21} model is the best performing point-mass model in this case. In Chapter 5 we have seen that 1a models perform better than BB models in terms of adverse event detection, when the differences in rates are small and we find similar results here.

- For low treatment rates model $1a_{21}$ detects more events early in the trial than any of the other methods. This is particularly true for severity 3+ events. Its control of the Type-1 error rate is tighter than the other 1a methods. The BB_{21} model is the best performing point-mass model in these cases, but it does not detect any severity 3+ events until much later in the trial when more events have accumulated (Table 7.73). This is the effect of the point-mass.

- The estimates of the underlying parameters produced by the models are quite accurate for high treatment rates, with the 1a models giving slightly better estimates than the BB models for the same data.

- The BB models generally need a strong signal in order to detect an adverse event with a raised treatment rate. This keeps the Type-I error rate very low but militates against the detection of events, particularly when the occurrence rates are low, and especially for low rate severity 3+ events. Some of the three-level models do not flag any events at all and the posterior probability that $\theta_{bj,h} > 0$ is very small. This also leads to poor point estimates. We have seen similar behaviour in Chapter 5 for point-mass models with low rates. The three-level BB models have too many parameters, leading to overfitting, and this makes overcoming the effect of the point-mass difficult.

Removing a level of the hierarchy, as in §6.7.3, reduces the number of parameters required and produces better results overall, sometimes on a par with the 1a models.

- The 1a models have less parameters and converge more quickly than the equivalent BB models. In the demonstration analyses we rarely had to alter any of the global simulation parameters for the 1a models. For the BB models on a number of occasions we needed to change the global defaults for certain body-systems and adverse events. Convergence and parameter tuning is discussed in Appendix C.

Based on the demonstration analyses it looks that overall model $1a_{21}$ is the better model for the data we considered. Its control of the Type-I error rate is the best of all 1a models. However, there are occasions where the other 1a models perform slightly better at certain stages of the analysis. For trials with low rates, or low rate differences between treatment and control, the BB models have difficulty overcoming the influence of the point-mass. We can also see the effect of this in the estimates of the parameters. For higher rate differences the BB models do detect more events but only BB_{21} gives a performance anywhere near as good as the 1a models in terms of event detection. Given that this model has better Type-I error control than the 1a models, we investigated further the effect of lowering the flagging threshold for this model and this did show improved results (§7.8.1).

The results for models $1a_{21}$ and $1a_{31}$ are very similar apart from the early intervals of the low frequency event trials (Tables 7.71, 7.73). For the data considered here the addition of the third level to the hierarchy does not appear to have any advantages compared to the equivalent two-level hierarchical model, and in fact the extra parameters appear to impede slightly the three-level model's performance compared to the two-level model. Looking at the estimates for a number of the parameters for the Trial II(a) at day 360 (the first interim analysis) for severity 1+ incidence data, given in Table 7.82, we see that the parameter estimates are very similar across the body-systems, and that for $1a_{31}$ the parameter $\mu_{\gamma 0}$, which is the third level parameter, also has a very similar value. The two-level model is able to adequately account for the variation in the trial data, and this is true even in the mixed background rate trials (§7.8.4.1).

Parameter	$1a_{21}$		$1a_{31}$	
	Mean	**95% HPI**	**Mean**	**95% HPI**
$\mu_{\gamma 1}$	-7.43	(-7.63, -7.24)	-7.44	(-7.63, -7.26)
$\mu_{\gamma 10}$	-7.23	(-7.37, -7.09)	-7.25	(-7.39, -7.11)
$\mu_{\gamma 11}$	-7.40	(-7.52, -7.28)	-7.41	(-7.53, -7.29)
$\mu_{\gamma 12}$	-8.22	(-8.51, -7.94)	-8.16	(-8.42, -7.88)
$\mu_{\gamma 13}$	-7.35	(-7.48, -7.21)	-7.36	(-7.49, -7.22)
$\mu_{\gamma 14}$	-7.48	(-7.63, -7.32)	-7.48	(-7.63, -7.33)
$\mu_{\gamma 15}$	-7.78	(-8.01, -7.55)	-7.77	(-7.99, -7.55)
$\mu_{\gamma 2}$	-7.52	(-7.75, -7.30)	-7.53	(-7.74, -7.31)
$\mu_{\gamma 3}$	-7.30	(-7.55, -7.05)	-7.32	(-7.56, -7.09)
$\mu_{\gamma 4}$	-7.48	(-7.70, -7.25)	-7.49	(-7.70, -7.27)
$\mu_{\gamma 5}$	-7.49	(-7.69, -7.28)	-7.50	(-7.70, -7.30)
$\mu_{\gamma 6}$	-7.80	(-7.99, -7.61)	-7.79	(-7.97, -7.61)
$\mu_{\gamma 7}$	-7.65	(-7.94, -7.38)	-7.66	(-7.91, -7.40)
$\mu_{\gamma 8}$	-7.23	(-7.49, -6.97)	-7.27	(-7.53, -7.02)
$\mu_{\gamma 9}$	-7.61	(-7.84, -7.38)	-7.62	(-7.84, -7.41)
$\mu_{\gamma 0}$	-	-	-7.53	(-7.74, -7.31)

Table 7.82. Demonstration Analysis: Trial II(a) model 1a level 1 parameter estimates.

7.10 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer

In this section we apply the methods to the safety data from GSK Trial EGF100151 (§1.8). The publicly available data does not contain the timings of patient recruitment or adverse event occurrence. However, we may still look to apply the interim analysis methods to final adverse event data, provided we make a number of assumptions. The median survival time of a subject under treatment (*lapatinib and capecitabine*) is 75 weeks (525 days) versus 64.7 weeks (452.9 days) for the control (*capecitabine*) [21]. We may use these as estimates of the time each subject remains in the trial and to calculate total exposure times. As we do not know when each patient experienced their adverse events, and thus dropped out of the risk-set for

that adverse event, this will give an over-estimate of the exposure times for the trial, and we should expect that this will have some impact on our results. The top 10 adverse events in terms of posterior probability for the methods $1a_{21}$ and BB_{21} are given in Tables 7.83 and 7.84.

System Organ Class	**Adverse Event**	**Posterior[1] probability $\theta > 0$**
Skin and subcutaneous tissue disorders	Rash	1.000
Gastrointestinal disorders	Diarrhoea	0.998
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.995
Gastrointestinal disorders	Dyspepsia	0.990
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.983
Skin and subcutaneous tissue disorders	Nail disorder	0.978
Respiratory, thoracic and mediastinal disorders	Dyspnoea	0.972
Musculoskeletal and connective tissue disorders	Arthralgia	0.965
Hepatobiliary disorders	Hyperbilirubinaemia	0.964
Musculoskeletal and connective tissue disorders	Back pain	0.957

Table 7.83. Trial EGF100151: Top 10 adverse events by posterior probability for $1a_{21}$.

[1] θ is the log relative risk (§6.8).

Comparing Table 7.83 to Table 4.4 we see very similar results with one exception. *Hyperbilirubinaemia* is now in the top 10. However, for the c212.1a model *Hyperbilirubinaemia* also had a high posterior probability but was not in the top 10 adverse events.

System Organ Class	Adverse Event	Posterior probability $\theta > 0$
Skin and subcutaneous tissue disorders	Rash	0.996
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.893
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.863
Gastrointestinal disorders	Diarrhoea	0.800
Gastrointestinal disorders	Dyspepsia	0.796
Skin and subcutaneous tissue disorders"	Nail disorder	0.788
Skin and subcutaneous tissue disorders	Pruritus	0.652
Skin and subcutaneous tissue disorders	Rash macular	0.651
Hepatobiliary disorders	Hyperbilirubinaemia	0.601
Skin and subcutaneous tissue disorders	Dry skin	0.600

Table 7.84. Trial EGF100151: Top 10 adverse events by posterior probability for BB_{21}.

Comparing Tables 7.84 and 4.2 we can see that the top 6 are the same, with slightly different orderings, but after that the posterior probabilities drop off. However the overall trend is similar to results from $1a_{21}$. We know from the simulation study in Chapter 5, and the results above, that point-mass models generally do not flag as many events as models without a point-mass and, in this case, the over-estimation of the time at risk for each patient will also have had a greater effect on the point-mass model as it will lower the estimated rates, so the fall off in posterior probability is not unexpected. Additionally, the fact that we have only one interval in the analysis means that there is no possibility of borrowing strength across the intervals. Overall the results are consistent with those in Chapter 4.

7.11 Conclusions

In this demonstration analysis we have looked at a number of possible different scenarios which could occur during the course of a clinical trial, as well as one real

world example. As for the simulation study in Chapter 5, care should be taken when drawing conclusions based on a small number of examples. Bearing this in mind, we can see that for point-mass models using a three-level hierarchy is not a suitable choice for these types of interim analyses. The proliferation of parameters and presence of the point-mass means that a very strong signal is required for the detection of adverse events with raised treatment rates. For two-level hierarchy point-mass models a lower threshold may be required to increase the power. For the data considered here the presence of the point-mass reduces the Type-I error rate dramatically compared to other models. An adverse event is either correctly flagged or not flagged at all, only rarely was an event incorrectly flagged.

For models without the point-mass the three-level and two-level hierarchies give similar results. For the data we looked at the two-level model was capable of adequately modelling the variation in the data. Here, the effect of missing out the point-mass is to increase the Type-I error rate. Where the point-mass models give a binary output, an event is generally correctly flagged or not flagged at all, the models without a point-mass correctly flag adverse events but with a corresponding number of Type-I errors which, in the demonstration analyses, are well controlled for models with common body-system means over the intervals (level 1 models).

Overall, the models which perform the best are those with a strong dependence between the body-system means over the trial. These are models $1a_{21}$, $1a_{31}$, and BB_{21}. For the same cut-off the models without a point-mass will correctly detect more adverse events than the point-mass models, but with higher Type-I error levels. The 1a models are also better at detecting events earlier in the trial, particularly for low frequency events (§7.8.3.1), which may be crucial in the early exploration of safety issues. The choice of which model to use, and the cut-off threshold, will likely be dependent on the type of data being analysed, how much data has accrued, the importance of detecting serious adverse events, the tolerance for error rates, and the acceptance or otherwise of using a lower threshold posterior probability when exploring the data. Finally, we again note that these are tentative conclusions from demonstration analyses which cover only a small number of trials and trial scenarios.

Chapter 8

Conclusions

Clinical trials are the standard method for establishing the efficacy and safety of new treatments. In this study we have looked at safety analysis in clinical trials with the main focus being the detection of adverse events. During the course of a clinical trial the occurrence and severity of patient adverse events are routinely recorded. The assessment of these adverse events is an important part of the safety analysis of a clinical trial and is also useful in establishing a safety profile for a new treatment. The statistical analysis of adverse events is complicated by two difficulties. Firstly, large numbers of different adverse events may be recorded during the trial. If a hypothesis testing approach is taken to compare between treatment arms, then there is the possibility of inflated Type-I error rates due to multiple comparisons. Secondly, many adverse events are expected to be rare. As trials are generally sized to answer the primary outcome, there is reduced power to detect differences in adverse rates between trial arms.

The International Council for Harmonisation (ICH) provides guidelines for many aspects of clinical trials. With regard to safety, the ICH recommends summary tables of the most common adverse events, along with comparisons of rates between the different trial arms (§1.4.3). Comparison methods may range from crude rate comparisons to more sophisticated techniques, such as survival analysis. While the ICH state (E3) that not every event need be the subject of a statistical evaluation, for longer-term treatment (§1.4.2) there should be some characterisation of the safety profile for the treatment. As there are many different types of adverse event, the ICH recommends that where a hypothesis testing approach is taken for comparing events, multiple comparison error controlling procedures should be used.

The conduct of a clinical trial is generally detailed in the trial protocol and this will include important aspects of the safety analysis to be performed. The protocol may also define certain adverse events as being of special interest. These may be

events which, based on the results of earlier trials or the clinical properties of the treatment, may be expected to occur over the course of the trial. Additionally, protocol defined serious events may require special attention, such as notification to a regulatory authority. While specific hypotheses for certain adverse events may be defined in the trial protocol (often referred to as Tier 1 events), this will not be the case for the majority of adverse events.

Given the expected rarity of adverse events, one approach to increasing the power to detect increased event rates is to use groupings of related adverse events into body-systems or system organ classes (§1.9), and use this additional relationship in a statistical analysis. Handling potential multiple comparison issues can be approached in a number of ways. In Chapter 2 we looked at a number of approaches to testing multiple hypotheses. In particular we looked at the Double False Discover Rate (DFDR) and Grouped Benjamini-Hochberg (GBH) procedures which are recent methods for error control when testing multiple hypotheses. These are based on the False Discovery Rate (FDR) but use grouping of related hypotheses to increase their power, under the appropriate assumptions, compared to the FDR. In Chapter 3 we looked at general approaches to adverse event modelling, including a number of models which used event groupings or analysed related events, including the Berry and Berry hierarchical model (§3.6.2) and Multivariate Bayesian Logistic Regression (MBLR) (§3.6.3). Many modelling approaches allow body-system adverse event groupings, either directly, or through random effects. Bayesian approaches to fitting models have a number of potentially attractive properties. The assumption of relationships in the data, such as body-system groupings, allows us to share information about the adverse events within the model (borrowing strength), and this is particularly important for the case of rare events. The Bayesian modelling approaches also contains the idea that, given the data and prior distributions, the posterior distributions tend towards the true distribution of the parameters, and that this in effect provides multiple comparison robustness [27], [66]. In particular, the Berry and Berry model implements these ideas [5], the hierarchical structure of the model reflects an assumed relationship between the adverse events, and, given the data, we should expect the posteriors to be close to the true parameter distributions, in effect controlling for multiple comparisons.

In Chapter 4 we illustrated these ideas by applying a number of the grouped methods to real trial data from GSK Trial EGF100151 (Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast

Cancer). The Berry and Berry models gave results which were very similar to those from applying Fisher exact tests to the adverse event data. However, there was a body-system effect, with *Localised infection* and *Back pain* missing from the top 10 adverse events by posterior probability for the point-mass model, and *Localised infection* missing from the top 10 adverse events for the model without a point-mass. The direct error controlling procedures flagged two events as significant at the 5% level, with both the DFDR and GBH flagging the same events. Both of these methods are affected by the presence of many very low count adverse events within some of the body-systems. Removing the 326 events which affected less than 1% of the patients resulted in more events being flagged. This idea of removing very low count events is explicitly considered to be part of the DFDR procedure ([3]), and the ICH guidelines also allow for the removal of low count adverse events (§1.4.4). For the Bayesian methods this type of step is not required.

In Chapter 5 a simulation study was used to investigate further how the methods compared with regard to flagging adverse events with raised treatment rates and overall error control. Here we found that, where a grouping structure exists within the data, the grouped methods performed better than unadjusted testing or error controlling procedures such as the Bonferroni correction or the Benjamini-Hochberg procedure. For small trials and low adverse event rates the Berry and Berry model without the point-mass (c212.1a) could be considered the best, correctly detecting more events, but with lower Type-I error rate than unadjusted testing. All the methods performed well for large trial sizes, although for c212.1a there were increased Type-I error rates. The point-mass plays an important role in the Berry and Berry models. When using the same posterior probability cut-off point for event flagging for both the model with point-mass (c212.BB) and without (c212.1a), the effect of the point-mass is to both reduce the numbers of correctly detected adverse events, and also the Type-I error rates. c212.BB performs best when the differences between control and treatment are large, and the trial size is large. In this case it is able to control both the Type-I and Type-II error rates. For smaller rate increases, c212.BB does not detect differences between treatment and control as well as c212.1a.

The methods compared in the simulation study are suitable for end of trial analysis and are not generally directly applicable for use at interim analyses. For short trials this may be acceptable, but for longer studies, where the Data Monitoring Committee may meet on a number of occasions, we would like to be able to say

something about the occurrences of adverse events at the times of the interim safety analyses. For analyses which occur early we may have unbalanced trial arms in terms of recruitment rates or time in study, and the adverse event rates themselves may vary over different periods of the trial as it progresses. For an approach that may be more useful at interim analyses we need to take into account the facts that the adverse event data accumulates over time, that the adverse event rates may change over time, and that, as there are multiple types of adverse events, we need some way of controlling for multiple comparisons. A Bayesian approach is natural way of handling accumulating data, and we have seen in the simulation study that, where applicable, a suitable choice of model and priors may help control for multiple comparisons. To do this we looked in particular at the Bayesian three-level hierarchy proposed by Berry and Berry [5].

The Berry and Berry model has a number of interesting features. The hierarchical approach, use of body-systems, and choice of priors provide a level of built-in multiple-comparison robustness to the model (§3.6). The data model is conditionally binomial and has a simple interpretation. The model also contains a point-mass term which explicitly takes into account the possibility of there being no difference between the treatment and control arms of the trial. With this in mind, we looked at conditional Poisson models for event occurrences which have a relatively simple interpretation in terms of average event rates (Chapter 6). The approach taken was to divide the trial time period into intervals, and use piecewise constant rates with a body-system hierarchy of random parameters to model the incidence (and total) adverse event occurrence. We considered models which treated each interval as independent, and models with relationships between the intervals. Models with and without a point-mass were investigated. These are summary level models which are in keeping with ICH guidelines. Patient level models were considered but not developed for reasons given in §6.4.1.

One consequence of hierarchical models is the large number of parameters needed to define the models, and splitting the trial duration into intervals greatly increases the number of parameters needed compared to the end of trial models. This over-parameterisation of the model has the potential to affect a model's usefulness, so, for each model considered, we also included models where we removed the lowest level of the hierarchy, leaving an equivalent two-level hierarchical model (§6.7). The Bayesian approach to fitting the models allows accumulating data to be handled in a straightforward manner. However, unlike error controlling procedures, the modelling approach does not come with a definitive method for flagging an adverse

event. In the Berry and Berry model the posterior probability of an increased event risk is used to determine if an event is to be flagged and we follow this approach for the interim analysis Poisson models.

To illustrate the methods a number of demonstration analyses were performed on simulated trial data (Chapter 7). The demonstration analyses consisted of a number of trials with differing event rates between treatment and control for a single body-system. In order to flag an adverse event as significant, we used threshold or cut-off points for the posterior probability that the increase in rate of event occurrence on the treatment arm is positive. The choice of cut-off point has an effect on the results. For the same cut-off point models with and without the point-mass give different results. Due to the presence of the point-mass a lower cut-off point was needed to give broadly similar numbers of correctly identified adverse events to the models without a point-mass (§7.8.1). We did not see an inflation of the Type-I error rate when this was done. As our approach has been to develop methods which may be suitable for use at an interim analysis in an exploratory sense (§1.10), rather than trying to define trial stopping-rules, the determination of a suitable cut-off point was not part of the study.

Overall, we found that for the interim analyses the models that performed best had a stronger dependence between the body-system means over the duration of the trial, and the use of the body-system approach, when applicable to the data, improved the performance of the models in terms of adverse event detection. The three-level hierarchy model with the point-mass was not suitable for the type of interim analyses we performed. The number of parameters, along with the presence of the point-mass, made detection of significant adverse events difficult, especially early in the trial. For the models without the point-mass, we did not find any great difference between the two-level and three-level hierarchies in terms of performance. The two-level hierarchy was capable of accounting for the variation in the data we considered. For lower event rates, the models without a point-mass were able to flag adverse events correctly much earlier in the trial than those with a point-mass. This was particularly noticeable for severity 3+ events. In one case it was day 1080 of the trial before any of the point-mass models were capable of flagging an adverse event as having a raised rate on the treatment arm (Table 7.73). This alone may make the models without a point-mass more suitable for early interim analyses. We also looked at the effect of missing data on the model results. Here we assumed that a patient who had a serious adverse event within a certain number of days of an upcoming interim analysis would be censored (at

that analysis). Under our assumptions the models held up quite well. In this case, the relatively low rates of event occurrence meant that not enough events were removed from the higher rate trial arm to affect detection greatly.

Model inferences about adverse event rates are based on MCMC simulations for the models. Within each MCMC chain the generated samples are correlated, and inference based on correlated samples may be less precise than those drawn independently, but at convergence this is generally not a problem [122]. However, the correlation may cause inefficiencies in the simulations, leading to slow convergence. In our simulations and demonstration analyses we chose not to thin the samples. Although the models have large numbers of parameters, we did have enough memory to store all generated simulations. Once the chains have reached approximate convergence they can be used directly for inference about the various model parameters, regardless of whether they have been thinned or not, and this is the approach we followed [122].

All of the methods used in the study have been implemented in the `R` package `c212` [139]. The `coda` package was used for convergence checking (Gelman-Rubin statistic), and a number of other post sampling tasks [148]. Overall, the models without a point-mass run more quickly than their point-mass counterparts, and generally appear to converge faster, with less simulation parameter tuning required. Tuning the simulations for the point-mass models may not be straightforward. For each non-standard distribution there are one (for Metropolis-Hastings sampling) or two (for slice sampling) parameters which control how the samples are generated and, in Metropolis-Hastings sampling, the acceptance rates. Consideration needs to be given to the size of any potential treatment differences and any target acceptance rates when adjusting the parameters. In order to check the convergence of the model all samples generated must be retained. This requires a large amount of memory, particularly if multiple chains are run. The software does include the possibility of just retaining certain families of parameters if there are memory constraints on the system on which the software is being executed. Potential improvements to the software include a step to identify suitable simulation parameter values based on the acceptance rates and convergence diagnostics (auto-tuning), the ability to retain only individual model parameters rather than whole families, bringing improved memory performance, and the implementation of a number of the convergence diagnostics as native methods in `c212`, thereby improving the overall runtime. The sampling approaches implemented are Metropolis-Hastings and slice sampling within a Gibbs sampler, and these worked well with the data we

considered. However a number of alternatives approaches, such as reversible jump MCMC methods [124], were investigated but not implemented in the current software release. We also investigated methods by Ji and Schmidler who introduced a number of what are termed Adaptive Metropolized Independence Samplers, which include methods suitable for distributions which contain point-masses, and which are claimed to produce samples with very little auto-correlation [149].

We may also consider what further work can be achieved with regard to the current models. The possibility exists to explore further data scenarios in order to identify when a three-level model is more useful than a two-level model. We have seen for our data, with its limited number of simulations, that there is very little difference between the 1a level 1 models and investigating this further would be of interest. We could also look to investigate further the differences between similar models. The models without a point-mass are nested within an equivalent point-mass model, and one interesting aspect of the study is the difference the point-mass makes between the two similar models in terms both of error control and event detection. For small trials and low rates the point-mass models have very low detection rates, while for large differences between treatment and control the point-mass models are capable of identifying the adverse events with very low error rates. In effect, the point-mass acts as a barrier which must be overcome in order for an event to be considered as having an increased treatment arm rate. The models without a point-mass have a different effect, they may correctly detect more events but have an increased error rate. It appears, for the data studied, that point-mass models have possibly too high a barrier for very rare events, but the models without a point-mass have a higher than hoped for error rate. One approach to this is to lower the threshold for flagging an event in point-mass models, and we have seen that this may be possible without inflating the error rates (§7.8.1). However, alternative approaches to the models would also be worth considering, while keeping the basic underlying body-system structure. We could reduce the point-mass parameters to a single overall trial parameter and see how this affects the model performance. Also, given that we are potentially looking for a model which is in some way between one with a point-mass and without a point-mass, we could consider the effect of Bayesian Model Averaging (BMA) on our results [150]. In this approach we have a number of different possible models for the data, and the posterior distribution of the effect of interest is an average of its posterior distribution under each of the models considered, weighted by the posterior probability of the these models. Here, for example, we could assume that the effect of

interest is the flagging of an adverse event as having a raised treatment event rate, and it would be interesting to compare the results under this approach.

The models we have considered are summary conditional Poisson models, where the rates do not apply to the individual patients in the study, but are overall rates on the individual trial arms. The point-mass term is used to model the possibility of no differences between the trial arms. As an alternative, in order to handle the non-increase of events on the treatment arms, or general low or zero event counts, we could look again at patient level models. In §6.4.1 we considered that these models have the potential to very complicated. However, they do allow the possibility of using zero-inflated Poisson (ZIP) or hurdle models ([151]) to handle zero counts at the patient level, and a prudent choice of a limited number of model parameters, while still maintaining a body-system relationship, may allow potentially interesting models to be investigated.

Finally, we consider how the methods may be used over the course of a clinical trial. For the Data Monitoring Committee (DMC) (§1.3) ensuring the safety of patients in the trial is of paramount importance [2]. The DMC must expect unforeseen adverse events and must be prepared to alter its procedures in response. Once a trial is under way the DMC will meet regularly to look at the accumulating data, including adverse events. The DMC must decide if the adverse events reported are such that continuing the trial cannot be justified, even if these are not statistically significant, or have not been specified as of special interest in the trial protocol [2]. Making these decisions is not a clear cut process. This is often a medical or ethical judgement rather than a clear statistical decision. Stopping a trial too early or too late can cause patient harm. In addition to this, repeated looks at the data leads to an increase in the possibility of seeing a significant result by chance, and the potentially large numbers of adverse events recorded during a trial complicates the statistical analysis. The methods developed in this project are designed to help the DMC achieve these safety goals. A Bayesian framework ensures that accumulating data is naturally handled and the body-system based hierarchy ensures that multiple types of adverse events are catered for. The methods as implemented (Appendix A) are simple to apply to the accumulating data and the study statistician, or the statistical centre analysing the trial data, could include the model outputs or any potentially interesting adverse events indicated by the models, in the presentation to the DMC at the relevant interim analysis meeting. The DMC could then choose to consider this additional information when making any decisions about the future conduct of the trial.

Appendix A

Software Implementations and Model Fitting Algorithms

Unless otherwise stated, all the methods and models used in this study are implemented in the `c212` package for `R` [139]. Table A.1 gives the methods and their corresponding `R` functions:

Method Name[1]	`R` **Function**	**Description**
NOADJ	c212.NOADJ	No error controlling procedure.
BONF	c212.BONF	Bonferroni correction [12].
BH	c212.BH	Benjamini-Hochberg procedure [31].
DFDR	c212.DFDR	Double false discovery rate [3].
GBH	c212.GBH	Group Benjamini-Hochberg [30].
ssBH	c212.ssBH	Subset Benjamini-Hochberg [57].
c212.BB	c212.BB	Berry and Berry model [5].
c212.1a	c212.1a	Berry and Berry model without point-mass [5], [60].
$1a_{2l}$	c212.interim.1a.hier2	Two-level-hierarchy, no point-mass (§6.5.3).
$1a_{3l}$	c212.interim.1a.hier3	Three-level-hierarchy, no point-mass (§6.5.3).
BB_{2l}	c212.interim.BB.hier2	Two-level-hierarchy, point-mass (§6.5.3).
BB_{3l}	c212.interim.BB.hier3	Three-level-hierarchy, point-mass (§6.5.3).

Table A.1. Methods implemented in the package `c212`.

[1] The subscript l on the 1a and BB models refers to the model dependency level (Table 6.1).

A.1 MCMC Approach to Bayesian Model Fitting

The Bayesian models are fitted using a Gibbs sampling Markov Chain Monte Carlo (MCMC) method [124]. This approach, in which the complete conditionals are sampled, is particularly useful for decomposing high-dimensional target distributions into smaller targets for sampling. The joint posterior distributions and complete conditionals for the Bayesian models are given in Appendix B.

A.2 MCMC Sampling Algorithms

Many of the complete conditionals are from standard distributions and may be sampled directly in `R`. Within the Gibbs sampler, for non-standard distributions we either use a Metropolis-Hastings (MH) ([124]) or a slice sampler ([152]) step.

The presence of the point-mass in a number of the models makes the implementation of the sampler slightly difficult as the complete conditional distribution of the θ parameters will contain a point-mass, and this means that any proposal distribution used in an MH step will need a similar term. Technically the proposal must be absolutely continuous with respect to the dominating measure, a mixture of the point-mass at zero and the Lebesgue measure. The approach taken in [5] is to use a mixture proposal distribution consisting of a point-mass at zero, with probability weighting 0.5, and a normal distribution centred on the current value. We follow this as the default approach for θ in the point-mass models. For the non-standard distributions without a point-mass, we use either an MH step with a normal distribution centred on the current value as the proposal distribution, or a slice sampler. For the MH steps the variance (σ^2_{MH}) must be sepcified, and may be tuned if necessary. For slice sampling a width parameter (w) is required and also, if chosen, a control parameter (m), which limits the distance the algorithm traverses within the domain of the distribution from the current location when looking for the next sample [124]. Table A.2 lists the parameters with non-standard complete conditional distributions and their implemented sampling steps. The default numbers of chains, burn-in period, and total iterations in each chain are given in Table A.3.

Method Name	Parameter Family	Sampler
c212.BB	θ	MH
c212.BB	$\gamma, \alpha_\pi, \beta_\pi$	Slice or MH
c212.1a	θ, γ	Slice or MH
$1a_{2l}$	θ, γ	Slice or MH
$1a_{3l}$	θ, γ	Slice or MH
BB_{2l}	θ	MH
BB_{2l}	γ	Slice or MH
BB_{3l}	θ	MH
BB_{3l}	$\gamma, \alpha_\pi, \beta_\pi$	Slice or MH

Table A.2. Model parameters with non-standard distributions.

Method Name	Parallel Chains	Burn-in	Total Iterations
c212.BB	3	20000	60000
c212.1a	3	10000	40000
$1a_{2l}$	3	10000	40000
$1a_{3l}$	3	10000	40000
BB_{2l}	5	20000	60000
BB_{3l}	5	20000	60000

Table A.3. Model MCMC defaults.

A.3 Default Simulation Parameter Values

The global parameter values required by the MH and SLICE steps, used in the fitting algorithms for the simulation study and demonstration analyses in Chapters 5 and 7 respectively, are given in Tables A.4 and A.5. The choice of default parameter values was guided by their function within the MCMC simulation, and expectations of the input data. The σ_{MH} and w parameters control the exploration of the target distribution and, for the input data we are dealing with, we expect that non-standard distributions may lie close to some underlying true parameter value. The value of defaults for σ_{MH} and w were chosen to be relatively small to reflect this, and then tuned further over a number of different data sets. The m parameter also reflects the overall width of the distribution, and the relatively large value chosen should ensure that the distribution is explored as far as possible.

Model	Model Parameters	Sampling Parameter	Value
c212.1a (SLICE)	γ_{bj}, θ_{bj}	w	1.00
c212.1a (SLICE)	γ_{bj}, θ_{bj}	m	6.00
c212.1a (MH)	γ_{bj}, θ_{bj}	σ_{MH}	0.35
c212.BB	θ_{bj}	σ_{MH}	0.20
c212.BB (SLICE)	γ_{bj}	w_γ	1.00
c212.BB (SLICE)	γ_{bj}	m_γ	6.00
c212.BB (SLICE)	α_π	w_α	1.00
c212.BB (SLICE)	α_π	m_α	6.00
c212.BB (SLICE)	β_π	w_β	1.00
c212.BB (SLICE)	α_π	m_α	6.00
c212.BB (MH)	γ_{bj}	σ_{MH}	0.20
c212.BB (MH)	α_π	σ_{MH}	3.00
c212.BB (MH)	β_π	σ_{MH}	3.00

Table A.4. Global MCMC parameters for end of trial models.

Model	Model Parameters	Sampling Parameter	Value
$\text{BB}_{2\text{l}}$ (2-level hierarchy)	$\theta_{bj,h}$	σ_{MH}	0.5
$\text{BB}_{3\text{l}}$ (3-level hierarchy)	$\theta_{bj,h}$	σ_{MH}	0.25
BB_{hl} (all models)	$\gamma_{bj,h}$	w_γ	1.0
BB_{hl} (all models)	$\gamma_{bj,h}$	m_γ	6.0
BB_{30}, BB_{31}, BB_{32}	$\alpha_{\pi,h}$, α_π	w_α	1.0
BB_{30}, BB_{31}, BB_{32}	$\alpha_{\pi,h}$, α_π	m_α	6.0
BB_{30}, BB_{31}, BB_{32}	$\beta_{\pi,h}$, β_π	w_β	1.0
BB_{30}, BB_{31}, BB_{32}	$\beta_{\pi,h}$, β_π	m_β	6.0
1a_{hl} (all models)	$\gamma_{bj,h}, \theta_{bj,h}$	w	1.0
1a_{hl} (all models)	$\gamma_{bj,h}, \theta_{bj,h}$	m	6.0

Table A.5. Global MCMC parameters for the interim analysis models.

A.4 Initial Values for MCMC Chains

The model parameters are defined in §3.6.2 for the end of trial methods, and in §6.6 and §6.7 for the interim analysis methods. The Gibbs sampler requires an initial value for each parameter in the model. If these are not supplied by the user then initial values are generated by the software.

A.4.1 Top-Level Parameters

The top-level parameters are used to model the trial data and the initial values are derived from the data for the first chain in the MCMC simulation.

A.4.1.1 End of Trial Methods

The general method we use for calculating the initial values of the top-level parameters $(\gamma_{bj}, \theta_{bj})$, for a given set of trial data (X_{bj}, Y_{bj}) (§3.6.2), is:

1. $C_{bj} = \begin{cases} \frac{X_{bj}}{N_C} & X_{bj} \neq 0 \\ \frac{1}{N_C} & X_{bj} = 0 \\ \frac{N_C - 1}{N_C} & X_{bj} = N_C \end{cases}$

2. $T_{bj} = \begin{cases} \frac{Y_{bj}}{N_T} & Y_{bj} \neq 0 \\ \frac{1}{N_T} & Y_{bj} = 0 \\ \frac{N_T - 1}{N_T} & Y_{bj} = N_T \end{cases}$

3. $\gamma_{bj} = \log \frac{C_{bj}}{1 - C_{bj}}$

4. $\theta_{bj} = \log \frac{T_{bj}}{1 - T_{bj}} - \gamma_{bj}$

Table A.6 summarises the approach.

Model Parameters	First Chain	Subsequent Chains
γ_{bj}, θ_{bj}	Use steps 1 to 4 above with the trial data.	Use steps 1 to 4 above where X_{bj} and Y_{bj} are sampled with replacement from $0, 1, ..., N_C$ and $0, 1, ..., N_T$ respectively.

Table A.6. End of trial methods: top-level initial value generation for MCMC simulation.

A.4.1.2 Interim Analysis Methods

The approach for the first chain is similar to that for the end of trial (§A.4.1.1). For a given set of trial data (§6.5.1) we proceed as follows:

1. $L_{bj,h}^{(1)} = \frac{X_{bj,h}^{(1)}}{T_{bj,h}^{(1)}}$
2. $L_{bj,h}^{(2)} = \frac{X_{bj,h}^{(2)}}{T_{bj,h}^{(2)}}$
3. $\gamma_{bj,h} = \begin{cases} \log\left(L_{bj,h}^{(1)}\right) & L_{bj,h}^{(1)} \neq 0 \\ -10 & L_{bj,h}^{(1)} = 0 \end{cases}$
4. $\theta_{bj,h} = \begin{cases} \log\left(L_{bj,h}^{(2)}\right) - \gamma_{bj,h} & L_{bj,h}^{(1)} \neq 0 \\ & L_{bj,h}^{(2)} \neq 0 \\ -10 & \text{otherwise} \end{cases}$

Table A.7 summarises the approach.

Model Parameters	First Chain	Subsequent Chains[1]
$\gamma_{bj,h}$, $\theta_{bj,h}$	Use steps 1 to 4 above with the trial data.	Sample from $\mathrm{U}(-10, 10)$.

Table A.7. Interim analysis methods: top-level parameter initial value generation for MCMC simulation.

[1] $\mathrm{U}(a, b)$ is the continuous uniform distribution on $[a, b]$.

A.4.2 Hyperparameters

Tables A.8 and A.9 describe how the initial values for the model hyperparameters are generated for the different chains in the MCMC simulation.

Model Parameters	First Chain	Subsequent Chains
$\mu_{\gamma 0}$	0	Sample from U$(-50, 50)$.
$\tau^2_{\gamma 0}$	10	Sample from U$(5, 20)$.
$\mu_{\theta 0}$	0	Sample from U$(-50, 50)$.
$\tau^2_{\theta 0}$	10	Sample from U$(5, 20)$.
$\mu_{\gamma b}$	0	Sample from U$(-50, 50)$.
$\sigma_{\gamma b}$	10	Sample from U$(5, 20)$.
$\mu_{\theta b}$	0	Sample from U$(-50, 50)$.
$\sigma_{\theta b}$	10	Sample from U$(5, 20)$.
α_π	1.5	Sample from U$(1.25, 100)$.
β_π	1.5	Sample from U$(1.25, 100)$.
π_b	0.5	Sample from U$(0, 1)$.

Table A.8. End of trial methods: hyper-parameter initial value generation for MCMC simulation.

Model Parameters	First Chain	Subsequent Chains
$\mu_{\gamma 0,h}$, $\mu_{\gamma 0}$	0	Sample from U$(-10, 10)$
$\tau^2_{\gamma 0,h}$, $\tau^2_{\gamma 0}$	10	Sample from U$(5, 20)$
$\mu_{\theta 0,h}$, $\mu_{\theta 0}$	0	Sample from U$(-10, 10)$
$\tau^2_{\theta 0,h}$, $\tau^2_{\theta 0}$	10	Sample from U$(5, 20)$
$\mu_{\gamma b,h}$, $\mu_{\gamma b}$	0	Sample from U$(-10, 10)$
$\sigma_{\gamma b,h}$, $\sigma_{\gamma b}$	10	Sample from U$(5, 20)$
$\mu_{\theta b,h}$, $\mu_{\theta b}$	0	Sample from U$(-10, 10)$
$\sigma_{\theta b,h}$, $\sigma_{\theta b}$	10	Sample from U$(5, 20)$
$\alpha_{\pi,h}$, α_π	1.5	Sample from U$(1.25, 100)$
$\beta_{\pi,h}$, β_π	1.5	Sample from U$(1.25, 100)$
$\pi_{b,h}$, π_b	0.5	Sample from U$(0, 1)$

Table A.9. Interim analysis methods: hyper-parameter initial value generation for MCMC simulation.

A.5 Convergence Diagnostics and Summary Statistics

The main convergence diagnostic available within the package, and used to assess convergence in this study, is the Gelman-Rubin statistic ([122]), discussed in more detail in Appendix C. Summary statistics and Highest Probability Intervals (HPI) are available for the posterior distributions.

A.6 Inverse Gamma Distribution

The joint posterior distribution and complete conditional distributions for the parameters of the model c212.BB are given in the appendix of [5] as well as a description of the simulation algorithms used in the paper. In our implementation we use an inverse-gamma distribution, $\mathrm{IG}(\alpha, \beta)$, where β is a scale parameter. This has the following density function [122]:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

In the original model description in [5] Berry and Berry use $\frac{1}{\beta}$ as the scale parameter with corresponding density function:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{1}{\beta x}\right)$$

With the default parameters from [5] (§3.21), all the scale parameters for the inverse-gamma distributions are set to 1 and there is no difference between the distributions.

A.7 Direct Error Controlling Methods

The error controlling methods in Table A.1 are purely deterministic and the translation of the methods to `R` code is straightforward, with each method fully specified in the paper which introduced it. A common interface to each function was used with the data supplied to the function required to be in a tabular format specified either in a data frame or a file. Each function returns the subset of the data passed in which corresponds to the hypotheses deemed significant.

A.8 BUGs Models

The BUGS modelling language and related software (WinBUGS, OpenBUGS) provide a generic approach to implementing Bayesian models using Gibbs sampling [126]. The models below are those provided by [60].

A.8.1 c212.BB

```
model{
    for (i in 1:Nae) {
        X[i] ~ dbin(c[b[i], j[i]], Nc)
        Y[i] ~ dbin(t[b[i], j[i]], Nt)
        logit(c[b[i], j[i]]) <- gamma[b[i], j[i]]
        logit(t[b[i], j[i]]) <- gamma[b[i], j[i]] + theta[b[i], j[i]]

        gamma[b[i], j[i]] ~ dnorm(mu.gamma[b[i]], tau.gamma [b[i]])
        p0[i] ~ dbern(pi[b[i]]) # prob of point mass
        theta1[b[i], j[i]] ~ dnorm(mu.theta[b[i]], tau.theta[b[i]])
        theta[b[i], j[i]] <- (1- p0[i]) * theta1[b[i], j[i]]

        OR[b[i],j[i]] <- exp(theta[b[i],j[i]] )
        PGT0[b[i], j[i]] <- 1 - step(0 - theta[b[i],j[i]])

        D[i]<-X[i]*log(c[b[i], j[i]])+(Nc-X[i])*log(1-c[b[i], j[i]])
                + Y[i]*log(t[b[i], j[i]])+(Nt- Y[i])*log(1-t[b[i],j[i]])

    }
    Dbar<- -2* sum(D[]) # -2logL without normalizing constant

    # SOC level parameters
    for(k in 1:B) {
        pi[k] ~ dbeta(alpha.pi, beta.pi)
        mu.gamma[k] ~ dnorm(mu.gamma.0, tau.gamma.0)
        tau.gamma[k] ~ dgamma(3,1)
        mu.theta[k] ~ dnorm(mu.theta.0, tau.theta.0)
        tau.theta[k] ~ dgamma(3,1)
    }
    mu.gamma.0 ~ dnorm(0, 0.1)
    tau.gamma.0 ~ dgamma(3,1)
    mu.theta.0 ~ dnorm(0, 0.1)
    tau.theta.0 ~ dgamma(3,1)
    alpha.pi ~ dexp(1.0) I(1, )
    beta.pi ~ dexp(1.0) I(1, )
}
```

Figure A.1. c212.BB BUGs model.

A.8.2 c212.1a

```
model{
    for (i in 1:Nae) {
        X[i] ~ dbin(c[b[i], j[i]], Nc)
        Y[i] ~ dbin(t[b[i], j[i]], Nt)
        logit(c[b[i], j[i]]) <- gamma[b[i], j[i]]
        logit(t[b[i], j[i]]) <- gamma[b[i], j[i]] + theta[b[i], j[i]]
        gamma[b[i], j[i]] ~ dnorm(mu.gamma[b[i]], tau.gamma[b[i]])
        theta[b[i], j[i]]~ dnorm(mu.theta[b[i]], tau.theta[b[i]])
        OR[b[i],j[i]] <- exp(theta[b[i],j[i]])
        PGT0[b[i], j[i]] <- 1 - step(0 - theta[b[i],j[i]])
    }

    for(k in 1:B){
        mu.gamma[k] ~ dnorm(mu.gamma.0, tau.gamma.0)
        tau.gamma[k] ~ dgamma(alpha.gamma, beta.gamma)
        mu.theta[k] ~ dnorm(mu.theta.0, tau.theta.0)
        tau.theta[k] ~ dgamma(alpha.theta, beta.theta)
    }

    mu.gamma.0 ~ dnorm(mu.gamma.0.0, tau.gamma.0.0)
    tau.gamma.0 ~ dgamma(alpha.gamma.0.0, beta.gamma.0.0)
    mu.theta.0 ~ dnorm(mu.theta.0.0, tau.theta.0.0)
    tau.theta.0 ~ dgamma(alpha.theta.0.0, beta.theta.0.0)

    # Hyperparameters
    mu.gamma.0.0 <- 0    #
    tau.gamma.0.0 <- 0.1 #
    alpha.gamma.0.0 <- 3 #
    beta.gamma.0.0 <- 1 #
    mu.theta.0.0 <- 0     #
    tau.theta.0.0 <- 0.1   #
    alpha.theta.0.0 <- 3   #
    beta.theta.0.0 <- 1 #
    alpha.gamma <- 3
    beta.gamma <- 1
    alpha.theta <- 3
    beta.theta <- 1
}
```

Figure A.2. c212.1a BUGs model.

Appendix B

Joint Distributions and Complete Conditional Distributions

B.1 Distributions

Table B.1 lists the probability distributions used in this study.

Name	Description
$\mathrm{U}(a, b)$	Continuous uniform distribution on [a,b] [122].
$\mathrm{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2 [122].
$\mathrm{Gamma}(\alpha, \beta)$	Gamma dsistribution [122].
$\mathrm{IG}(\alpha, \beta)$	Inverse-Gamma distribution (§A.6).
$\mathrm{Poisson}(\lambda)$	Poisson distribution with rate λ [122].
$\mathrm{Beta}(a, b)$	Beta distribution [122].
$\mathrm{Bin}(n, p)$	Binomial distribution [122].

Table B.1. Probability distributions.

B.2 General Results

Many of the complete conditionals for the models in this study may be derived from known results about conjugate priors. We use the following general results in the derivation of the complete conditionals for the Bayesian models used in the study:

Result 1. *([122]) If* $X_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, ..., N$, *and* $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, *then*

$$\mu \Big| X_1, \ldots, X_N \sim \mathrm{N}\left(\frac{\mu_0 \sigma^2 + \sigma_0^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_0^2}, \frac{\sigma^2 \sigma_0^2}{\sigma^2 + N\sigma_0^2} \right)$$

Equivalently, in terms of density functions we have:

$$\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{i=1}^{N}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
$$\propto \exp\left[-\frac{1}{2}\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+N\sigma_0^2}\right)^{-1}\left(\mu-\frac{\left(\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N}x_i\right)}{(\sigma^2+N\sigma_0^2)}\right)\right]$$

Result 2. *If* $X_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(\mu,\sigma^2)$, $i=1,...,N,\ j=1,...,k_i$, *and* $\mu \sim \mathrm{N}(\mu_0,\sigma_0^2)$, *then*

$$\mu\Big|X_{11},...,X_{N,k_N} \sim \mathrm{N}\left(\frac{\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N}\sum_{j=1}^{k_i}x_{ij}}{\sigma^2+\sigma_0^2\sum_{i=1}^{N}k_i}, \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2\sum_{i=1}^{N}k_i}\right)$$

This is a straightforward generalisation of Result 1.

Result 3. *If* $X_i \overset{i.i.d.}{\sim} \left[p\,\mathrm{I}_{[x_i=0]}+(1-p)\,\mathrm{I}_{[x_i\neq 0]}\,\mathrm{N}(\mu,\sigma^2)\right]$, $i=1,...,N$, *and* $\mu \sim \mathrm{N}(\mu_0,\sigma_0^2)$, *then*

$$\mu\Big|X_1,\ldots,X_N \sim \mathrm{N}\left(\frac{\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N}x_i}{\sigma^2+K_N\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2+K_N\sigma_0^2}\right)$$

where $K_N=\sum_{i=1}^{N}\mathrm{I}_{[x_i\neq 0]}$. *Equivalently, in terms of density functions we have:*

$$\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{i=1}^{N}\left[p\,\mathrm{I}_{[x_i=0]}+(1-p)\,\mathrm{I}_{[x_i\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]$$
$$\propto \exp\left[-\frac{1}{2}\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+N_K\sigma_0^2}\right)^{-1}\left(\mu-\frac{\left(\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N}x_i\right)}{(\sigma^2+N_K\sigma_0^2)}\right)\right]$$

Proof. The distribution of $\mu \Big| X_1, \ldots, X_N$ is proportional to:

$$\begin{aligned}
&\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{i=1}^{N}\left[p\,\mathrm{I}_{[x_i=0]}+(1-p)\,\mathrm{I}_{[x_i\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]\\
&=\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{i=1}^{N}\left[p^{\mathrm{I}_{[x_i=0]}}\left[(1-p)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^{\mathrm{I}_{[x_i\neq 0]}}\right]\\
&\propto\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{i=1}^{N}\left[\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^{\mathrm{I}_{[x_i\neq 0]}}\\
&=\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{\substack{i=1\\x_i\neq 0}}^{N}\left[\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]
\end{aligned}$$

Denoting the non-zero values of $\{x_i\}_{i=1}^{N}$ by $\{y_i\}_{i=1}^{N_K}$ and applying Result 1 with y_i and $N = K_N$ we have:

$$\mu\Big| X_1, \ldots, X_N \sim \mathrm{N}\left(\frac{\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{K_N} y_i}{\sigma^2+K_N\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2+K_N\sigma_0^2}\right)$$

As

$$\sum_{i=1}^{K_N} y_i = \sum_{\substack{i=1\\x_i\neq 0}}^{N} x_i = \sum_{i=1}^{N} x_i$$

we have:

$$\mu\Big| X_1, \ldots, X_N \sim \mathrm{N}\left(\frac{\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N} x_i}{\sigma^2+K_N\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2+K_N\sigma_0^2}\right) \tag{B.1}$$

□

Result 4. *If* $X_{ij} \overset{i.i.d.}{\sim} p\,\mathrm{I}_{[x_{ij}=0]}+(1-p)\,\mathrm{I}_{[x_{ij}\neq 0]}\,\mathrm{N}(\mu,\sigma^2)$, $i = 1, ..., N$, $j = 1, ..., k_i$ *and* $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, *then*

$$\mu\Big| X_{11}, ..., X_{N,k_N} \sim \mathrm{N}\left(\frac{\mu_0\sigma^2+\sigma_0^2\sum_{i=1}^{N}\sum_{j=1}^{k_i} x_{ij}}{\sigma^2+\sigma_0^2\sum_{i=1}^{N} K_i}, \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2\sum_{i=1}^{N} K_i}\right)$$

where $K_i = \sum_{j=1}^{k_i}\mathrm{I}_{[x_{ij}\neq 0]}$.

This is a straightforward generalisation of Result 3.

Result 5. *([122]) If* $X_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, ..., N$, *and* $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$, *then*

$$\sigma^2 \Big| X_1, \ldots, X_N \sim \mathrm{IG}\left(\alpha + \frac{N}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

Result 6. *If* $X_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, ..., N$, $j = 1, ..., k_i$ *and* $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$, *then*

$$\sigma^2 \Big| X_{11}, \ldots, X_{N,k_N} \sim \mathrm{IG}\left(\alpha + \frac{\sum_{i=1}^{N} k_i}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{k_i}(x_{ij} - \mu)^2\right)$$

This is a straightforward generalisation of Result 5.

Result 7. *If* $X_i \overset{i.i.d.}{\sim} p\,\mathrm{I}_{[x_i=0]} + (1-p)\,\mathrm{I}_{[x_i\neq 0]}\,\mathrm{N}(\mu, \sigma^2)$, $i = 1, ..., N$ *and* $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$. *Then*

$$\sigma^2 \Big| X_1, \ldots, X_N \sim \mathrm{IG}\left(\alpha + \frac{K_N}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N}\mathrm{I}_{[x_i\neq 0]}(x_i - \mu)^2\right)$$

where $K_N = \sum_{i=1}^{N}\mathrm{I}_{[x_i\neq 0]}$.

Proof. The distribution of $\sigma^2 \Big| X_1, \ldots, X_N$ is proportional to:

$$\begin{aligned}
&\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{(\sigma^2)}\right)\prod_{i=1}^{N}\left[p\,\mathrm{I}_{[x_i=0]} + (1-p)\,\mathrm{I}_{[x_i\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \\
&\propto \left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{(\sigma^2)}\right)\prod_{i=1}^{N}\left[p^{\mathrm{I}_{[x_i=0]}}\left[(1-p)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^{\mathrm{I}_{[x_i\neq 0]}}\right] \\
&\propto \left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{(\sigma^2)}\right)\prod_{i=1}^{N}\left[\frac{1}{\sqrt{\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]^{\mathrm{I}_{[x_i\neq 0]}}
\end{aligned} \tag{B.2}$$

Denoting the non-zero values of $\{x_i\}_{i=1}^{N}$ by $\{y_i\}_{i=1}^{N_K}$ and applying Result 5 with y_i and $N = K_N$ we have:

$$\sigma^2 \Big| X_1, \ldots, X_N \sim \mathrm{IG}\left(\alpha + \frac{K_N}{2}, \beta + \frac{1}{2}\sum_{i=1}^{K_N}(y_i - \mu)^2\right)$$

which may be written as:

$$\sigma^2 \Big| X_1, \ldots, X_N \sim \text{IG}\left(\alpha + \frac{K_N}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N} \text{I}_{[x_i \neq 0]}(x_i - \mu)^2\right)$$

□

Result 8. *If* $X_{ij} \overset{i.i.d.}{\sim} p\,\text{I}_{[x_{ij}=0]} + (1-p)\,\text{I}_{[x_{ij}\neq 0]}\,\text{N}(\mu, \sigma^2)$, $i = 1, ..., N$, $j = 1, ..., k_i$ *and* $\sigma^2 \sim \text{IG}(\alpha, \beta)$, *then*

$$\sigma^2 \Big| X_{11}, \ldots, X_{N,k_N} \sim \text{IG}\left(\alpha + \frac{\sum_{i=1}^{N} K_i}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{k_i} \text{I}_{[x_{ij}\neq 0]}(x_{ij} - \mu)^2\right)$$

where $K_i = \sum_{j=1}^{N} \text{I}_{[x_{ij}\neq 0]}$.

This is a straightforward generalisation of Result 7.

B.3 Model c212.1a

From §3.6.2.2 we have:

$$\begin{aligned} c_{bj} &= \frac{e^{\gamma_{bj}}}{1 + e^{\gamma_{bj}}} \\ t_{bj} &= \frac{e^{\theta_{bj} + \gamma_{bj}}}{1 + e^{\theta_{bj} + \gamma_{bj}}} \end{aligned}$$

and the probability function for X_{bj} is proportional to:

$$\begin{aligned} c_{bj}^{x_{bj}}(1 - c_{bj})^{N_C - x_{bj}} &= \left(\frac{e^{\gamma_{bj}}}{1 + e^{\gamma_{bj}}}\right)^{x_{bj}} \left(1 - \frac{e^{\gamma_{bj}}}{1 + e^{\gamma_{bj}}}\right)^{N_C - x_{bj}} \\ &= \left(\frac{e^{\gamma_{bj}}}{1 + e^{\gamma_{bj}}}\right)^{x_{bj}} \left(\frac{1}{1 + e^{\gamma_{bj}}}\right)^{N_C - x_{bj}} \\ &= \left(e^{\gamma_{bj}}\right)^{x_{bj}} \left(\frac{1}{1 + e^{\gamma_{bj}}}\right)^{N_C} \\ &= \frac{e^{\gamma_{bj} x_{bj}}}{\left(1 + e^{\gamma_{bj}}\right)^{N_C}} \end{aligned}$$

Similarly for Y_{bj} we have:

$$t_{bj}^{y_{bj}}(1-t_{bj})^{N_T-y_{bj}} = \frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}}$$

The joint posterior distribution for the parameters is proportional to:

$$\begin{aligned}
&\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{e^{\gamma_{bj}x_{bj}}}{(1+e^{\gamma_{bj}})^{N_C}}\right] \\
&\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}}\right] \\
&\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{(\gamma_{bj}-\mu_{\gamma b})^2}{2\sigma_{\gamma b}^2}\right)\right] \\
&\times\prod_{b=1}^{B}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right] \\
&\times\prod_{b=1}^{B}\left[\frac{{\beta_\gamma}^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}(\sigma_{\gamma b}^2)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\sigma_{\gamma b}^2}\right)\right] \\
&\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right] \\
&\times\prod_{b=1}^{B}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \\
&\times\prod_{b=1}^{B}\left[\frac{{\beta_\theta}^{\alpha_\theta}}{\Gamma(\alpha_\theta)}(\sigma_{\theta b}^2)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\sigma_{\theta b}^2}\right)\right] \\
&\times\frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau_{\gamma 00}^2}\right) \\
&\times\frac{1}{\sqrt{2\pi\tau_{\theta 00}^2}}\exp\left(-\frac{(\mu_{\theta 0}-\mu_{\theta 00})^2}{2\tau_{\theta 00}^2}\right) \\
&\times\frac{{\beta_{\gamma 00}}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}(\tau_{\gamma 0}^2)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\tau_{\gamma 0}^2}\right) \\
&\times\frac{{\beta_{\theta 00}}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}(\tau_{\theta 0}^2)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\tau_{\theta 00}^2}\right)
\end{aligned} \tag{B.3}$$

The complete conditionals may be read directly from the joint distribution.

B.3.1 Complete Conditional Distributions

$$[\gamma_{bj}\Big|...] \propto \frac{e^{\gamma_{bj}x_{bj}}}{(1+e^{\gamma_{bj}})^{N_C}} \frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}} \exp\left(-\frac{(\gamma_{bj}-\mu_{\gamma b})^2}{2\sigma^2_{\gamma b}}\right) \tag{B.4}$$

$$[\theta_{bj}\Big|...] \propto \frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right) \tag{B.5}$$

$$[\mu_{\gamma b}\Big|...] \propto \exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau^2_{\gamma 0}}\right) \prod_{j=1}^{k_b} \exp\left(-\frac{(\gamma_{bj}-\mu_{\gamma b})^2}{2\sigma^2_{\gamma b}}\right)$$

From Result 1 with $\mu = \mu_{\gamma b}$, $\sigma^2 = \sigma^2_{\gamma b}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = k_b$, and $x_b = \gamma_{bj}$ we have:

$$[\mu_{\gamma b}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{j=1}^{k_b}\gamma_{bj}}{\sigma^2_{\gamma b}+k_b\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b}+k_b\tau^2_{\gamma 0}}\right) \tag{B.6}$$

$$[\mu_{\theta b}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau^2_{\theta 0}}\right) \prod_{j=1}^{k_b} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)$$

$$[\mu_{\theta b}\Big|...] \sim N\left(\frac{\mu_{\theta 0}\sigma^2_{\theta b} + \tau^2_{\theta 0}\sum_{j=1}^{k_b}(\theta_{bj})}{\sigma^2_{\theta b}+k_b\tau^2_{\theta 0}}, \frac{\tau^2_{\theta 0}\sigma^2_{\theta b}}{\sigma^2_{\theta b}+k_b\tau^2_{\theta 0}}\right) \tag{B.7}$$

$$[\mu_{\gamma 0}\Big|...] \propto \exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau^2_{\gamma 00}}\right) \prod_{b=1}^{B} \exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau^2_{\gamma 0}}\right)$$

From Result 1 with $\mu = \mu_{\gamma 0}$, $\sigma^2 = \tau^2_{\gamma 0}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = B$, and $x_b = \mu_{\gamma b}$ we have:

$$[\mu_{\gamma 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau_{\gamma 0}^2 + \tau_{\gamma 00}^2\sum_{b=1}^{B}\mu_{\gamma b}}{\tau_{\gamma 0}^2 + B\tau_{\gamma 00}^2}, \frac{\tau_{\gamma 0}^2\tau_{\gamma 00}^2}{\tau_{\gamma 0}^2 + B\tau_{\gamma 00}^2}\right) \tag{B.8}$$

$$[\mu_{\theta 0}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta 0} - \mu_{\theta 00})^2}{2\tau_{\theta 00}^2}\right)\prod_{b=1}^{B}\exp\left(-\frac{(\mu_{\theta b} - \mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)$$

$$[\mu_{\theta 0}\Big|...] \sim N\left(\frac{\tau_{\theta 0}^2\mu_{\theta 00} + \tau_{\theta 00}^2\sum_{b=1}^{B}\mu_{\theta b}}{\tau_{\theta 0}^2 + B\tau_{\theta 00}^2}, \frac{\tau_{\theta 00}^2\tau_{\theta 0}^2}{\tau_{\theta 0}^2 + B\tau_{\theta 00}^2}\right) \tag{B.9}$$

$$[\sigma_{\gamma b}^2\Big|...] \propto (\sigma_{\gamma b}^2)^{-\alpha_\gamma - 1}\exp\left(-\frac{\beta_\gamma}{\sigma_{\gamma b}^2}\right)\prod_{j=1}^{k_b}\frac{1}{\sqrt{\sigma_{\gamma b}^2}}\exp\left(-\frac{(\gamma_{bj} - \mu_{\gamma b})^2}{2\sigma_{\gamma b}^2}\right)$$

Applying Result 5 with $\sigma^2 = \sigma_{\gamma b}^2$, $\mu = \mu_{\gamma b}$, $x_j = \gamma_{bj}$, $\alpha = \alpha_\gamma$, $\beta = \beta_\gamma$, $N = k_b$ gives:

$$[\sigma_{\gamma b}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma + \frac{k_b}{2}, \beta_\gamma + \frac{1}{2}\sum_{j=1}^{k_b}(\gamma_{bj} - \mu_{\gamma b})^2\right) \tag{B.10}$$

$$[\sigma_{\theta b}^2\Big|...] \propto (\sigma_{\theta b}^2)^{-\alpha_\theta - 1}\exp\left(-\frac{\beta_\theta}{\sigma_{\theta b}^2}\right)\prod_{j=1}^{k_b}\left[\frac{1}{\sqrt{\sigma_{\theta b}^2}}\exp\left(-\frac{(\theta_{bj} - \mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right]$$

$$[\sigma_{\theta b}^2\Big|...] \sim IG\left(\alpha_\theta + \frac{k_b}{2}, \beta_\theta + \frac{1}{2}\sum_{j=1}^{k_b}(\theta_{bj} - \mu_{\theta b})^2\right) \tag{B.11}$$

$$[\tau_{\gamma 0}^2\Big|...] \propto (\tau_{\gamma 0}^2)^{-\alpha_{\gamma 00} - 1}\exp\left(-\frac{\beta_{\gamma 00}}{\tau_{\gamma 0}^2}\right)\prod_{b=1}^{B}\frac{1}{\sqrt{\tau_{\gamma 0}^2}}\exp\left(-\frac{(\mu_{\gamma b} - \mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)$$

Applying Result 5 with $\sigma^2 = \tau_{\gamma 0}^2$, $\mu = \mu_{\gamma 0}$, $x_b = \mu_{\gamma b}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, $N = B$ gives:

$$[\tau_{\gamma 0}^2 \Big| ...] \sim \text{IG}\left(\alpha_{\gamma 00} + \frac{B}{2}, \beta_{\gamma 00} + \frac{1}{2}\sum_{i=1}^{B}(\mu_{\gamma b} - \mu_{\gamma 0})^2\right) \tag{B.12}$$

$$[\tau_{\theta 0}^2 \Big| ...] \propto (\tau_{\theta 0}^2)^{-\alpha_{\theta 00}-1} \exp\left(-\frac{\beta_{\theta 00}}{\tau_{\theta 00}^2}\right) \prod_{b=1}^{B}\left[\frac{1}{\sqrt{\tau_{\theta 0}^2}} \exp\left(-\frac{(\mu_{\theta b} - \mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \tag{B.13}$$

$$[\tau_{\theta 0}^2 \Big| ...] \sim IG\left(\alpha_{\theta 00} + \frac{B}{2}, \beta_{\theta 00} + \frac{1}{2}\sum_{b=1}^{B}(\mu_{\theta b} - \mu_{\theta 0})^2\right)$$

B.4 Model c212.BB

The joint posterior distribution is proportional to:

$$
\begin{aligned}
&\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{e^{\gamma_{bj}x_{bj}}}{\left(1+e^{\gamma_{bj}}\right)^{N_C}}\right]\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}}\right]\\
&\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{(\gamma_{bj}-\mu_{\gamma b})^2}{2\sigma_{\gamma b}^2}\right)\right]\\
&\times\prod_{b=1}^{B}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\times\prod_{b=1}^{B}\left[\frac{{\beta_\gamma}^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}(\sigma_{\gamma b}^2)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\sigma_{\gamma b}^2}\right)\right]\\
&\times\prod_{b=1}^{B}\prod_{j=1}^{k_b}\left[\pi_b\,\mathrm{I}_{[\theta_{bj}=0]}+(1-\pi_b)\,\mathrm{I}_{[\theta_{bj}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right]\\
&\times\prod_{b=1}^{B}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\times\prod_{b=1}^{B}\left[\frac{{\beta_\theta}^{\alpha_\theta}}{\Gamma(\alpha_\theta)}(\sigma_{\theta b}^2)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\sigma_{\theta b}^2}\right)\right]\\
&\times\frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau_{\gamma 00}^2}\right)\\
&\times\frac{1}{\sqrt{2\pi\tau_{\theta 00}^2}}\exp\left(-\frac{(\mu_{\theta 0}-\mu_{\theta 00})^2}{2\tau_{\theta 00}^2}\right)\\
&\times\frac{{\beta_{\gamma 00}}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}(\tau_{\gamma 0}^2)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\tau_{\gamma 0}^2}\right)\\
&\times\frac{{\beta_{\theta 00}}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}(\tau_{\theta 0}^2)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\tau_{\theta 00}^2}\right)\\
&\times\prod_{b=1}^{B}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}\pi_b^{\alpha_\pi-1}\left(1-\pi_b\right)^{\beta_\pi-1}\right]\\
&\times\lambda_\alpha\frac{\exp(-\lambda_\alpha\alpha_\pi)}{\exp(-\lambda_\alpha)}\,\mathrm{I}_{[\alpha_\pi>1]}\\
&\times\lambda_\beta\frac{\exp(-\lambda_\beta\beta_\pi)}{\exp(-\lambda_\beta)}\,\mathrm{I}_{[\beta_\pi>1]}
\end{aligned}
\tag{B.14}
$$

B.4.1 Complete Conditional Distributions

As for model c212.1a the complete conditional can be read from the posterior. The only differences from model c212.1a are:

$$
\begin{aligned}
[\theta_{bj}\Big|...] \propto & \frac{e^{\left(\gamma_{bj}+\theta_{bj}\right)y_{bj}}}{\left(1+e^{\left(\gamma_{bj}+\theta_{bj}\right)}\right)^{N_T}} \\
& \times \left[\pi_b \, \mathrm{I}_{[\theta_{bj}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)\right]
\end{aligned}
\tag{B.15}
$$

$$
\begin{aligned}
[\pi_b\Big|...] &\propto {\pi_b}^{\alpha_\pi - 1} \left(1-\pi_b\right)^{\beta_\pi - 1} \\
&\quad \times \prod_{j=1}^{k_b} \left[\pi_b \, \mathrm{I}_{[\theta_{bj}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)\right] \\
&= {\pi_b}^{\alpha_\pi - 1} \left(1-\pi_b\right)^{\beta_\pi - 1} \\
&\quad \times \left[\pi_b^{\sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}} \left((1-\pi_b)\frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)\right)^{\sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}\neq 0]}}\right] \\
&\propto {\pi_b}^{\alpha_\pi - 1} \left(1-\pi_b\right)^{\beta_\pi - 1} \left[\pi_b^{\sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}} (1-\pi_b)^{\sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}\neq 0]}}\right] \\
&= {\pi_b}^{\alpha_\pi - 1 + \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}} \left(1-\pi_b\right)^{\beta_\pi - 1 + k_b - \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}}
\end{aligned}
$$

$$
[\pi_b\Big|...] \sim \mathrm{Beta}\left(\alpha_\pi + \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}, \beta_\pi + k_b - \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}=0]}\right)
\tag{B.16}
$$

$$
\begin{aligned}
[\alpha_\pi \Big| ...] &\propto \lambda_\alpha \frac{\exp(-\lambda_\alpha \alpha_\pi)}{\exp(-\lambda_\alpha)} \, \mathrm{I}_{[\alpha_\pi > 1]} \prod_{b=1}^{B} \left[\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} {\pi_b}^{\alpha_\pi - 1} \left(1 - \pi_b\right)^{\beta_\pi - 1} \right] \\
&\propto \exp(-\lambda_\alpha \alpha_\pi) \, \mathrm{I}_{[\alpha_\pi > 1]} \left(\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \right)^B \prod_{b=1}^{B} {\pi_b}^{\alpha_\pi - 1} \\
&\propto \exp(-\lambda_\alpha \alpha_\pi) \left(\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)} \right)^B \left[\prod_{b=1}^{B} \pi_b \right]^{\alpha_\pi - 1} \mathrm{I}_{[\alpha_\pi > 1]}
\end{aligned}
\tag{B.17}
$$

$$
\begin{aligned}
[\beta_\pi \Big| ...] &\propto \lambda_\beta \frac{\exp(-\lambda_\beta \beta_\pi)}{\exp(-\lambda_\beta)} \, \mathrm{I}_{[\beta_\pi > 1]} \prod_{b=1}^{B} \left[\left(\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \right) {\pi_b}^{\alpha_\pi - 1} \left(1 - \pi_b\right)^{\beta_\pi - 1} \right] \\
&\propto \exp(-\lambda_\beta \beta_\pi) \, \mathrm{I}_{[\beta_\pi > 1]} \left(\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \right)^B \prod_{b=1}^{B} \left(1 - \pi_b\right)^{\beta_\pi - 1} \\
&\propto \exp(-\lambda_\beta \beta_\pi) \left(\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\beta_\pi)} \right)^B \left[\prod_{b=1}^{B} \left(1 - \pi_b\right) \right]^{\beta_\pi - 1} \mathrm{I}_{[\beta_\pi > 1]}
\end{aligned}
\tag{B.18}
$$

$$
\begin{aligned}
[\mu_{\theta b} \Big| ...] &\propto \exp\left(-\frac{\left(\mu_{\theta b} - \mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2} \right) \\
&\quad \times \prod_{j=1}^{k_b} \left[\pi_b \, \mathrm{I}_{[\theta_{bj} = 0]} + (1 - \pi_b) \, \mathrm{I}_{[\theta_{bj} \neq 0]} \, \frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}} \exp\left(-\frac{\left(\theta_{bj} - \mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2} \right) \right]
\end{aligned}
$$

Applying Result 3 with $K_b = \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj} \neq 0]}$ and $\mu = \mu_{\theta b}$, $\sigma^2 = \sigma_{\theta b}^2$, $\mu_0 = \mu_{\theta 0}$, $\sigma_0^2 = \tau_{\theta 0}^2$, $N = k_b$, $p = \pi_b$, and $x_j = \theta_{bj}$ we have:

$$
[\mu_{\theta b} \Big| ...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b}^2 + \tau_{\theta 0}^2 \sum_{j=1}^{k_b} \theta_{bj}}{\sigma_{\theta b}^2 + K_b \tau_{\theta 0}^2}, \frac{\sigma_{\theta b}^2 \tau_{\theta 0}^2}{\sigma_{\theta b}^2 + K_b \tau_{\theta 0}^2} \right) \tag{B.19}
$$

$$[\sigma^2_{\theta b}\Big|...] \propto (\sigma^2_{\theta b})^{-\alpha_\theta - 1} \exp\left(-\frac{\beta_\theta}{\sigma^2_{\theta b}}\right)$$
$$\times \prod_{j=1}^{k_b} \left[\pi_b \, \mathrm{I}_{[\theta_{bj}=0]} + (1-\pi_b) \, \mathrm{I}_{[\theta_{bj}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)\right]$$

Applying Result 7 with $K_b = \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}\neq 0]}$, $\sigma^2 = \sigma^2_{\theta b}$, $\mu = \mu_{\theta b}$, $x_j = \theta_{bj}$, $\alpha = \alpha_\theta$, $\beta = \beta_\theta$, and $N = k_b$ we have:

$$[\sigma^2_{\theta b}\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{K_b}{2}, \beta_\theta + \frac{1}{2}\sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj}\neq 0]}(\theta_{bj}-\mu)^2\right) \tag{B.20}$$

B.5 Model $1a_{20}$

From §6.7.4 the joint posterior distribution of the parameters is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b,h}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 0}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 0}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0}\right)^2}{2\tau^2_{\theta 0}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma^2_{\theta b,h}\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]
\end{aligned}
\tag{B.21}
$$

The complete conditionals may be read directly from the joint distribution.

B.5.1 Complete Conditional Distributions

$$
\begin{aligned}
\left[\gamma_{bj,h}\middle|...\right] &\propto \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right]\left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
&\quad\times\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right] \\
\left[\gamma_{bj,h}\middle|...\right] &\propto \left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}}C_{bj,h}}\left(e^{\gamma_{bj,h}}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}} \\
&\quad\times\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)
\end{aligned}
\tag{B.22}
$$

$$
\begin{aligned}
\left[\theta_{bj,h}\middle|...\right] &\propto \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
&\quad\times\left[\left(\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right)\right] \\
\left[\theta_{bj,h}\middle|...\right] &\propto \left(e^{\left(\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\left(\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right)
\end{aligned}
\tag{B.23}
$$

$$
\left[\mu_{\gamma b,h}\middle|...\right] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right]\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]
$$

Applying Result 1 with $\mu = \mu_{\gamma b,h}$, $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma_0^2 = \tau^2_{\gamma 0}$, $N = k_{bh}$, and $x_j = \gamma_{bj,h}$ we have:

$$
\left[\mu_{\gamma b,h}\middle|...\right] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b,h} + \tau^2_{\gamma 0}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b,h}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}\right) \tag{B.24}
$$

$$[\mu_{\theta b,h}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right]\prod_{j=1}^{k_{bh}}\left[\left(\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right)\right]$$

$$[\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b,h}^2+\tau_{\theta 0}^2\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma_{\theta b,h}^2+k_{bh}\tau_{\theta 0}^2}, \frac{\tau_{\theta 0}^2\sigma_{\theta b,h}^2}{\sigma_{\theta b,h}^2+k_{bh}\tau_{\theta 0}^2}\right) \tag{B.25}$$

$$[\sigma_{\gamma b,h}^2\Big|...] \propto \left[\left(\sigma_{\gamma b,h}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b,h}^2\right)}\right)\right]\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\gamma b,h}^2}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b,h})^2}{2\sigma_{\gamma b,h}^2}\right)\right]$$

Applying Result 5 with $\sigma^2=\sigma_{\gamma b,h}^2$, $\mu=\mu_{\gamma b,h}$, $x_j=\gamma_{bj,h}$, $\alpha=\alpha_{\gamma,h}$ and $\beta=\beta_{\gamma,h}$, $N=k_{bh}$ gives:

$$[\sigma_{\gamma b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma+\frac{k_{bh}}{2}, \beta_\gamma+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b,h})^2\right) \tag{B.26}$$

$$[\sigma_{\theta b,h}^2\Big|...] \propto \left[\left(\sigma_{\theta b,h}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b,h}^2\right)}\right)\right]\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\theta b,h}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right]$$

$$[\sigma_{\theta b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\theta+\frac{k_{bh}}{2}, \beta_\theta+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\theta_{bj,h}-\mu_{\theta b,h})^2\right) \tag{B.27}$$

B.6 Model $1a_{21}$

From §6.7.4 the joint posterior distribution of the parameters is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b}^2\right)}\right)\right]
\end{aligned}
\tag{B.28}
$$

B.6.1 Complete Conditional Distributions

$$[\gamma_{bj,h}|...] \propto \left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\, e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right] \times \left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right]$$

$$[\gamma_{bj,h}|...] \propto \left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}}\, e^{-e^{\gamma_{bj,h}}C_{bj,h}}\, \left(e^{\gamma_{bj,h}}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right) \tag{B.29}$$

$$[\theta_{bj,h}|...] \propto \left(e^{\theta_{bj,h}}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\left(\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma^2_{\theta b}}\right)\right) \tag{B.30}$$

$$[\mu_{\gamma b}|...] \propto \exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right]$$

Applying Result 2 with $\mu = \mu_{\gamma b}$, $\sigma^2 = \sigma^2_{\gamma b}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = H$, $\sum_{i=1}^{N} k_i = \sum_{h=1}^{H} k_{bh}$, and $x_{hj} = \gamma_{bj,h}$ we have:

$$[\mu_{\gamma b}|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{h=1}^{H}k_{bh}}, \frac{\sigma^2_{\gamma b}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{h=1}^{H}k_{bh}}\right) \tag{B.31}$$

$$
\begin{aligned}
\left[\mu_{\theta b}\middle|...\right] &\propto \exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right] \\
\left[\mu_{\theta b}\middle|...\right] &\sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b}^2+\tau_{\theta 0}^2\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma_{\theta b}^2+\tau_{\theta 0}^2\sum_{h=1}^{H}k_{bh}},\frac{\tau_{\gamma 0}^2\sigma_{\theta b}^2}{\sigma_{\theta b}^2+\tau_{\theta 0}^2\sum_{h=1}^{H}k_{bh}}\right)
\end{aligned}
\tag{B.32}
$$

$$
\left[\sigma_{\gamma b}^2\middle|...\right] \propto \left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\gamma b}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]
$$

Applying Result 6 with $\sigma^2=\sigma_{\gamma b}^2$, $\mu=\mu_{\gamma b}$, $x_{hj}=\gamma_{bj,h}$, $\alpha=\alpha_\gamma$, $\beta=\beta_\gamma$, and $N=H$ we have:

$$
\left[\sigma_{\gamma b}^2\middle|...\right] \sim \mathrm{IG}\left(\alpha_\gamma+\frac{\sum_{h=1}^{H}k_{bh}}{2},\beta_\gamma+\frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b})^2\right) \tag{B.33}
$$

$$
\left[\sigma_{\theta b}^2\middle|...\right] \propto \left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b}^2\right)}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\theta b}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right]
$$

$$
\left[\sigma_{\theta b}^2\middle|...\right] \sim \mathrm{IG}\left(\alpha_\theta+\frac{1}{2}\sum_{h=1}^{H}k_{bh},\beta_\theta+\frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\left(\theta_{bj,h}-\mu_{\theta b}\right)^2\right) \tag{B.34}
$$

B.7 Model $\mathbf{BB_{20}}$

The joint posterior distribution is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b,h}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 0}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 0}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0}\right)^2}{2\tau^2_{\theta 0}}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma^2_{\theta b,h}\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right]
\end{aligned}
\tag{B.35}
$$

B.7.1 Complete Conditional Distributions

$$\begin{aligned}
\left[\theta_{bjh}\middle|...\right] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
& \times \left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]
\end{aligned} \tag{B.36}$$

$$\begin{aligned}
\left[\pi_{bh}\middle|...\right] \propto & \left[\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right] \\
& \times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right] \\
= & \left[\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right] \\
& \times \left[\pi_{b,h}^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}\left((1-\pi_{b,h})\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right)^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}}\right] \\
\propto & \left[\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right]\left[\pi_{b,h}^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}(1-\pi_{b,h})^{k_{bh}-\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}\right] \\
\propto & \left[\pi_{b,h}^{\alpha_\pi+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}-1}(1-\pi_{b,h})^{\beta_\pi+k_{bh}-\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}-1}\right]
\end{aligned}$$

$$\left[\pi_{bh}\middle|...\right] \sim \mathrm{Beta}\left(\alpha_\pi + \sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}, \beta_\pi + k_{bh} - \sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}\right) \tag{B.37}$$

$$[\mu_{\theta b,h}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b,h}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right]$$

Applying Result 3 with $\mu = \mu_{\theta_{b,h}}$, $\sigma^2 = \sigma_{\theta b,h}^2$, $\mu_0 = \mu_{\theta 0}$, $\sigma_0^2 = \tau_{\theta 0}^2$, $x_j = \theta_{bj,h}$, and $N = K_b$, where $K_b = \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj,h}\neq 0]}$ we have:

$$[\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b,h}^2 + \tau_{\theta 0}^2\sum_{i=j}^{k_b}\theta_{bj,h}}{\sigma_{\theta b,h}^2 + K_b\tau_{\theta 0}^2}, \frac{\sigma_{\theta b,h}^2\tau_{\theta 0}^2}{\sigma_{\theta b,h}^2 + K_b\tau_{\theta 0}^2}\right) \tag{B.38}$$

$$[\sigma_{\theta b,h}^2\Big|...] \propto \left[\left(\sigma_{\theta b,h}^2\right)^{-\alpha_\theta - 1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b,h}^2\right)}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b,h}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right]$$

Applying Result 7 with $\sigma^2 = \sigma_{\theta b,h}^2$, $\mu = \mu_{\theta b,h}$, $x_j = \theta_{bj,h}$, $\alpha = \alpha_\theta$, $\beta = \beta_\theta$, and $N = K_b$, where $K_b = \sum_{j=1}^{k_b} \mathrm{I}_{[\theta_{bj,h}\neq 0]}$ we have:

$$[\sigma_{\theta b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{K_b}{2}, \beta_\theta + \frac{1}{2}\sum_{j=1}^{k_b}\mathrm{I}_{[\theta_{bj,h}\neq 0]}(\theta_{bj,h}-\mu_{\theta b,h})^2\right) \tag{B.39}$$

B.8 Model $\mathbf{BB_{21}}$

The joint posterior distribution is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\pi_b\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_b)\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b}^2\right)}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}\pi_b^{\alpha_\pi-1}(1-\pi_b)^{\beta_\pi-1}\right]
\end{aligned}
\tag{B.40}
$$

B.8.1 Complete Conditional Distributions

The complete conditionals may be read from the joint distribution. Many are the same as for model $1a_{21}$. The different distributions are:

$$[\gamma_{bj,h}\Big|...] \propto \left[(e^{\gamma_{bj,h}})^{x_{bj,h}}\, e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right]\left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \times \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right] \tag{B.41}$$

$$[\theta_{bj,h}\Big|...] \propto \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \times \left[\pi_b\, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}=0]}\, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma^2_{\theta b}}\right)\right] \tag{B.42}$$

$$[\mu_{\gamma b}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right] \tag{B.43}$$

$$[\mu_{\gamma b}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b}+\tau^2_{\gamma 0}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b}+\tau^2_{\gamma 0}\sum_{h=1}^{H}k_{bh}}, \frac{\tau^2_{\gamma 0}\sigma^2_{\gamma b}}{\sigma^2_{\gamma b}+\tau^2_{\gamma 0}\sum_{h=1}^{H}k_{bh}}\right)$$

$$[\mu_{\theta b}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau^2_{\theta 0}}\right) \times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\pi_b\, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}\neq 0]}\, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma^2_{\theta b}}\right)\right] \tag{B.44}$$

Applying Result 4 with $p = \pi_b$, $x_{hj} = \theta_{bj,h}$, $N = H$, $\mu = \mu_{\theta b}$, $\sigma^2 = \sigma^2_{\theta b}$, $\mu_0 = \mu_{\theta 0}$,

$\sigma_0^2 = \tau_{\theta 0}^2$ we have:

$$[\mu_{\theta b}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b}^2 + \tau_{\theta 0}^2 \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \theta_{bj,h}}{\sigma_{\theta b}^2 + \tau_{\theta 0}^2 \sum_{h=1}^{H} K_{bh}}, \frac{\tau_{\theta 0}^2 \sigma_{\theta b}^2}{\sigma_{\theta b}^2 + \tau_{\theta 0}^2 \sum_{h=1}^{H} K_{bh}}\right)$$

where $K_{bh} = \sum_{k=1}^{k_{bj}} \mathrm{I}_{[\theta_{bj,h} \neq 0]}$.

$$[\sigma_{\gamma b}^2\Big|...] \propto \left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma - 1} \exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right) \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}} \left[\frac{1}{\sqrt{\sigma_{\gamma b}^2}} \exp\left(-\frac{\left(\gamma_{bj,h} - \mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]$$

$$[\sigma_{\gamma b}^2\Big|...] \sim IG\left(\alpha_\gamma + \frac{1}{2}\sum_{h=1}^{H} k_{bh}, \beta_\gamma + \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \left(\gamma_{bj,h} - \mu_{\gamma b}\right)^2\right) \tag{B.45}$$

$$[\sigma_{\theta b}^2\Big|...] \propto \left(\sigma_{\theta b}^2\right)^{-\alpha_\theta - 1} \exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b}^2\right)}\right) \times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}} \left[\left(\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}} \exp\left(-\frac{\left(\theta_{bj,h} - \mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right)\right] \tag{B.46}$$

Applying Result 8 we have:

$$[\sigma_{\theta b}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{1}{2}\sum_{h=1}^{H} K_{bh}, \beta_\theta + \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h} \neq 0]} \left(\theta_{bj,h} - \mu_{\theta b}\right)^2\right) \tag{B.47}$$

where $K_{bh} = \sum_{j=1}^{k_{bj}} \mathrm{I}_{[\theta_{bj,h} \neq 0]}$.

$$
\begin{aligned}
[\pi_b \Big| ...] &\propto \pi_b^{\alpha_\pi - 1}(1-\pi_b)^{\beta_\pi - 1} \\
&\times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}} \left[\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}} \right) \right] \\
&\propto \pi_b^{\alpha_\pi - 1}(1-\pi_b)^{\beta_\pi - 1} \left[\pi_b^{\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]}} (1-\pi_b)^{\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]}} \right] \\
&\propto \pi_b^{\alpha_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]} - 1} (1-\pi_b)^{\beta_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]} - 1}
\end{aligned}
$$

$$
[\pi_b \Big| ...] \sim \mathrm{Beta}\left(\alpha_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]}, \beta_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]} \right) \tag{B.48}
$$

B.9 Model $1a_{30}$

From §6.6.4 the joint posterior distribution for the parameters is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b,h}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 0,h}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\gamma,h}^{\alpha_{\gamma,h}}}{\Gamma(\alpha_{\gamma,h})}\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 00}}}\exp\left(-\frac{\left(\mu_{\gamma 0,h}-\mu_{\gamma 00}\right)^2}{2\tau^2_{\gamma 00}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau^2_{\gamma 0,h}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0,h}\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 0,h}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0,h}\right)^2}{2\tau^2_{\theta 0,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\theta,h}^{\alpha_{\theta,h}}}{\Gamma(\alpha_{\theta,h})}\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 00}}}\exp\left(-\frac{\left(\mu_{\theta 0,h}-\mu_{\theta 00}\right)^2}{2\tau^2_{\theta 00}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau^2_{\theta 0,h}\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau^2_{\theta 0,h}\right)}\right)\right]
\end{aligned}
\tag{B.49}
$$

B.9.1 Complete Conditional Distributions

$$
\begin{aligned}
[\gamma_{bj,h}\Big|...] \propto & \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right] \\
& \times \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
& \times \exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)
\end{aligned}
\tag{B.50}
$$

$$
\begin{aligned}
[\theta_{bj,h}\Big|...] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
& \times \left[\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]
\end{aligned}
\tag{B.51}
$$

$$
[\mu_{\gamma b,h}\Big|...] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right] \prod_{j=1}^{k_{bh}} \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right] \tag{B.52}
$$

Using Result 1 with $x_j = \gamma_{bj,h}$, $\mu = \mu_{\gamma b,h}$, $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu_0 = \mu_{\gamma 0,h}$, $\sigma_0^2 = \tau^2_{\gamma 0,h}$, and $N = k_{bh}$ we have:

$$
\mu_{\gamma b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0,h}\sigma^2_{\gamma b,h} + \tau^2_{\gamma 0,h}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0,h}}, \frac{\sigma^2_{\gamma b,h}\tau^2_{\gamma 0,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0,h}}\right)
$$

$$[\mu_{\theta b,h}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0,h})^2}{2\tau^2_{\theta 0,h}}\right) \times \prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma^2_{\theta b,h}}\right)\right]$$

$$\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0,h}\sigma^2_{\theta b,h}+\tau^2_{\theta 0,h}\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma^2_{\theta b,h}+k_{bh}\tau^2_{\theta 0,h}}, \frac{\sigma^2_{\theta b,h}\tau^2_{\theta 0,h}}{\sigma^2_{\theta b,h}+k_{bh}\tau^2_{\theta 0,h}}\right) \tag{B.53}$$

$$\begin{aligned}[\sigma^2_{\gamma b,h}\Big|...] \propto &\left[\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right] \\ &\times \prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma^2_{\gamma b,h}}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b,h})^2}{2\sigma^2_{\gamma b,h}}\right)\right]\end{aligned} \tag{B.54}$$

Applying Result 5 with $\sigma^2=\sigma^2_{\gamma b,h}$, $\mu=\mu_{\gamma b,h}$, $x_j=\gamma_{bj,h}$, $\alpha=\alpha_{\gamma,h}$ and $\beta=\beta_{\gamma,h}$, $N=k_{bh}$ gives:

$$[\sigma^2_{\gamma b,h}\Big|...] \sim \mathrm{IG}\left(\alpha_{\gamma,h}+\frac{k_{bh}}{2}, \beta_{\gamma,h}+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b,h})^2\right) \tag{B.55}$$

$$\begin{aligned}[\sigma^2_{\theta b,h}\Big|...] \propto &\left[\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right] \\ &\times \prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma^2_{\theta b,h}}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma^2_{\theta b,h}}\right)\right]\end{aligned} \tag{B.56}$$

$$[\sigma^2_{\theta b,h}\Big|...] \sim \mathrm{IG}\left(\alpha_{\theta,h}+\frac{k_{bh}}{2}, \beta_{\theta,h}+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\theta_{bj,h}-\mu_{\theta b,h})^2\right)$$

$$[\mu_{\gamma 0,h}\Big|...] \propto \exp\left(-\frac{(\mu_{\gamma 0,h}-\mu_{\gamma 00})^2}{2\tau^2_{\gamma 00}}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\gamma b,h}-\mu_{\gamma 0,h})^2}{2\tau^2_{\gamma 0,h}}\right)\right] \tag{B.57}$$

Applying Result 1 with $\mu = \mu_{\gamma 0,h}$, $\sigma^2 = \tau^2_{\gamma 0,h}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = B_h$, and $x_b = \mu_{\gamma b,h}$ we have:

$$[\mu_{\gamma 0,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0,h}+\tau^2_{\gamma 00}\sum_{b=1}^{B_h}\mu_{\gamma b,h}}{\tau^2_{\gamma 0,h}+B_h\tau^2_{\gamma 00}}, \frac{\tau^2_{\gamma 0,h}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0,h}+B_h\tau^2_{\gamma 00}}\right) \tag{B.58}$$

$$[\mu_{\theta 0,h}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta 0,h}-\mu_{\theta 00})^2}{2\tau^2_{\theta 00}}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0,h})^2}{2\tau^2_{\theta 0,h}}\right)\right]$$

$$[\mu_{\theta 0,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau^2_{\theta 0,h}+\tau^2_{\theta 00}\sum_{b=1}^{B_h}\mu_{\theta b,h}}{\tau^2_{\theta 0,h}+B_h\tau^2_{\theta 00}}, \frac{\tau^2_{\theta 0,h}\tau^2_{\theta 00}}{\tau^2_{\theta 0,h}+B_h\tau^2_{\theta 00}}\right) \tag{B.59}$$

$$[\tau^2_{\gamma 0,h}\Big|...] \propto \left(\tau^2_{\gamma 0,h}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0,h}\right)}\right)\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\gamma 0,h}}}\exp\left(-\frac{(\mu_{\gamma b,h}-\mu_{\gamma 0,h})^2}{2\tau^2_{\gamma 0,h}}\right)\right] \tag{B.60}$$

Applying Result 5 with $\sigma^2 = \tau^2_{\gamma 0,h}$, $\mu = \mu_{\gamma 0,h}$, $x_b = \mu_{\gamma b,h}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, and $N = B_h$ gives:

$$[\tau^2_{\gamma 0,h}\Big|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00}+\frac{B_h}{2}, \beta_{\gamma 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\gamma b,h}-\mu_{\gamma 0,h})^2\right) \tag{B.61}$$

$$[\tau^2_{\theta 0,h}\Big|...] \propto \left(\tau^2_{\theta 0,h}\right)^{-\alpha_{\theta 00}-1} \exp\left(-\frac{\beta_{\theta 00}}{\left(\tau^2_{\theta 0,h}\right)}\right) \prod_{b=1}^{B_h} \left[\frac{1}{\sqrt{\tau^2_{\theta 0,h}}} \left(\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0,h}\right)^2}{2\tau^2_{\theta 0,h}}\right)\right)\right]$$

$$[\tau^2_{\theta 0,h}\Big|...] \sim \text{IG}\left(\alpha_{\theta 00}+\frac{B_h}{2}, \beta_{\theta 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\theta b,h}-\mu_{\theta 0,h})^2\right) \tag{B.62}$$

B.10 Model $1a_{31}$

From §6.6.4 the joint posterior distribution for the parameters is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_{\gamma}^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\right]\\
&\quad\times\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{\left(\mu_{\gamma 0}-\mu_{\gamma 00}\right)^2}{2\tau_{\gamma 00}^2}\right)\right]\\
&\quad\times\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau_{\gamma 0}^2\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau_{\gamma 0}^2\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\quad\times\prod_{b=1}^{B_h}\left[\frac{\beta_{\theta}^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{(\sigma_{\theta b}^2)}\right)\right]\\
&\quad\times\left[\frac{1}{\sqrt{2\pi\tau_{\theta 00}^2}}\exp\left(-\frac{\left(\mu_{\theta 0}-\mu_{\theta 00}\right)^2}{2\tau_{\theta 00}^2}\right)\right]\\
&\quad\times\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{(\tau_{\theta 0}^2)}\right)\right]
\end{aligned}
\tag{B.63}
$$

B.10.1 Complete Conditional Distributions

$$[\gamma_{bj,h}\Big|...] \propto \left[(e^{\gamma_{bj,h}})^{x_{bj,h}}\, e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right]\left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \times \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right] \tag{B.64}$$

$$[\theta_{bj,h}\Big|...] \propto \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}}\, e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \times \left[\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma^2_{\theta b}}\right)\right] \tag{B.65}$$

$$[\mu_{\gamma b}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right] \tag{B.66}$$

Applying Result 2 with $\mu = \mu_{\gamma b}$, $\sigma^2 = \sigma^2_{\gamma b}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = H$, $\sum_{i=1}^{N} k_i = \sum_{h=1}^{H} k_{bh}$, and $x_{hj} = \gamma_{bj,h}$ we have:

$$[\mu_{\gamma b}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b} + \sum_{h=1}^{H}k_{bh}\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b} + \sum_{h=1}^{H}k_{bh}\tau^2_{\gamma 0}}\right) \tag{B.67}$$

$$[\mu_{\theta b}|...] \propto \exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right]$$

$$[\mu_{\theta b}|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b}^2+\tau_{\theta 0}^2\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma_{\theta b}^2+\sum_{h=1}^{H}k_{bh}\tau_{\theta 0}^2},\frac{\sigma_{\theta b}^2\tau_{\theta 0}^2}{\sigma_{\theta b}^2+\sum_{h=1}^{H}k_{bh}\tau_{\theta 0}^2}\right) \tag{B.68}$$

$$[\sigma_{\gamma b}^2|...] \propto \left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\gamma b}^2}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b})^2}{2\sigma_{\gamma b}^2}\right)\right] \tag{B.69}$$

Applying Result 5 with $\sigma^2=\sigma_{\gamma b}^2$, $\mu=\mu_{\gamma b}$, $x_{hj}=\gamma_{bj,h}$, $\alpha=\alpha_\gamma$, $\beta=\beta_\gamma$, $N=H$, and $\sum_{i=1}k_i=\sum_{h=1}^{H}k_{bh}$ gives:

$$[\sigma_{\gamma b}^2|...] \sim \mathrm{IG}\left(\alpha_\gamma+\frac{\sum_{h=1}^{H}k_{bh}}{2},\beta_\gamma+\frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b})^2\right) \tag{B.70}$$

$$[\sigma_{\theta b}^2|...] \propto \left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{(\sigma_{\theta b}^2)}\right)\prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\theta b}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right]$$

$$[\sigma_{\theta b}^2|...] \sim \mathrm{IG}\left(\alpha_\theta+\frac{\sum_{h=1}^{H}k_{bh}}{2},\beta_\theta+\frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}(\theta_{bj,h}-\mu_{\theta b})^2\right) \tag{B.71}$$

$$[\mu_{\gamma 0}|...] \propto \frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau_{\gamma 00}^2}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right] \tag{B.72}$$

Applying Result 1 with $\mu = \mu_{\gamma 0}$, $\sigma^2 = \tau^2_{\gamma 0}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = B$, and $x_b = \mu_{\gamma b}$ we have:

$$[\mu_{\gamma 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0} + \tau^2_{\gamma 00}\sum_{b=1}^{B}\mu_{\gamma b}}{\tau^2_{\gamma 0} + B\tau^2_{\gamma 00}}, \frac{\tau^2_{\gamma 0}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0} + B\tau^2_{\gamma 00}}\right) \tag{B.73}$$

$$[\mu_{\theta 0}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta 0} - \mu_{\theta 00})^2}{2\tau^2_{\theta 00}}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\theta b} - \mu_{\theta 0})^2}{2\tau^2_{\theta 0}}\right)\right]$$

$$[\mu_{\theta 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau^2_{\theta 0} + \tau^2_{\theta 00}\sum_{b=1}^{B}\mu_{\theta b}}{\tau^2_{\theta 0} + B\tau^2_{\theta 00}}, \frac{\tau^2_{\theta 0}\tau^2_{\theta 00}}{\tau^2_{\theta 0} + B\tau^2_{\theta 00}}\right) \tag{B.74}$$

$$[\tau^2_{\gamma 0}\Big|...] \propto \left(\tau^2_{\gamma 0}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{(\tau^2_{\gamma 0})}\right)\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\gamma 0}}}\exp\left(-\frac{(\mu_{\gamma b} - \mu_{\gamma 0})^2}{2\tau^2_{\gamma 0}}\right)\right] \tag{B.75}$$

Applying Result 5 with $\sigma^2 = \tau^2_{\gamma 0}$, $\mu = \mu_{\gamma 0}$, $x_b = \mu_{\gamma b}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, and $N = B$ gives:

$$[\tau^2_{\gamma 0}\Big|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00} + \frac{B}{2}, \beta_{\gamma 00} + \frac{1}{2}\sum_{b=1}^{B}(\mu_{\gamma b} - \mu_{\gamma 0})^2\right) \tag{B.76}$$

$$[\tau^2_{\theta 0}\Big|...] \propto \left(\tau^2_{\theta 0}\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{(\tau^2_{\theta 0})}\right)\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\theta 0}}}\exp\left(-\frac{(\mu_{\theta b} - \mu_{\theta 0})^2}{2\tau^2_{\theta 0}}\right)\right]$$

$$[\tau^2_{\theta 0}\Big|...] \sim \mathrm{IG}\left(\alpha_{\theta 00} + \frac{B}{2}, \beta_{\theta 00} + \frac{1}{2}\sum_{b=1}^{B}(\mu_{\theta b} - \mu_{\theta 0})^2\right) \tag{B.77}$$

B.11 Model $1a_{32}$

From §6.6.4 the joint posterior distribution for the parameters is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b,h}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 0}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\gamma,h}^{\alpha_{\gamma,h}}}{\Gamma(\alpha_{\gamma,h})}\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right]\\
&\quad\times\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 00}}}\exp\left(-\frac{\left(\mu_{\gamma 0}-\mu_{\gamma 00}\right)^2}{2\tau^2_{\gamma 00}}\right)\right]\\
&\quad\times\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau^2_{\gamma 0}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0}\right)}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 0}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0}\right)^2}{2\tau^2_{\theta 0}}\right)\right]\\
&\quad\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\theta,h}^{\alpha_{\theta,h}}}{\Gamma(\alpha_{\theta,h})}\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]\\
&\quad\times\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 00}}}\exp\left(-\frac{\left(\mu_{\theta 0}-\mu_{\theta 00}\right)^2}{2\tau^2_{\theta 00}}\right)\right]\\
&\quad\times\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau^2_{\theta 0}\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau^2_{\theta 0}\right)}\right)\right]
\end{aligned}
\tag{B.78}
$$

B.11.1 Complete Conditional Distributions

$$
\begin{aligned}
[\gamma_{bj,h}\Big|...] \propto & \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right] \\
& \times \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
& \times \left[\left(\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right)\right]
\end{aligned}
\tag{B.79}
$$

$$
\begin{aligned}
[\theta_{bj,h}\Big|...] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\
& \times \left[\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]
\end{aligned}
\tag{B.80}
$$

$$
[\mu_{\gamma b,h}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right] \tag{B.81}
$$

Applying Result 1 with $\mu = \mu_{\gamma b,h}$, $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = k_{bh}$, and $x_j = \gamma_{bj,h}$ we have:

$$
[\mu_{\gamma b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b,h} + \tau^2_{\gamma 0}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b,h}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}\right) \tag{B.82}
$$

$$[\mu_{\theta b,h}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right]$$

$$[\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b,h}^2+\tau_{\theta 0}^2\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma_{\theta b,h}^2+k_{bh}\tau_{\theta 0}^2}, \frac{\sigma_{\theta b,h}^2\tau_{\theta 0}^2}{\sigma_{\theta b,h}^2+k_{bh}\tau_{\theta 0}^2}\right) \tag{B.83}$$

$$[\sigma_{\gamma b,h}^2\Big|...] \propto \left(\sigma_{\gamma b,h}^2\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma_{\gamma b,h}^2\right)}\right)\times\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\gamma b,h}^2}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b,h})^2}{2\sigma_{\gamma b,h}^2}\right)\right] \tag{B.84}$$

Applying Result 5 with $\sigma^2=\sigma_{\gamma b,h}^2$, $\mu=\mu_{\gamma b,h}$, $x_j=\gamma_{bj,h}$, $\alpha=\alpha_\gamma$, $\beta=\beta_\gamma$, and $N=k_{bh}$ gives:

$$[\sigma_{\gamma b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma+\frac{k_{bh}}{2}, \beta_\gamma+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b,h})^2\right) \tag{B.85}$$

$$[\sigma_{\theta b,h}^2\Big|...] \propto \left(\sigma_{\theta b,h}^2\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma_{\theta b,h}^2\right)}\right)\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\theta b,h}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right]$$

$$[\sigma_{\theta b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\theta+\frac{k_{bh}}{2}, \beta_\theta+\frac{1}{2}\sum_{j=1}^{k_{bh}}(\theta_{bj,h}-\mu_{\theta b,h})^2\right) \tag{B.86}$$

$$[\mu_{\gamma 0}\Big|...] \propto \exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau_{\gamma 00}^2}\right)\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\gamma b,h}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right] \tag{B.87}$$

Applying Result 2 with $\mu = \mu_{\gamma 0}$, $\sigma^2 = \tau^2_{\gamma 0}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = H$, $\sum_{i=1}^{N} k_i = \sum_{h=1}^{H} B_h$, and $x_{hb} = \mu_{\gamma b,h}$ we have:

$$[\mu_{\gamma 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0} + \tau^2_{\gamma 00}\sum_{h=1}^{H}\sum_{b=1}^{B_h}\mu_{\gamma b,h}}{\tau^2_{\gamma 0} + \sum_{h=1}^{H} B_h \tau^2_{\gamma 00}}, \frac{\tau^2_{\gamma 0}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0} + \sum_{h=1}^{H} B_h \tau^2_{\gamma 00}}\right) \tag{B.88}$$

$$[\mu_{\theta 0}\Big|...] \propto \exp\left(-\frac{(\mu_{\theta 0} - \mu_{\theta 00})^2}{2\tau^2_{\theta 00}}\right)\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\theta b,h} - \mu_{\theta 0})^2}{2\tau^2_{\theta 0}}\right)\right] \tag{B.89}$$

$$[\mu_{\theta 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau^2_{\theta 0} + \tau^2_{\theta 00}\sum_{h=1}^{H}\sum_{b=1}^{B_h}\mu_{\theta b,h}}{\tau^2_{\theta 0} + \sum_{h=1}^{H} B_h \tau^2_{\theta 00}}, \frac{\tau^2_{\theta 0}\tau^2_{\theta 00}}{\tau^2_{\theta 0} + \sum_{h=1}^{H} B_h \tau^2_{\theta 00}}\right)$$

$$[\tau^2_{\gamma 0}\Big|...] \propto \left(\tau^2_{\gamma 0}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0}\right)}\right)\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\gamma 0}}}\exp\left(-\frac{(\mu_{\gamma b,h} - \mu_{\gamma 0})^2}{2\tau^2_{\gamma 0}}\right)\right] \tag{B.90}$$

Applying Result 6 with $\sigma^2 = \tau^2_{\gamma 0}$, $\mu = \mu_{\gamma 0}$, $x_{hb} = \mu_{\gamma b,h}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, $N = H$, and $\sum_{i=1}^{N} k_i \sum_{h=1}^{H} B_h$ gives:

$$[\tau^2_{\gamma 0}\Big|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00} + \frac{\sum_{h=1}^{H} B_h}{2}, \beta_{\gamma 00} + \frac{1}{2}\sum_{h=1}^{H}\sum_{b=1}^{B_h}(\mu_{\gamma b,h} - \mu_{\gamma 0})^2\right) \tag{B.91}$$

$$[\tau_{\theta 0}^2 \Big| ...] \propto \left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1} \exp\left(-\frac{\beta_{\theta 00}}{(\tau_{\theta 0}^2)}\right) \prod_{h=1}^{H} \prod_{b=1}^{B_h} \left[\left(\frac{1}{\sqrt{\tau_{\theta 0}^2}} \exp\left(-\frac{\left(\mu_{\theta b,h} - \mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right)\right]$$

$$[\tau_{\theta 0}^2 \Big| ...] \sim \text{IG}\left(\alpha_{\theta 00} + \frac{\sum_{h=1}^{H} B_h}{2}, \beta_{\theta 00} + \frac{1}{2}\sum_{h=1}^{H}\sum_{b=1}^{B_h}(\mu_{\theta b,h} - \mu_{\theta 0})^2\right) \tag{B.92}$$

B.12 Model $\mathbf{BB_{30}}$

The joint posterior distribution is proportional to:

$$
\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_{\pi,h}+\beta_{\pi,h})}{\Gamma(\alpha_{\pi,h})\Gamma(\beta_{\pi,h})}\pi_{b,h}^{\alpha_{\pi,h}-1}(1-\pi_{b,h})^{\beta_{\pi,h}-1}\right]
$$

$$
\times\prod_{h=1}^{H}\left[\frac{\lambda_\alpha\exp(-\alpha_{\pi,h}\lambda_\alpha)}{\exp(-\lambda_\alpha)}I(\alpha_{\pi,h}>1)\right]\prod_{h=1}^{H}\left[\frac{\lambda_\beta\exp(-\beta_{\pi,h}\lambda_\beta)}{\exp(-\lambda_\beta)}I(\beta_{\pi,h}>1)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\gamma b,h}}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 0,h}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\gamma,h}^{\alpha_{\gamma,h}}}{\Gamma(\alpha_{\gamma,h})}\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\left[\frac{1}{\sqrt{2\pi\tau^2_{\gamma 00}}}\exp\left(-\frac{\left(\mu_{\gamma 0,h}-\mu_{\gamma 00}\right)^2}{2\tau^2_{\gamma 00}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau^2_{\gamma 0,h}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0,h}\right)}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 0,h}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0,h}\right)^2}{2\tau^2_{\theta 0,h}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\theta,h}^{\alpha_{\theta,h}}}{\Gamma(\alpha_{\theta,h})}\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\left[\frac{1}{\sqrt{2\pi\tau^2_{\theta 00}}}\exp\left(-\frac{\left(\mu_{\theta 0,h}-\mu_{\theta 00}\right)^2}{2\tau^2_{\theta 00}}\right)\right]
$$

$$
\times\prod_{h=1}^{H}\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau^2_{\theta 0,h}\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau^2_{\theta 0,h}\right)}\right)\right] \tag{B.93}
$$

B.12.1 Complete Conditional Distributions

$$
\begin{aligned}
\left[\gamma_{bj,h}\middle|...\right] \propto & \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}} C_{bj,h}}\right] \\
& \times \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)} T_{bj,h}}\right] \\
& \times \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]
\end{aligned}
\tag{B.94}
$$

$$
\begin{aligned}
\left[\theta_{bj,h}\middle|...\right] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)} T_{bj,h}}\right] \\
& \times \left[\pi_{b,h}\, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\, \mathrm{I}_{[\theta_{bj,h}\neq 0]} \frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}} \exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]
\end{aligned}
\tag{B.95}
$$

$$
\left[\mu_{\gamma b,h}\middle|...\right] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right] \prod_{j=1}^{k_{bh}} \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right] \tag{B.96}
$$

Applying Result 1 with $\mu = \mu_{\gamma b,h}$, $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu_0 = \mu_{\gamma 0,h}$, $\sigma^2_0 = \tau^2_{\gamma 0,h}$, $N = k_{bh}$, and $x_j = \gamma_{bj,h}$ we have:

$$
\left[\mu_{\gamma b,h}\middle|...\right] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0,h}\sigma^2_{\gamma b,h} + \tau^2_{\gamma 0,h}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0,h}}, \frac{\sigma^2_{\gamma b,h}\tau^2_{\gamma 0,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0,h}}\right) \tag{B.97}
$$

$$[\mu_{\theta b,h}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0,h})^2}{2\tau^2_{\theta 0,h}}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma^2_{\theta b,h}}\right)\right] \tag{B.98}$$

Applying Result 3 with $\mu = \mu_{\theta b,h}$, $\sigma^2 = \sigma^2_{\theta b,h}$, $\mu_0 = \mu_{\theta 0,h}$, $\sigma^2_0 = \tau^2_{\theta 0,h}$, $N = k_{bh}$, and $x_j = \theta_{bj,h}$ we have:

$$[\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0,h}\sigma^2_{\theta b,h} + \tau^2_{\theta 0,h}\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma^2_{\theta b,h} + K_{bh}\tau^2_{\theta 0,h}}, \frac{\sigma^2_{\theta b,h}\tau^2_{\theta 0,h}}{\sigma^2_{\theta b,h} + K_{bh}\tau^2_{\theta 0,h}}\right) \tag{B.99}$$

where $K_{bh} = \sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]}$.

$$[\sigma_{\gamma b,h}\Big|...] \propto \left[\left(\sigma^2_{\gamma b,h}\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma^2_{\gamma b,h}\right)}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\left(\frac{1}{\sqrt{\sigma^2_{\gamma b,h}}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b,h})^2}{2\sigma^2_{\gamma b,h}}\right)\right)\right] \tag{B.100}$$

Applying Result 5 with $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu = \mu_{\gamma b,h}$, $x_j = \gamma_{bj,h}$, $\alpha = \alpha_{\gamma,h}$, $\beta = \beta_{\gamma,h}$, $N = k_{bh}$ gives:

$$[\sigma_{\gamma b,h}\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma + \frac{k_{bh}}{2}, \beta_\gamma + \frac{1}{2}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b,h})^2\right) \tag{B.101}$$

$$[\sigma_{\theta b,h}\Big|...] \propto \left[\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]$$
$$\times\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right] \tag{B.102}$$

Applying Result 8 with $\sigma^2=\sigma^2_{\theta b,h}$, $\mu=\mu_{\theta b,h}$, $x_j=\theta_{bj,h}$, $\alpha=\alpha_{\theta,h}$, $\beta=\beta_{\theta,h}$, $N=k_{bh}$ gives:

$$\sigma_{\theta b,h}\Big|...] \sim \mathrm{IG}\left(\alpha_\theta+\frac{K_{bh}}{2},\beta_\theta+\frac{1}{2}\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}(\theta_{bj,h}-\mu_{\theta b,h})^2\right) \tag{B.103}$$

where $K_{bh}=\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}$.

$$[\mu_{\gamma 0,h}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\gamma 0,h}-\mu_{\gamma 00,h}\right)^2}{2\tau^2_{\gamma 00}}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right] \tag{B.104}$$

Applying Result 1 with $\mu=\mu_{\gamma 0,h}$, $\sigma^2=\tau^2_{\gamma 0,h}$, $\mu_0=\mu_{\gamma 00}$, $\sigma_0^2=\tau^2_{\gamma 00}$, $N=B_h$, and $x_b=\mu_{\gamma b,h}$ we have:

$$[\mu_{\gamma 0,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0,h}+\tau^2_{\gamma 00}\sum_{b=1}^{B_h}\mu_{\gamma b,h}}{\tau^2_{\gamma 0,h}+B_h\tau^2_{\gamma 00}},\frac{\tau^2_{\gamma 0,h}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0,h}+B_h\tau^2_{\gamma 00}}\right) \tag{B.105}$$

$$[\mu_{\theta 0,h}\Big|...] \propto \exp\left(-\frac{\left(\mu_{\theta 0,h}-\mu_{\theta 00}\right)^2}{2\tau^2_{\theta 00}}\right)\prod_{b=1}^{B_h}\left[\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0,h}\right)^2}{2\tau^2_{\theta 0,h}}\right)\right]$$

$$[\mu_{\theta 0,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau^2_{\theta 0,h}+\tau^2_{\theta 00}\sum_{b=1}^{B_h}\mu_{\theta b,h}}{\tau^2_{\theta 0,h}+B_h\tau^2_{\theta 00}}, \frac{\tau^2_{\theta 0,h}\tau^2_{\theta 00}}{\tau^2_{\theta 0,h}+B_h\tau^2_{\theta 00}}\right) \tag{B.106}$$

$$[\tau^2_{\gamma 0,h}\Big|...] \propto \left[\left(\tau^2_{\gamma 0,h}\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau^2_{\gamma 0,h}\right)}\right)\right]\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\gamma 0,h}}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0,h}\right)^2}{2\tau^2_{\gamma 0,h}}\right)\right] \tag{B.107}$$

Applying Result 5 with $\sigma^2 = \tau^2_{\gamma 0,h}$, $\mu = \mu_{\gamma 0,h}$, $x_b = \mu_{\gamma b,h}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, $N = B_h$ gives:

$$\tau^2_{\gamma 0,h}\Big|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00}+\frac{B_h}{2}, \beta_{\gamma 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\gamma b,h}-\mu_{\gamma 0,h})^2\right) \tag{B.108}$$

$$[\tau^2_{\theta 0,h}\Big|...] \propto \left[\left(\tau^2_{\theta 0,h}\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau^2_{\theta 0,h}\right)}\right)\right]\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau^2_{\theta 0,h}}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0,h}\right)^2}{2\tau^2_{\theta 0,h}}\right)\right]$$

$$\tau^2_{\theta 0,h}\Big|...] \sim \mathrm{IG}\left(\alpha_{\theta 00}+\frac{B_h}{2}, \beta_{\theta 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\theta b,h}-\mu_{\theta 0,h})^2\right) \tag{B.109}$$

$$
\begin{aligned}
[\pi_{b,h}\Big|...] &\propto \left[\pi_{b,h}^{\alpha_{\pi,h}-1}(1-\pi_{b,h})^{\beta_{\pi,h}-1}\right] \\
&\times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}=0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma^2_{\theta b,h}}\right)\right] \\
&\propto \left[\pi_{b,h}^{\alpha_{\pi,h}-1}(1-\pi_{b,h})^{\beta_{\pi,h}-1}\right]\left[\pi_{b,h}^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}(1-\pi_{b,h})^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}}\right] \\
&= \pi_{b,h}^{\alpha_{\pi,h}-1+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}(1-\pi_{b,h})^{\beta_{\pi,h}-1+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}} \\
&= \pi_{b,h}^{\alpha_{\pi,h}-1+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}(1-\pi_{b,h})^{\beta_{\pi,h}-1+k_{bh}-\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}
\end{aligned}
$$

$$
[\pi_{b,h}\Big|...] \sim \text{Beta}\left(\alpha_{\pi,h}+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}, \beta_{\pi,h}+k_{bh}-\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}\right) \tag{B.110}
$$

$$
[\alpha_{\pi,h}\Big|...] \propto \left[\exp(-\alpha_{\pi,h}\lambda_\alpha)\,\mathrm{I}_{[\alpha_{\pi,h}>1]}\right]\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_{\pi,h}+\beta_{\pi,h})}{\Gamma(\alpha_{\pi,h})}\pi_{b,h}^{\alpha_{\pi,h}-1}\right] \tag{B.111}
$$

$$
[\beta_{\pi,h}\Big|...] \propto \left[\exp(-\beta_{\pi,h}\lambda_\beta)\,\mathrm{I}_{[\beta_{\pi,h}>1]}\right]\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_{\pi,h}+\beta_{\pi,h})}{\Gamma(\beta_{\pi,h})}(1-\pi_{b,h})^{\beta_{\pi,h}-1}\right] \tag{B.112}
$$

B.13 Model $\mathbf{BB_{31}}$

The joint posterior distribution is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\times\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}\pi_b^{\alpha_\pi-1}(1-\pi_b)^{\beta_\pi-1}\right]\\
&\times\left[\frac{\lambda_\alpha\exp(-\alpha_\pi\lambda_\alpha)}{\exp(-\lambda_\alpha)}I(\alpha_\pi>1)\right]\\
&\times\left[\frac{\lambda_\beta\exp(-\beta_\pi\lambda_\beta)}{\exp(-\lambda_\beta)}I(\beta_\pi>1)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma_{\gamma b}^2}\right)\right]\\
&\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\times\prod_{b=1}^{B_h}\left[\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma-1}\exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\right]\\
&\times\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{\left(\mu_{\gamma 0}-\mu_{\gamma 00}\right)^2}{2\tau_{\gamma 00}^2}\right)\right]\\
&\times\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau_{\gamma 0}^2\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau_{\gamma 0}^2\right)}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\pi_b\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_b)\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma_{\theta b}^2}\right)\right]\\
&\times\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{\left(\mu_{\theta b}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\times\prod_{b=1}^{B_h}\left[\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\left(\sigma_{\theta b}^2\right)^{-\alpha_\theta-1}\exp\left(-\frac{\beta_\theta}{\left(\sigma_{\theta b}^2\right)}\right)\right]\\
&\times\left[\frac{1}{\sqrt{2\pi\tau_{\theta 00}^2}}\exp\left(-\frac{\left(\mu_{\theta 0}-\mu_{\theta 00}\right)^2}{2\tau_{\theta 00}^2}\right)\right]\\
&\times\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau_{\theta 0}^2\right)}\right)\right]
\end{aligned}
\tag{B.113}
$$

B.13.1 Complete Conditionals Distributions

$$
\begin{aligned}
\left[\gamma_{bj,h}\middle|...\right] \propto & \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}} C_{bj,h}}\right] \\
& \times \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)} T_{bj,h}}\right] \\
& \times \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right]
\end{aligned}
\tag{B.114}
$$

$$
\begin{aligned}
\left[\theta_{bj,h}\middle|...\right] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)} T_{bj,h}}\right] \\
& \times \left[\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b}\right)^2}{2\sigma^2_{\theta b}}\right)\right]
\end{aligned}
\tag{B.115}
$$

$$
\left[\mu_{\gamma b}\middle|...\right] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma b}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right] \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}} \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b}\right)^2}{2\sigma^2_{\gamma b}}\right)\right] \tag{B.116}
$$

Applying Result 2 with $\mu = \mu_{\gamma b}$, $\sigma^2 = \sigma^2_{\gamma b}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = H$, and $x_{hj} = \gamma_{bj,h}$ we have:

$$
\left[\mu_{\gamma b}\middle|...\right] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b} + \tau^2_{\gamma 0}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b} + \sum_{h=1}^{H} k_{bh}\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b} + \sum_{h=1}^{H} k_{bh}\tau^2_{\gamma 0}}\right) \tag{B.117}
$$

$$[\mu_{\theta b}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}} \exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right] \tag{B.118}$$

Applying Result 4 with $\mu = \mu_{\theta b}$, $\sigma^2 = \sigma_{\theta b}^2$, $\mu_0 = \mu_{\theta 0}$, $\sigma_0^2 = \tau_{\theta 0}^2$, $N = H$, and $x_{hj} = \theta_{bj,h}$ we have:

$$[\mu_{\theta b}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b}^2 + \tau_{\theta 0}^2 \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \theta_{bj,h}}{\sigma_{\theta b}^2 + \sum_{h=1}^{H} K_{bh}\tau_{\theta 0}^2}, \frac{\sigma_{\theta b}^2\tau_{\theta 0}^2}{\sigma_{\theta b}^2 + \sum_{h=1}^{H} K_{bh}\tau_{\theta 0}^2}\right) \tag{B.119}$$

where $K_{bh} = \sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}]\neq 0}$.

$$[\sigma_{\gamma b}^2\Big|...] \propto \left[\left(\sigma_{\gamma b}^2\right)^{-\alpha_\gamma - 1} \exp\left(-\frac{\beta_\gamma}{\left(\sigma_{\gamma b}^2\right)}\right)\right] \times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\left(\frac{1}{\sqrt{\sigma_{\gamma b}^2}} \exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b})^2}{2\sigma_{\gamma b}^2}\right)\right)\right] \tag{B.120}$$

Applying Result 6 with $\sigma^2 = \sigma_{\gamma b}^2$, $\mu = \mu_{\gamma b}$, $x_{hj} = \gamma_{bj,h}$, $\alpha = \alpha_\gamma$, $\beta = \beta_\gamma$, and $N = H$ gives:

$$\sigma_{\gamma b}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma + \frac{\sum_{h=1}^{H} k_{bh}}{2}, \beta_\gamma + \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b})^2\right) \tag{B.121}$$

$$[\sigma^2_{\theta b}\Big|...] \propto \left[\left(\sigma^2_{\theta b}\right)^{-\alpha_\theta - 1} \exp\left(-\frac{\beta_\theta}{(\sigma^2_{\theta b})}\right)\right]$$
$$\times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}\neq 0]} \frac{1}{\sqrt{2\pi\sigma^2_{\theta b}}} \exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma^2_{\theta b}}\right)\right]$$

$$\sigma^2_{\theta b}\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{\sum_{h=1}^{H} k_{bh}}{2}, \beta_\theta + \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}}(\theta_{bj,h}-\mu_{\theta b})^2\right) \tag{B.122}$$

Applying Result 8 with $\alpha = \alpha_\theta$, $\beta = \beta_\theta$, $\mu = \mu_{\theta b}$, $\sigma^2 = \sigma^2_{\theta b}$, $x_{hj} = \theta_{bj,h}$, and $N = H$ we have:

$$\sigma^2_{\theta b}\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{\sum_{h=1}^{H} K_{bh}}{2}, \beta_\theta + \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}]\neq 0}(\theta_{bj,h}-\mu_{\theta b})^2\right) \tag{B.123}$$

where $K_{bh} = \sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}]\neq 0}$.

$$[\mu_{\gamma 0}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\gamma 0}-\mu_{\gamma 00})^2}{2\tau^2_{\gamma 00}}\right)\right]\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau^2_{\gamma 0}}\right)\right] \tag{B.124}$$

Applying Result 1 with $\mu = \mu_{\gamma 0}$, $\sigma^2 = \tau^2_{\gamma 0}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = B_h$, and $x_b = \mu_{\gamma b}$ we have:

$$[\mu_{\gamma 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0} + \tau^2_{\gamma 00}\sum_{b=1}^{B_h}\mu_{\gamma,b}}{\tau^2_{\gamma 0} + B_h\tau^2_{\gamma 00}}, \frac{\tau^2_{\gamma 0}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0} + B_h\tau^2_{\gamma 00}}\right) \tag{B.125}$$

$$[\mu_{\theta 0}|...] \propto \left[\exp\left(-\frac{(\mu_{\theta 0}-\mu_{\theta 00})^2}{2\tau_{\theta 00}^2}\right)\right]\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right]$$

$$[\mu_{\theta 0}|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau_{\theta 0}^2+\tau_{\theta 00}^2\sum_{b=1}^{B_h}\mu_{\theta,b}}{\tau_{\theta 0}^2+B_h\tau_{\theta 00}^2}, \frac{\tau_{\theta 0}^2\tau_{\theta 00}^2}{\tau_{\theta 0}^2+B_h\tau_{\theta 00}^2}\right) \tag{B.126}$$

$$[\tau_{\gamma 0}^2|...] \propto \left[\left(\tau_{\gamma 0}^2\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau_{\gamma 0}^2\right)}\right)\right]\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau_{\gamma 0}^2}}\exp\left(-\frac{(\mu_{\gamma b}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right] \tag{B.127}$$

Applying Result 5 with $\sigma^2 = \tau_{\gamma 0}^2$, $\mu = \mu_{\gamma 0}$, $x_b = \mu_{\gamma b}$, $\alpha = \alpha_{\gamma 00}$, $\beta = \beta_{\gamma 00}$, and $N = B_h$ gives:

$$[\tau_{\gamma 0}^2|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00}+\frac{B_h}{2}, \beta_{\gamma 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\gamma b}-\mu_{\gamma 0})^2\right) \tag{B.128}$$

$$[\tau_{\theta 0}^2|...] \propto \left[\left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau_{\theta 0}^2\right)}\right)\right]\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau_{\theta 0}^2}}\exp\left(-\frac{(\mu_{\theta b}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right]$$

$$[\tau_{\theta 0}^2|...] \sim \mathrm{IG}\left(\alpha_{\theta 00}+\frac{B_h}{2}, \alpha_{\theta 00}+\frac{1}{2}\sum_{b=1}^{B_h}(\mu_{\theta b}-\mu_{\theta 0})^2\right) \tag{B.129}$$

$$
\begin{aligned}
[\pi_b \Big| ...] &\propto \left[\pi_b^{\alpha_\pi - 1}(1-\pi_b)^{\beta_\pi - 1}\right] \\
&\times \prod_{h=1}^{H}\prod_{j=1}^{k_{bh}}\left[\pi_b \, \mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_b)\, \mathrm{I}_{[\theta_{bj,h}\neq 0]} \, \frac{1}{\sqrt{2\pi\sigma_{\theta b}^2}} \exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b})^2}{2\sigma_{\theta b}^2}\right)\right] \\
&\propto \left[\pi_b^{\alpha_\pi - 1}(1-\pi_b)^{\beta_\pi - 1}\right] \\
&\times \left[\left(\pi_b^{\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]}} (1-\pi_b)^{\sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]}}\right)\right] \\
&= \left[\pi_b^{\alpha_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]} - 1} (1-\pi_b)^{\beta_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]} - 1}\right]
\end{aligned}
$$

$$
[\pi_b \Big| ...] \sim \mathrm{Beta}\left(\alpha_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}=0]}, \beta_\pi + \sum_{h=1}^{H}\sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}\neq 0]}\right) \tag{B.130}
$$

$$
[\alpha_\pi \Big| ...] \propto \times \left[\exp(-\alpha_\pi \lambda_\alpha)\, \mathrm{I}_{[\alpha_\pi > 1]}\right] \prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)} \pi_b^{\alpha_\pi - 1}\right] \tag{B.131}
$$

$$
[\beta_\pi] \propto \left[\exp(-\beta_\pi \lambda_\beta)\, \mathrm{I}_{[\beta_\pi > 1]}\right] \prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\beta_\pi)} (1-\pi_b)^{\beta_\pi - 1}\right] \tag{B.132}
$$

B.14 Model $\mathbf{BB_{32}}$

The joint posterior distribution is proportional to:

$$
\begin{aligned}
&\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\gamma_{bj,h}}C_{bj,h}\right)^{x_{bj,h}}\,e^{-e^{\gamma_{bj,h}}C_{bj,h}}}{x_{bj,h}!}\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}\right)^{y_{bj,h}}\,e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}}{y_{bj,h}!}\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)}\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right]\\
&\times\prod_{h=1}^{H}\left[\frac{\lambda_\alpha\exp(-\alpha_\pi\lambda_\alpha)}{\exp(-\lambda_\alpha)}I(\alpha_\pi>1)\right]\prod_{h=1}^{H}\left[\frac{\lambda_\beta\exp(-\beta_\pi\lambda_\beta)}{\exp(-\lambda_\beta)}I(\beta_\pi>1)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{2\pi\sigma_{\gamma b,h}^2}}\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma_{\gamma b,h}^2}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 0}^2}}\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau_{\gamma 0}^2}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\gamma,h}^{\alpha_{\gamma,h}}}{\Gamma(\alpha_{\gamma,h})}\left(\sigma_{\gamma b,h}^2\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma_{\gamma b,h}^2\right)}\right)\right]\\
&\times\left[\frac{1}{\sqrt{2\pi\tau_{\gamma 00}^2}}\exp\left(-\frac{\left(\mu_{\gamma 0}-\mu_{\gamma 00}\right)^2}{2\tau_{\gamma 00}^2}\right)\right]\\
&\times\left[\frac{\beta_{\gamma 00}^{\alpha_{\gamma 00}}}{\Gamma(\alpha_{\gamma 00})}\left(\tau_{\gamma 0}^2\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau_{\gamma 0}^2\right)}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b,h}^2}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma_{\theta b,h}^2}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{2\pi\tau_{\theta 0}^2}}\exp\left(-\frac{\left(\mu_{\theta b,h}-\mu_{\theta 0}\right)^2}{2\tau_{\theta 0}^2}\right)\right]\\
&\times\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\beta_{\theta,h}^{\alpha_{\theta,h}}}{\Gamma(\alpha_{\theta,h})}\left(\sigma_{\theta b,h}^2\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma_{\theta b,h}^2\right)}\right)\right]\\
&\times\left[\frac{1}{\sqrt{2\pi\tau_{\theta 00}^2}}\exp\left(-\frac{\left(\mu_{\theta 0}-\mu_{\theta 00}\right)^2}{2\tau_{\theta 00}^2}\right)\right]\\
&\times\left[\frac{\beta_{\theta 00}^{\alpha_{\theta 00}}}{\Gamma(\alpha_{\theta 00})}\left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau_{\theta 0}^2\right)}\right)\right]
\end{aligned}
\tag{B.133}
$$

B.14.1 Complete Conditional Distributions

$$\begin{aligned}\left[\gamma_{bj,h}\middle|...\right] \propto & \left[\left(e^{\gamma_{bj,h}}\right)^{x_{bj,h}} e^{-e^{\gamma_{bj,h}}C_{bj,h}}\right] \\ & \times \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\ & \times \left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right]\end{aligned} \tag{B.134}$$

$$\begin{aligned}\left[\theta_{bj,h}\middle|...\right] \propto & \left[\left(e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}\right)^{y_{bj,h}} e^{-e^{\left(\gamma_{bj,h}+\theta_{bj,h}\right)}T_{bj,h}}\right] \\ & \times \left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right]\end{aligned} \tag{B.135}$$

$$\left[\mu_{\gamma b,h}\middle|...\right] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\exp\left(-\frac{\left(\gamma_{bj,h}-\mu_{\gamma b,h}\right)^2}{2\sigma^2_{\gamma b,h}}\right)\right] \tag{B.136}$$

Applying Result 1 with $\mu = \mu_{\gamma b,h}$, $\sigma^2 = \sigma^2_{\gamma b,h}$, $\mu_0 = \mu_{\gamma 0}$, $\sigma^2_0 = \tau^2_{\gamma 0}$, $N = k_{bh}$, and $x_j = \gamma_{bj,h}$ we have:

$$\left[\mu_{\gamma b,h}\middle|...\right] \sim \mathrm{N}\left(\frac{\mu_{\gamma 0}\sigma^2_{\gamma b,h} + \tau^2_{\gamma 0}\sum_{j=1}^{k_{bh}}\gamma_{bj,h}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}, \frac{\sigma^2_{\gamma b,h}\tau^2_{\gamma 0}}{\sigma^2_{\gamma b,h} + k_{bh}\tau^2_{\gamma 0}}\right) \tag{B.137}$$

$$[\mu_{\theta b,h}\Big|...] \propto \left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]} + (1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma_{\theta b,h}^2}}\exp\left(-\frac{(\theta_{bj,h}-\mu_{\theta b,h})^2}{2\sigma_{\theta b,h}^2}\right)\right] \tag{B.138}$$

Applying Result 3 with $\mu = \mu_{\theta b,h}$, $\sigma^2 = \sigma_{\theta b,h}^2$, $\mu_0 = \mu_{\theta 0}$, $\sigma_0^2 = \tau_{\theta 0}^2$, $N = k_{bh}$, and $x_j = \theta_{bj,h}$ we have:

$$[\mu_{\theta b,h}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 0}\sigma_{\theta b,h}^2 + \tau_{\theta 0}^2\sum_{j=1}^{k_{bh}}\theta_{bj,h}}{\sigma_{\theta b,h}^2 + K_{bh}\tau_{\theta 0}^2}, \frac{\sigma_{\theta b,h}^2\tau_{\theta 0}^2}{\sigma_{\theta b,h}^2 + K_{bh}\tau_{\theta 0}^2}\right) \tag{B.139}$$

where $K_{bh} = \sum_{j=1}^{k_{bh}} \mathrm{I}_{[\theta_{bj,h}]\neq 0}$.

$$[\sigma_{\gamma b,h}^2\Big|...] \propto \left[\left(\sigma_{\gamma b,h}^2\right)^{-\alpha_{\gamma,h}-1}\exp\left(-\frac{\beta_{\gamma,h}}{\left(\sigma_{\gamma b,h}^2\right)}\right)\right] \times \prod_{j=1}^{k_{bh}}\left[\frac{1}{\sqrt{\sigma_{\gamma b,h}^2}}\exp\left(-\frac{(\gamma_{bj,h}-\mu_{\gamma b,h})^2}{2\sigma_{\gamma b,h}^2}\right)\right] \tag{B.140}$$

Applying Result 5 with $\sigma^2 = \sigma_{\gamma b,h}^2$, $\mu = \mu_{\gamma b,h}$, $x_j = \gamma_{bj,h}$, $\alpha = \alpha_{\gamma,h}$, $\beta = \beta_{\gamma,h}$, and $N = k_{bh}$ gives:

$$[\sigma_{\gamma b,h}^2\Big|...] \sim \mathrm{IG}\left(\alpha_\gamma + \frac{k_{bh}}{2}, \beta_\gamma + \frac{1}{2}\sum_{i=j}^{k_{bh}}(\gamma_{bj,h}-\mu_{\gamma b,h})^2\right) \tag{B.141}$$

$$[\sigma^2_{\theta b,h}\Big|...] \propto \left[\left(\sigma^2_{\theta b,h}\right)^{-\alpha_{\theta,h}-1}\exp\left(-\frac{\beta_{\theta,h}}{\left(\sigma^2_{\theta b,h}\right)}\right)\right]$$
$$\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right] \tag{B.142}$$

Applying Result 7 with $\sigma^2 = \sigma^2_{\theta b,h}$, $\mu = \mu_{\theta b,h}$, $x_j = \theta_{bj,h}$, $\alpha = \alpha_{\theta,h}$, $\beta = \beta_{\theta,h}$, and $N = k_{bh}$ gives:

$$[\sigma^2_{\theta b,h}\Big|...] \sim \mathrm{IG}\left(\alpha_\theta + \frac{K_{bh}}{2}, \beta_\theta + \frac{1}{2}\sum_{i=j}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}(\theta_{bj,h}-\mu_{\theta b,h})^2\right) \tag{B.143}$$

where $K_{bh} = \sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}]\neq 0}$.

$$[\mu_{\gamma 0}\Big|...] \propto \left[\exp\left(-\frac{\left(\mu_{\gamma 0}-\mu_{\gamma 00}\right)^2}{2\tau^2_{\gamma 00}}\right)\right] \times \prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\exp\left(-\frac{\left(\mu_{\gamma b,h}-\mu_{\gamma 0}\right)^2}{2\tau^2_{\gamma 0}}\right)\right] \tag{B.144}$$

Applying Result 2 with $\mu = \mu_{\gamma 0,h}$, $\sigma^2 = \tau^2_{\gamma 0}$, $\mu_0 = \mu_{\gamma 00}$, $\sigma^2_0 = \tau^2_{\gamma 00}$, $N = H$, and $x_{hb} = \mu_{\gamma b,h}$ we have:

$$[\mu_{\gamma 0}\Big|...] \sim \mathrm{N}\left(\frac{\mu_{\gamma 00}\tau^2_{\gamma 0} + \tau^2_{\gamma 00}\sum_{h=1}^{H}\sum_{b=1}^{B_h}\mu_{\gamma b,h}}{\tau^2_{\gamma 0,h} + \sum_{h=1}^{H}B_h\tau^2_{\gamma 00}}, \frac{\tau^2_{\gamma 0}\tau^2_{\gamma 00}}{\tau^2_{\gamma 0} + \sum_{h=1}^{H}B_h\tau^2_{\gamma 00}}\right) \tag{B.145}$$

$$[\mu_{\theta 0}|...] \propto \left[\exp\left(-\frac{(\mu_{\theta 0}-\mu_{\theta 00})^2}{2\tau_{\theta 00}^2}\right)\right]\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right] \tag{B.146}$$

$$[\mu_{\theta 0}|...] \sim \mathrm{N}\left(\frac{\mu_{\theta 00}\tau_{\theta 0}^2+\tau_{\theta 00}^2\sum_{h=1}^{H}\sum_{b=1}^{B_h}\mu_{\theta b,h}}{\tau_{\theta 0}^2+\sum_{h=1}^{H}B_h\tau_{\theta 00}^2}, \frac{\tau_{\theta 0}^2\tau_{\theta 00}^2}{\tau_{\theta 0}^2+\sum_{h=1}^{H}B_h\tau_{\theta 00}^2}\right)$$

$$[\tau_{\gamma 0}^2|...] \propto \left[\left(\tau_{\gamma 0}^2\right)^{-\alpha_{\gamma 00}-1}\exp\left(-\frac{\beta_{\gamma 00}}{\left(\tau_{\gamma 0}^2\right)}\right)\right]\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau_{\gamma 0}^2}}\exp\left(-\frac{(\mu_{\gamma b,h}-\mu_{\gamma 0})^2}{2\tau_{\gamma 0}^2}\right)\right] \tag{B.147}$$

Applying Result 6 with $\sigma^2=\tau_{\gamma 0}^2$, $\mu=\mu_{\gamma 0}$, $x_{bj}=\mu_{\gamma b,h}$, $\alpha=\alpha_{\gamma,00}$, $\beta=\beta_{\gamma,00}$, and $N=H$ gives:

$$[\tau_{\gamma 0}^2|...] \sim \mathrm{IG}\left(\alpha_{\gamma 00}+\frac{\sum_{h=1}^{H}B_h}{2}, \beta_{\gamma 00}+\frac{1}{2}\sum_{h=1}^{H}\sum_{b=1}^{B_h}(\mu_{\gamma b,h}-\mu_{\gamma 0})^2\right) \tag{B.148}$$

$$[\tau_{\theta 0}^2|...] \propto \left[\left(\tau_{\theta 0}^2\right)^{-\alpha_{\theta 00}-1}\exp\left(-\frac{\beta_{\theta 00}}{\left(\tau_{\theta 0}^2\right)}\right)\right]\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{1}{\sqrt{\tau_{\theta 0}^2}}\exp\left(-\frac{(\mu_{\theta b,h}-\mu_{\theta 0})^2}{2\tau_{\theta 0}^2}\right)\right]$$

$$[\tau_{\theta 0}^2|...] \sim \mathrm{IG}\left(\alpha_{\theta 00}+\frac{\sum_{h=1}^{H}B_h}{2}, \beta_{\theta 00}+\frac{1}{2}\sum_{h=1}^{H}\sum_{b=1}^{B_h}(\mu_{\theta b,h}-\mu_{\theta 0})^2\right) \tag{B.149}$$

$$
\begin{aligned}
\left[\pi_{b,h}\middle|...\right] &\propto \left[\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right] \\
&\quad\times\prod_{j=1}^{k_{bh}}\left[\pi_{b,h}\,\mathrm{I}_{[\theta_{bj,h}=0]}+(1-\pi_{b,h})\,\mathrm{I}_{[\theta_{bj,h}\neq 0]}\,\frac{1}{\sqrt{2\pi\sigma^2_{\theta b,h}}}\exp\left(-\frac{\left(\theta_{bj,h}-\mu_{\theta b,h}\right)^2}{2\sigma^2_{\theta b,h}}\right)\right] \\
&\propto \left[\pi_{b,h}^{\alpha_\pi-1}(1-\pi_{b,h})^{\beta_\pi-1}\right]\left[\left(\pi_{b,h}^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}}(1-\pi_{b,h})^{\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}}\right)\right] \\
&= \left[\pi_{b,h}^{\alpha_\pi+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}-1}(1-\pi_{b,h})^{\beta_\pi+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}\neq 0]}-1}\right]
\end{aligned}
$$

$$
\left[\pi_{b,h}\middle|...\right] \sim \mathrm{Beta}\left(\alpha_\pi+\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]},\beta_\pi+k_{bh}-\sum_{j=1}^{k_{bh}}\mathrm{I}_{[\theta_{bj,h}=0]}\right) \tag{B.150}
$$

$$
\left[\alpha_\pi\middle|...\right] \propto \left[\exp(-\alpha_\pi\lambda_\alpha)\,\mathrm{I}_{(\alpha_\pi>1)}\right]\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\alpha_\pi)}\pi_{b,h}^{\alpha_\pi-1}\right] \tag{B.151}
$$

$$
\left[\beta_\pi\middle|...\right] \propto \left[\exp(-\beta_\pi\lambda_\beta)\,\mathrm{I}_{[\beta_\pi>1]}\right]\prod_{h=1}^{H}\prod_{b=1}^{B_h}\left[\frac{\Gamma(\alpha_\pi+\beta_\pi)}{\Gamma(\beta_\pi)}(1-\pi_{b,h})^{\beta_\pi-1}\right] \tag{B.152}
$$

Appendix C

Model Tuning and Monitoring Convergence

C.1 Approach to Determining Approximate Convergence

In order to make inferences from Markov Chain Monte Carlo (MCMC) based models, such as those used in this study, the samples generated must have converged, at least approximately, to the Markov chain's stationary distribution [122]. There is no definitive diagnostic tool for assessing MCMC convergence but there are a number of approaches, both graphical and numerical, which may indicate non-convergence to stationarity. Assessing convergence of the MCMC fit by graphical means is suitable for a small number of parameters. However given the large number of parameters in the models in this study, and the large number of simulations, it is not feasible to inspect them all visually, and a more general approach is needed. In this study the Gelman-Rubin (GR) convergence diagnostic is the method chosen to assess convergence for the general simulations where applicable [122]. This statistic can be considered to provide a measure of similarity between the parallel chains when they start from over-dispersed initial values. Values close to 1 are considered to be consistent with convergence and Gelman et al. ([122, pg. 501]) use a check that the values are less than 1.2, while Kenneth ([153]) suggests a value of less than 1.1, when monitoring for convergence. While checking this statistic may indicate issues with convergence, it does not in itself indicate that the MCMC simulation has converged.

When using the GR statistic to assess the convergence of the θ parameters in models with a point-mass we need to consider how the point-mass may influence the statistic. The GR statistic is based on the means and variances of the generated MCMC samples and may not work as well for distributions which are far

from normal [122], which is the case for the θ parameters in the point-mass models. These parameters are absolutely continuous with respect to a mixture of the Lebesgue measure and a point-mass at zero. In this case transformation of the data to approximate normality is not an option, so we consider the GR statistic only as an assessment of a level of similarity between the different chains, which we might expect to see at approximate convergence. If we consider the case of adverse events where there is no difference between the control and treatment arms, we may expect that the chains in the simulation will contain mostly zeros. This means that only a small number of the samples actually contribute to the variability within the chains, and that differences between chains as measured by the GR statistic will thus be dependent on a small number of values, the number of chains and their lengths. Consequently, we should be careful about over-interpreting the GR statistic for the θ parameters for point-mass models. When dealing with a single data set a graphical assessment is a possible alternative, but for the large-scale simulations in this study where this is not an option we have chosen the GR statistic as a guide while acknowledging the above considerations. Another potential issue which may arise when fitting the θ parameters in the point-mass model is that it can be difficult to distinguish when the simulation is remaining at zero due to no differences between treatment and control, or if the model fit is not functioning correctly. For individual simulations we can investigate the fitting process by varying the simulation parameters and the sampling initial values to see what effect this has on the fit.

The general approach taken in Chapters 5 and 7 is to estimate a burn-in period and total number of iterations for the simulations, based on a number of preliminary runs, and then use this for all simulations. The inital values for the Markov chain may be highly dispersed and the chain may start and remain in a low probability region for a number of iterations. The burn-in period is used to allow the Markov chain to enter a higher probability region where the samples drawn should provide better approximations of the underlying distributions. If the GR statistics reported were less than 1.2 we considered the simulation to have reached approximate convergence. Simulations with larger GR statistics had their parameters tuned and were re-run until all the GR statistics were less than 1.2. The default numbers of chains, burn-in period, and total iterations in each chain are given in Table A.3.

C.2 Tuning Model Fitting Parameters

For MH sampling, using a normal proposal distribution centred on the current value, the variance (σ^2_{MH}) of the proposal must be supplied and tuned. For slice sampling a width parameter (w) is required and also, if chosen, a control parameter (m) [124]. For the simulation study in Chapter 5 a single parameter was generally suitable for all distributions of a particular type. The global values used in the simulations and demonstration analyses are given in Tables A.4 and A.5.

Generally, when using MH sampling, acceptance rates should be monitored with target acceptance rates of between 25% and 50% considered as optimal, depending on the dimensionality of the problem, for distributions absolutely continuous with respect to the Lebesgue measure [154]. For the MH steps as we have implemented them, the variance (σ^2_{MH}) controls how the next candidate is sampled. We may consider how this affects the acceptance rate in non-point mass models for the parameter θ. If there is no difference between treatment and control then we expect that θ should be close to 0. If the variance (σ^2_{MH}) is too large then many candidates will not be close to 0 and may be rejected. Conversely, if there is a difference between treatment and control, and σ^2_{MH} is too small, then the proposal may not adequately explore the domain of the target distribution.

For the case of point-mass models where the θ distributions are not absolutely continuous with respect to the Lebesgue measure, target acceptance rates are more difficult to assess. When considering proposal distributions for MH sampling there is the possibility of changing the weightings given to the mixture of the point-mass and normal distribution which, by default, is 0.5. With the default weightings, in the case where θ is actually non-zero approximately half of the proposed values will be zero and these may be rejected, especially if the difference between treatment and control is large, driving down the overall acceptance rate, with a similar problem when θ is actually zero but approximately 50% of the proposed values are non-zero. Further, we can consider that when there is no difference between treatment and control the possibility exists that most of the proposed zeros will be accepted. In this case the only way to control the acceptance rates for proposed zeros is by changing the weightings of the proposal distributions. To counter these sorts of issues the implementation allows both the MH variance (σ^2_{MH}), and the proposal point-mass weightings, to be overridden on a parameter by parameter basis.

C.3 Simulation Study

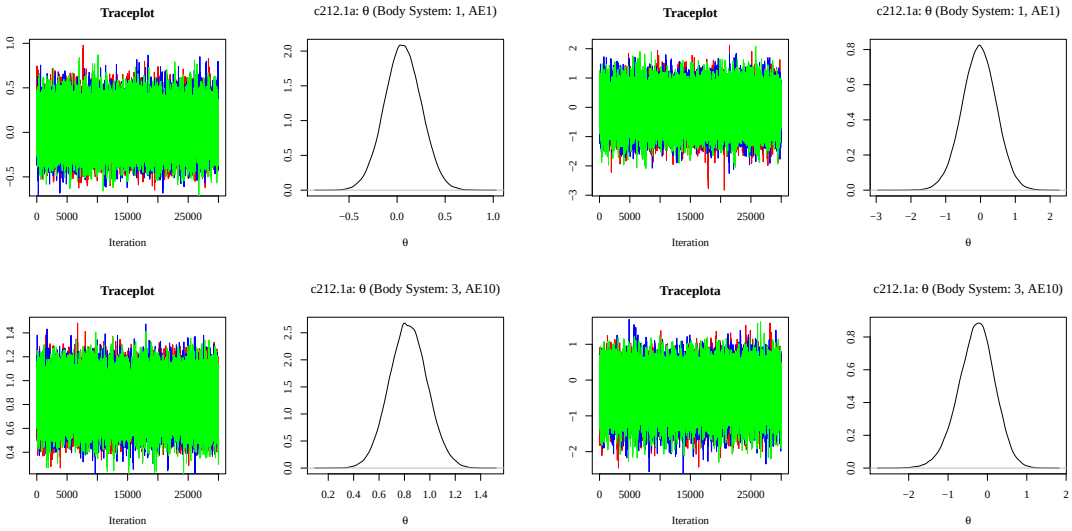
For the simulation study in Chapter 5 model c212.1a used three parallel chains of 40000 iterations, with the first 10000 iterations considered to be burn-in. Model c212.BB also used three parallel chains, but in this case of 60000 iterations with a 20000 burn-in period. The samples were not thinned [155]. The global defaults were generally suitable for all distributions of a particular type, particularly for the c212.1a models, but these global values were occasionally overridden by individual parameter values to improve model fit.

C.3.1 Individual Simulations

For the individual simulations described in §5.5.1, the Gelman-Rubin (GR) statistics from the models overall were consistent with MCMC convergence. For the c212.1a model with slice sampling just one simulation had a GR statistic greater than 1.2, with the largest value being 1.276755. For the model with MH sampling just two simulations had GR statistics exceeding 1.2, with the maximum GR statistics in these cases being 1.242543 and 1.273888. For the more complicated c212.BB model the GR statistic exceeded 1.2 in a maximum of 42 of the simulations (out of a possible 6000), with the maximum value being 1.294176. The vast majority of GR statistics in all simulations were less than 1.1.

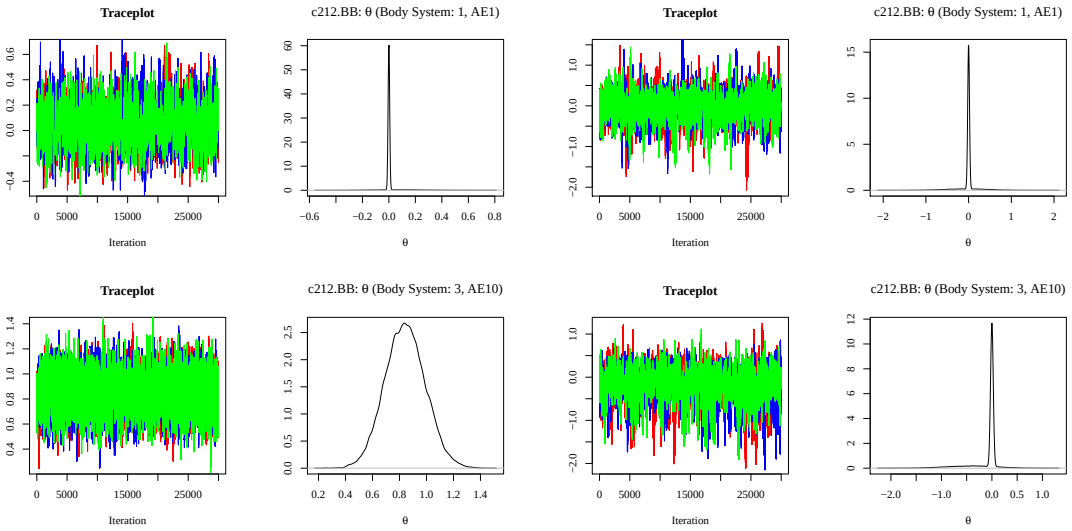
Figures C.1, C.2 show the posterior distributions and traceplots for the single adverse event in body system 1 (AE1), and the fifth adverse event in body system 3 (AE10), for a number of different simulations. In Figure C.1a, a large trial with a high increase in treatment rate, we can see that for c212.1a the posterior for θ for AE1 remains roughly centred at zero, whereas for AE10 it is almost entirely greater than zero. For the equivalent c212.BB plots (Figure C.2a) the posterior distribution for AE1 is effectively a point-mass at zero, whereas the posterior for AE10 is very similar to that for c212.1a. For the smaller trial with low increase in treatment rate for c212.1a (Figure C.1b) the posterior for AE1 is centred about 0 and for AE10 it looks to be centred about a negative value, the equivalent c212.BB posteriors (Figure C.2b) are effectively point-masses. This is not too surprising, for the smaller trial with low increased treatment event rate the models have trouble determining significant events. The plot for AE10 in Figure C.2a is not as smooth at that in Figure C.1a. We know from the parameters for the simulation that there is a large increase in the adverse event rate under treatment. Using a Metropolis-Hastings step with a proposal which gives equal weightings to both sides of the

mixture will, in this case, lead to many of the proposed zero values being rejected, and this contributes to lack of smoothness in the plot. A solution to this would be a re-weighting of the proposal distribution.



(a) Large Trial - High Rate. **(b)** Small Trial - Low Rate.

Figure C.1. c212.1a: Traceplots and posterior distributions for θ.



(a) Large Trial - High Rate. **(b)** Small Trial - Low Rate.

Figure C.2. c212.BB: Traceplots and posterior distributions for θ.

C.3.2 All Simulations

Overall, for all the simulations considered in Chapter 5, the Gelman-Rubin statistics were largely consistent with convergence. For model c212.1a (MH) 119 simulations had GR statistics which exceeded 1.2, with the maximum value being 1.300655. For c212.1a (SLICE) 109 simulations had statistics which exceeded 1.2, the largest being 1.276755. For c212.BB 2109 simulations had statistics which exceeded 1.2, with the maximum value being 1.746533. To put this in perspective, there are 5121000 sets of parameters families $(\theta_{bj}, \gamma_{bj}, ...)$ in the total c212.BB simulations, of which 2536 had a GR statistic greater than 1.2, covering 2109 separate simulations out of a total of 253500.

C.4 Lapatinib and Capecitabine versus Capecitabine in Women with Refractory Advanced or Metastatic Breast Cancer

In this section we look at the model fit for BB_{21} for the GSK trial results presented in §7.10. With the default parameters for BB_{21}, the largest reported GR statistic is 1.453476 for the θ parameter for the adverse event *Weight increased* in the *Investigations* body-system. The MH acceptance rates are between 15% and 68%. A traceplot for the data excluding the burn-in period is shown in Figure C.3.

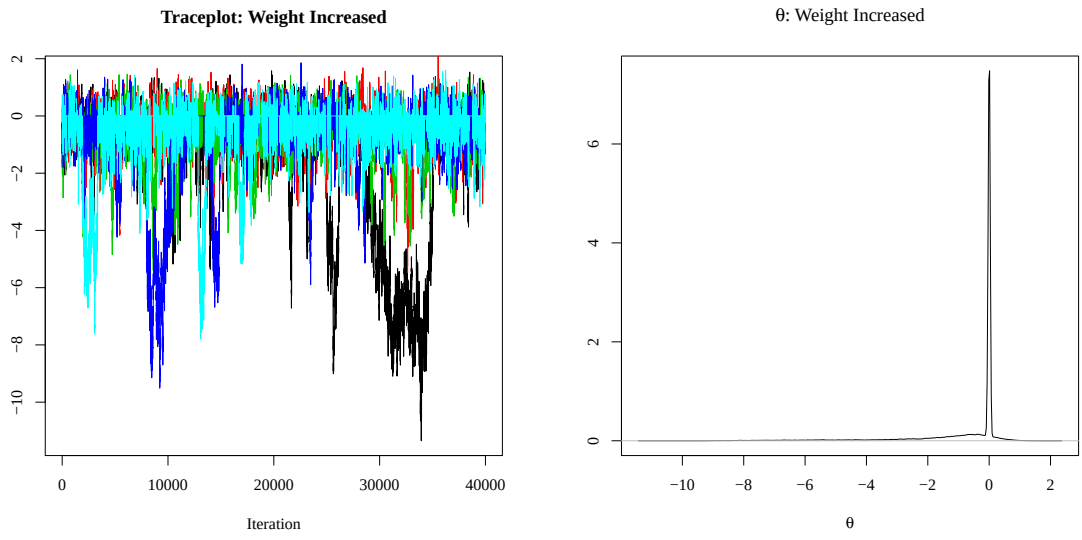


Figure C.3. Traceplot and posterior distribution for *Weight increased* with default simulation parameters.

We can see that while most of the simulated samples are gathered about 0, there are some quite significant departures for some of the chains, with some remaining in low probability regions for prolonged periods, and overall poor mixing. Even with this behaviour, the Top 10 adverse events by posterior probability are very nearly identical to those in Table 7.84, with only slight variation in the posterior probability values. Making a number of changes to the overall global defaults, and to a number of parameter specific values, and fitting the model again, gives an overall maximum GR statistic value 1.091287, and for *Weight increased* the GR statistic was 1.018458. The MH acceptance rates are now between 27% and 62%. The traceplot of the chains is shown in Figure C.4. With the updated simulation parameters the traceplot exhibits improved mixing and the chains do not remain in lower probability regions for as long as in the default parameter case. Further tuning is possible, but overall this traceplot may be considered to show consistency with approximate convergence. This fit of the model was used to generate the Top 10 adverse events reported in Table 7.84.

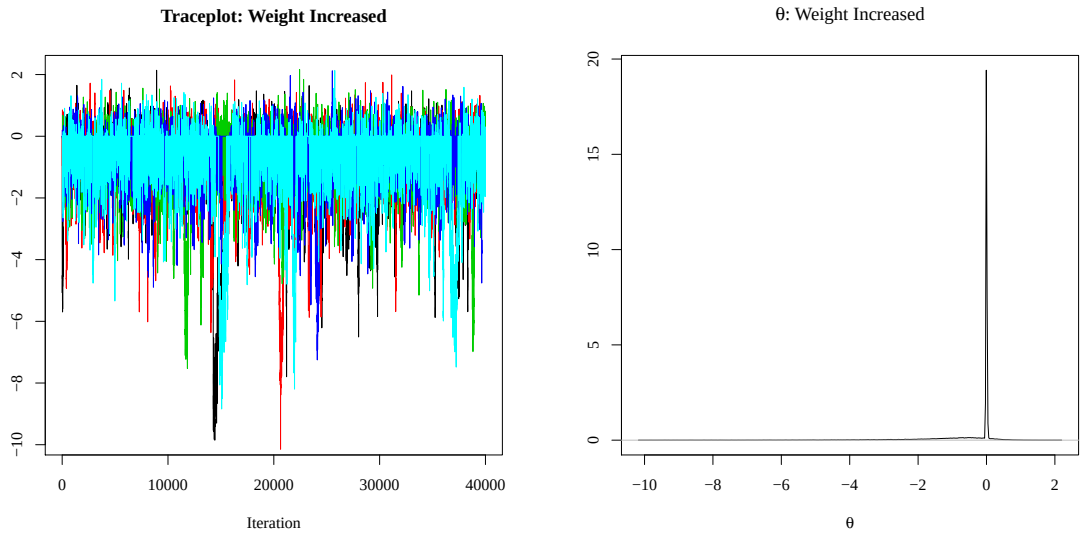


Figure C.4. Traceplot and posterior distribution plot for *Weight increased* with updated simulation parameters.

Increasing the number of chains from 5 to 10, and the number of iterations from 60000, with a burn-in of 20000, to 200000 with a burn-in of 40000, does not make any significant difference to the model outputs. For this case the traceplot for *Weight increased* is given in Figure C.5. The MH acceptance rates remain between 27% and 62%, and the Top 10 adverse events are almost identical to those in Table 7.84.

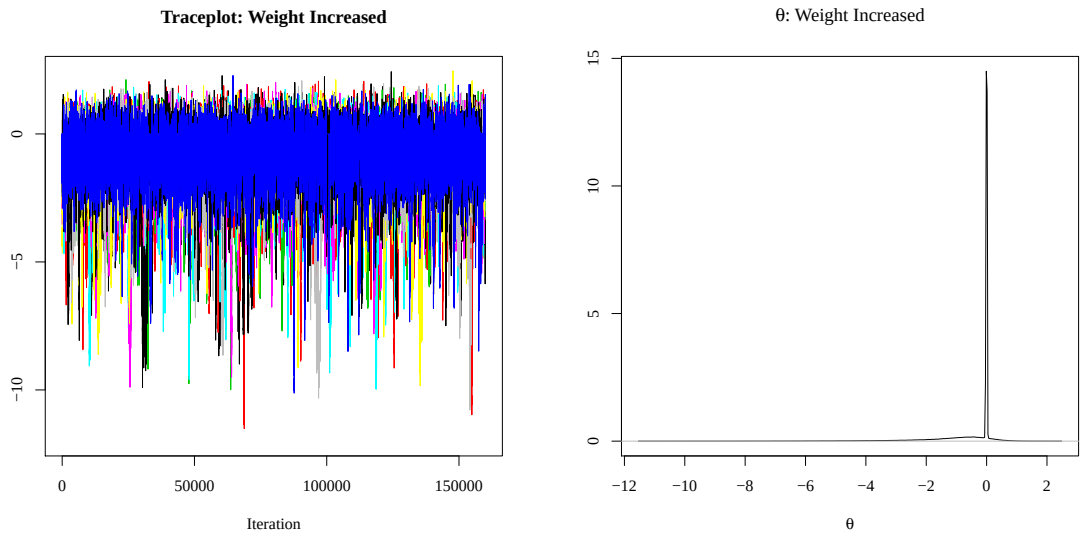


Figure C.5. Traceplot for *Weight increased* with 10 chains and increased iterations.

C.5 Demonstration Interim Analyses

For the 1a models in the demonstration interim analyses in Chapter 7, we ran three parallel chains of 40000 iterations, with the first 10000 iterations considered to be burn-in. The BB model fitting consisted of five parallel chains with 60000 iterations and a 20000 burn-in period. As in the simulation study (§C.3) the samples were not thinned [155].

The point-mass models always use Metropolis-Hastings sampling for the θ variables, and in this case it is also possible to monitor the MH acceptance rates for the sampler. For the purposes of this demonstration analysis we considered MH acceptance rates greater than 15% to be acceptable. Due to the much larger number of parameters in the interim analysis models, and their more complex relationships, it was expected that, at least for the BB_{hl} models, the default parameters in Table A.5 would have to be overridden.

Overall, for the 1a_{hl} models using the values from Table A.5 the largest Gelman-Rubin statistic was 1.116676. For the BB_{hl} model fitting a number of the global parameters in Table A.5 were overridden to enable better fit. In this case the largest Gelman-Rubin statistic was 1.199224. The acceptance rates for the MH sampler for the θ parameters varied between 15% and 75%, with most being between 20% and 60%. Overall these results are consistent with model convergence.

C.6 Sensitivity Analysis

C.6.1 Low Background Event Rate

For the 1a models a number of the defaults in Table A.5 had to be overridden. The largest Gelman-Rubin statistic was 1.195431, with the vast majority being under 1.1. For the BB_{hl} models with the default parameters most of the Gelman-Rubin statistics were under 1.2, but a small number did not converge using the default parameters values from Table A.5. This was particularly evident early in the trials when few events had occurred. This required that adjustments be made to the simulation parameters. With the overridden parameter values the largest Gelman-Rubin statistic was 1.199963, with MH acceptance rates ranging from 17% to 70%.

For example, with the default parameters, a model fit from the low background event rate analysis (§7.8.3.1) for model BB_{21} produced a maximum GR statistic 1.569185 for the θ model parameter for *Adv_131* in the interval *1260.0-1440.0*. The traceplot is shown in Figure C.6 where we can see that one of the chains remains in a low-probability region for a prolonged period.

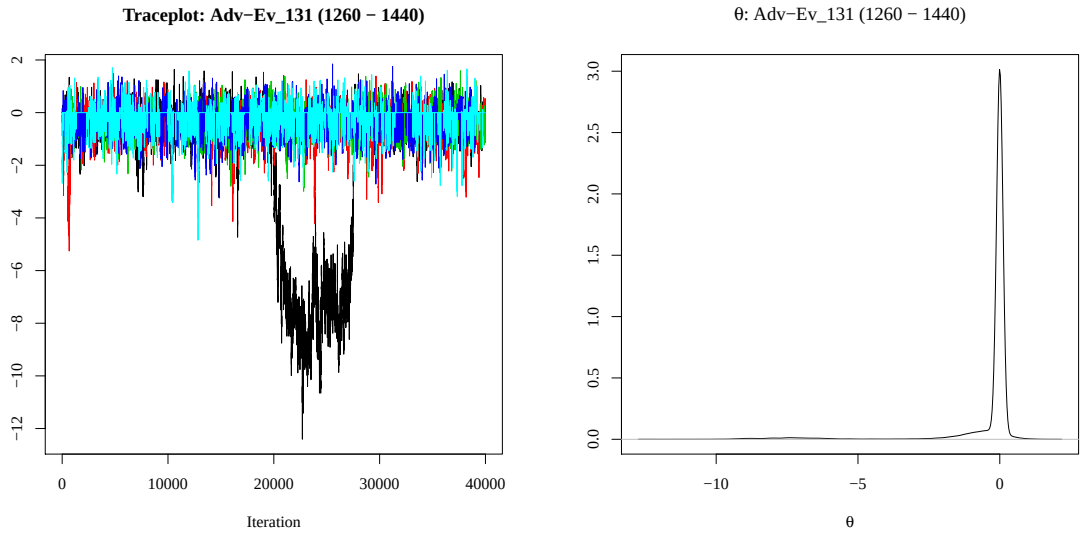


Figure C.6. Traceplot for *Adv_131* in interval 1260 - 1440 with default simulation parameters.

After tuning, the GR statistic for this parameter reduced to 1.074047, and the overall mixing improved. The traceplots are shown in Figure C.7. A number of the

chains remain in low probability regions for small numbers of iterations and while it is likely that further parameter tuning could yield quicker convergence, overall this traceplot is acceptable. Running the simulation with additional numbers of iterations (100000 in total, 20000 burn-in) (Figure C.8) yields a smaller GR statistic for the parameter (1.027956), but still gives a very similar traceplot with no chains remaining in low probability regions for long numbers of iterations.

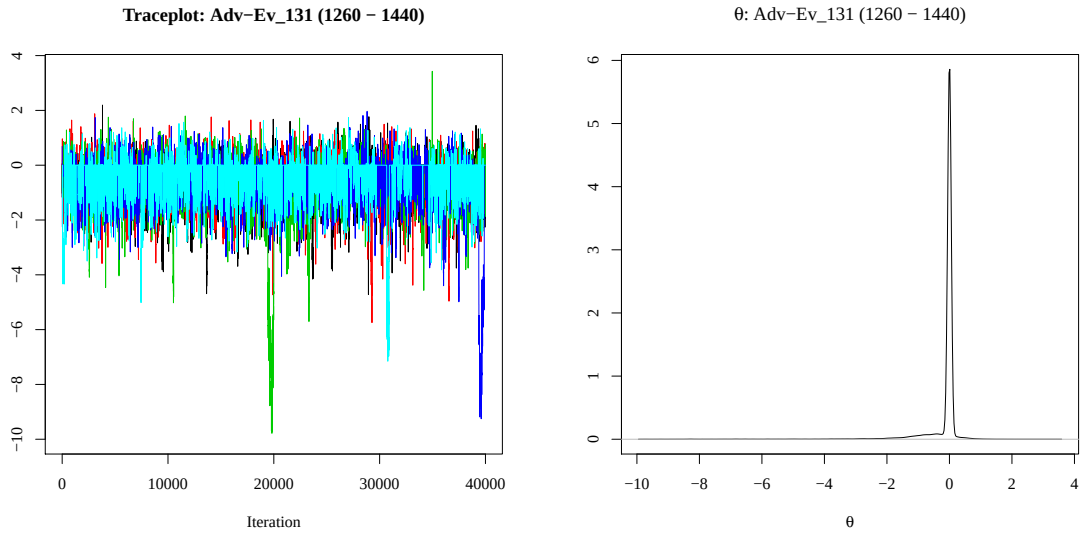


Figure C.7. Traceplot for *Adv_131* in interval 1260 - 1440 with tuned simulation parameters.

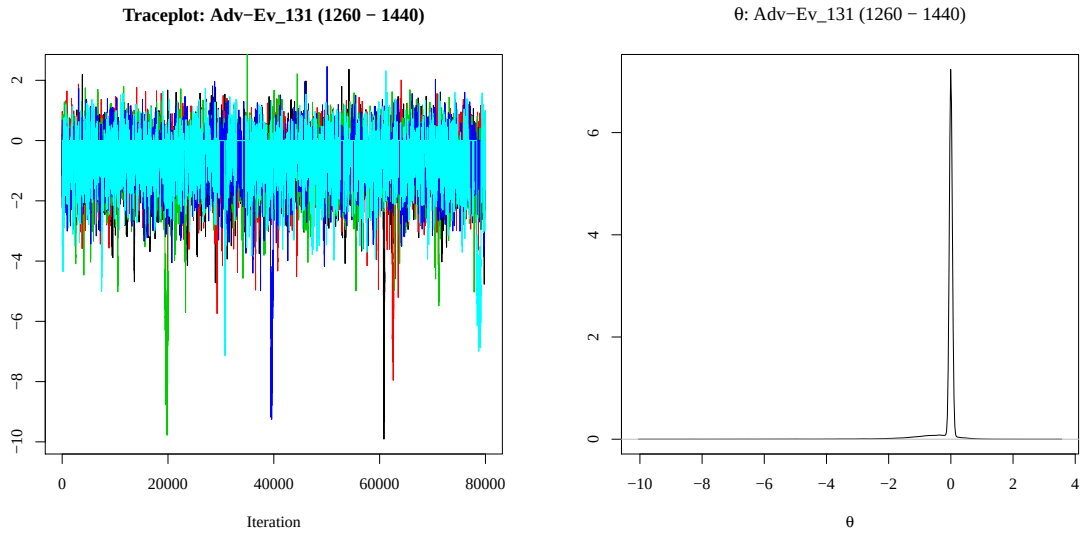


Figure C.8. Traceplot for *Adv_131* in interval 1260 - 1440 with tuned simulation parameters and additional iterations.

C.6.2 Mixed Background Event Rates

For the 1a models, as in the low frequency case, a number of the defaults in Table A.5 had to be overridden. The largest Gelman-Rubin statistic was 1.190395, with the vast majority being under 1.1. For the BB_{hl} models with the default parameters a number of the simulations did not converge using the default parameter values from Table A.5. Once adjustments were made to the simulation parameters the largest Gelman-Rubin statistic was 1.199455, with MH acceptance rates ranging from 15% to 75%.

C.7 Summary

The tuning of simulation parameters may be necessary to ensure approximate convergence is achieved in a reasonable period of time, particularly when performing large numbers of simulations. For the c212.1a, c212.BB, and $1a_{hl}$ models this is generally a straightforward procedure, with most of the simulations achieving convergence using the default global values, or common sets of values for particular types of simulations. For the BB_{hl} models this is less straightforward due to the existence of the point mass, the larger number of parameters, and their interdependent relationships. Unfortunately there is no straightforward globally applicable procedure to easily determine a set of simulation parameters suitable for any particular model and data set.

Appendix D

Grouped FDR Controlling Methods

D.1 Controlling Error Rates for DFDR

Any 2-STEP FDR controlling procedure as described in §2.3.2.5 controls the FDR at level α. This follows trivially by conditioning on the included families. Referring to Table 2.1, if PR is such a procedure used to select a set of families F, then assuming the conditions in [31] are met we have:

$$\mathrm{FDR}_{\mathrm{PR}} \left| F = E\left[FDP \middle| F \right] = E\left[\frac{V}{R} \middle| F \right] \leq \frac{m_0^F}{|F|}\alpha \leq \alpha \right.$$

where m_0^F is the number of true hypotheses in F and by taking expectations we have:

$$\mathrm{FDR}_{\mathrm{PR}} \leq \alpha$$

Unlike applying the BH-procedure, where we have control at level $\frac{m_0}{m}\alpha$, the actual level of control of the FDR is dependent on the distribution of the random set F. We could consider an approach to increasing the power by estimating m_0^F and $|F|$ which may allow control of the conditional FDR at level α as opposed to $\frac{m_0^F}{|F|}\alpha$. Although this is not investigated in this study, it is not dissimilar to the GBH approach to weighting p-values as discussed in §2.6, and to the approach in [43].

D.1.1 Large Body-System Properties for the DFDR Under Independence Assumptions

While the DFDR controls the FDR at both the group and overall level, there are currently no asymptotic results for the DFDR method with regard to power, but we can give some indication that the behaviour becomes stable even as the sizes of the groups increases.

One key element of the DFDR compared to the BH-procedure is the desire to exclude groups containing only true null hypotheses from the final set of hypotheses, F. This is controlled by the body-system representative p-value. Mehrotra and Adewale considered a number of alternative possibilities for the choice of body-system representative p-value [3] for a group g of hypotheses of size k_g, including $k_g p_{g(1)}$, which is the original DFDR representative p-value, weighted by the body-system size.

If we let g be a group of size $k_g > 1$, where all the null hypotheses are true, and let g_p be the corresponding p-values. Then if $p_{gi} \in g_p$ we have $p_{gi} \sim U(0,1)$, at least approximately (and exactly for continuous test statistics).

Assuming (approximate) independence of the p-values in group g, the l^{th} order statistic is Beta$(l, k_g + 1 - l)$ with

$$\mathrm{E}\left[\frac{k_g}{l} p_{g(1)}\right] = \frac{k_g}{l}\frac{l}{k_g+1} = \frac{k_g}{k_g+1}$$

The body-system representative p-value for the DFDR may be defined as:

$$P_g^* = \min\left(k_g p_{g(1)}, \frac{k_g}{2} p_{g(2)}, \ldots, p_{g(k_g)}\right) \tag{D.1}$$

and we can investigate the behaviour of P_g^*, as the body-system size increases, with a small simulation study.

The results in Table D.1 gives the estimated expected value of P_g^* as the body-size increases where, at each body-system size, the set of random variables

$$k_g p_{g(1)}, \frac{k_g}{2} p_{g(2)}, \ldots, p_{g(k_g)}$$

was sampled 20,000 times, and the minimum of each set of samples recorded. The estimated expected value of P_g^* is the mean of the recorded minimum values.

The results indicate that the expectation of the representative p-value does not decay to 0 as it does in the original DFDR, or approach 1, as it would do theoretically if $k_g p_{g(1)}$ was used as the representative p-value, but remains stable as the body-system size increases. This agrees with Mehrotra and Adewale's assertion that their chosen p-value is more powerful than $k_g p_{g(1)}$ in the sense that it may be less restrictive.

Body-System Size (k_g)	Estimated Expected Value of P_g^*
10	0.206655
100	0.237527
1000	0.250395
10000	0.255120
100000	0.258390

Table D.1. DFDR representative p-value as body-system size increases for body-systems containing only true null hypotheses.

This indicates that the probability of including a set which contains only true null hypotheses is controlled, at least in the case of independent or approximately independent test statistics, and that this behaviour is maintained as the sizes of the groups increase, leading to a potential increase in power over the standard BH-procedure. We will see this is the case in the simulation below (Table D.9).

D.2 Comparison of DFDR and GBH using Simulated Data

As discussed in §2.6 the Double False Discovery Rate (DFDR) and Group Benjamini-Hochberg (GBH) methods have some similar characteristics. The main purpose of this simulation is to examine the DFDR and GBH particularly with regard to the discussion in §2.6, and also to investigate the asymptotic properties of the DFDR (§D.1.1). In order to do this we look at the relative performance of these methods, the Benjamini-Hochberg (BH) procedure, and hypothesis testing unadjusted for multiplicities (NOADJ), when different proportions of adverse events have raised treatment rates within a body-system, and when the number of adverse events in each body-system becomes large but the proportion of adverse events with raised rates remains the same.

We are interested in the number of events correctly identified by the methods as having raised treatment rates, the number of type-I errors, and also the power of the DFDR as the body-system size increases.

D.2.1 Simulation Definition

We consider a Medium size trial (Table 5.3) with B body-systems and raised adverse events in one body-system, k, only. The proportion of events with raised

rates in k is q. We consider four body-system sizes: 10, 20, 50, and 100. The background probability of an adverse event occurring is p_1, the probability of a raised treatment rate adverse event occurring is p_2. The details of the simulations and parameter values are given in Table D.2.

Simulation Name	B	k	q	**Body-system Size**	p_1	p_2
LBS0	10	3	0	10, 20, 50, 100	0.047	0.047
LBS1	10	3	0.1	10, 20, 50, 100	0.047	0.052, 0.076, 0.12
LBS2	10	3	0.5	10, 20, 50, 100	0.047	0.052, 0.076, 0.12
LBS3	10	3	0.9	10, 20, 50, 100	0.047	0.052, 0.076, 0.12
LBS4	10	3	1	10, 20, 50, 100	0.047	0.052, 0.076, 0.12

Table D.2. Large body-system simulation parameter values.

Each simulation is repeated 500 times.

D.2.1.1 Adverse Events with Raised Treatment Rates

The total number of adverse events with raised treatment rates is given in the Table D.3. For example, for LBS2 we have 10 body-systems each containing $10, 20, 50,$ or 100 adverse events. There are 3 choices for p_2 and each simulation is repeated 500 times giving:

$$10 \times (10 + 20 + 50 + 100) \times 3 \times 500 = 2700000$$

adverse events in total. A proportion of 0.5 adverse events in body-system 3 have raised treatment rates:

$$(10 + 20 + 50 + 100) \times 0.5 \times 3 \times 500 = 135000$$

Simulation Name	Raised Rates[1]	Total Events[2]
LBS0	0	900000
LBS1	27000	2700000
LBS2	135000	2700000
LBS3	243000	2700000
LBS4	270000	2700000
Total	675000	11700000

Table D.3. Adverse events totals - all simulations

[1] Total number of adverse events with raised treatment rates in the simulation.
[2] Total number of adverse events in the simulation.

D.2.2 Results Summary

The overall results of the simulation are given in Table D.4. We used a significance level of 5% for BH and NOADJ, and 5% and 10% for DFDR and GBH. We can see that overall both the DFDR and GBH have performed better than the BH-procedure and the GBH is also the most powerful method, in agreement with [3], although because the simulations include tests where the known differences between control and treatment is small, overall the power is low. The Type-I error rate is inflated for the GBH compared to the other methods, at the 5% level it is more than 4 times that of the equivalent DFDR method, although still well below the nominal significance level.

Looking at the individual simulations we can see that in terms of correctly identifying significant adverse events that the GBH outperforms the DFDR at the same level in each simulation, but at the cost of a higher Type-I error rate. At the higher 10% significance level, recommended in [3], the Type-I error rate for GBH can be as high as 7 times that of DFDR (Table D.7).

Comparing LBS3 and LBS4, where we go from a proportion of 0.9 to 1.0 of significant events in body-system 3, we can see that there is both a reduction in overall total number of Type-I errors and the Type-I error rate for GBH as anticipated in §2.6. There is also a reduction in Type-I error for DFDR but, as might be expected, an increase for the BH-procedure, which does not have a group effect. The GBH performs best when we can group likely significant events into groups such

as by a cluster analysis in the gene expression experiment described in [30].

Method	Correct[1]	Type-I[2]	Type-II[3]	Raised[4] Rates	Total[5] Events
BH	192519(28.52%)	6262(0.06%)	482481(71.48%)	675000	11700000
DFDR(5%)	235685(34.92%)	2636(0.02%)	439315(65.08%)	675000	11700000
DFDR(10%)	262675(38.91%)	6317(0.06%)	412325(61.09%)	675000	11700000
GBH(5%)	262547(38.90%)	12128(0.11%)	412453(61.10%)	675000	11700000
GBH(10%)	301698(44.70%)	27673(0.25%)	373302(55.30%)	675000	11700000
NOADJ	315955(46.81%)	384998(3.49%)	359045(53.19%)	675000	11700000

Table D.4. Overall results.

[1] The total number of adverse events with raised rates that were correctly identified by the method as having a raised rate.
[2] The total number of adverse events without raised rates that were (incorrectly) identified by the method as having a raised rate.
[3] The total number of adverse events with raised rates that were not identified by the model as having a raised rate.
[4] Total from Table D.3.
[5] Total from Table D.3.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
BH	5908 (21.88%)	359(0.01%)	21092(78.12%)	27000	2700000
DFDR(5%)	7757(28.73%)	525(0.02%)	19243(71.27%)	27000	2700000
DFDR(10%)	8342(30.90%)	1203(0.05%)	18658(69.10%)	27000	2700000
GBH(5%)	8349(30.92%)	1996(0.07%)	18651(69.08%)	27000	2700000
GBH(10%)	9128(33.81%)	4383(0.16%)	17872(66.19%)	27000	2700000
NOADJ	12652(46.86%)	93255(3.49%)	14348(53.14%)	27000	2700000

Table D.5. LBS1 Results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
BH	36758(27.23%)	1300(0.05%)	98242(72.77%)	135000	2700000
DFDR(5%)	44808(33.19%)	1094(0.04%)	90192(66.81%)	135000	2700000
DFDR(10%)	48407(35.86%)	2538(0.10%)	86593(64.14%)	135000	2700000
GBH(5%)	48922(36.24%)	3412(0.13%)	86078(63.76%)	135000	2700000
GBH (10%)	54051(40.04%)	7604(0.30%)	80949(59.96%)	135000	2700000
NOADJ	63336 (46.92%)	89335(3.48%)	71664(53.08%)	135000	2700000

Table D.6. LBS2 Results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
BH	70481(29.00%)	2139(0.09%)	172519(71.00%)	243000	2700000
DFDR(5%)	86014(35.40%)	639(0.03%)	156986(64.60%)	243000	2700000
DFDR(10%)	96431(39.68%)	1570(0.06%)	146569(60.32%)	243000	2700000
GBH(5%)	96432(39.68%)	4549(0.19%)	146568(60.32%)	243000	2700000
GBH(10%)	111386(45.84%)	11172(0.45%)	131614(54.16%)	243000	2700000
NOADJ	113839(46.85%)	85843(3.49%)	129161(53.15%)	243000	2700000

Table D.7. LBS3 Results.

Method	Correct	Type-I	Type-II	Raised Rates	Total Events
BH	79372(29.40%)	2388(0.10%)	190628(70.60%)	270000	2700000
DFDR(5%)	97106(35.97%)	300(0.01%)	172894(64.03%)	270000	2700000
DFDR(10%)	109495(40.55%)	846(0.03%)	160505(59.45%)	270000	2700000
GBH(5%)	108844(40.31%)	1562(0.06%)	161156(59.69%)	270000	2700000
GBH(10%)	127133(47.09%)	3292(0.14%)	142867(52.91%)	270000	2700000
NOADJ	126128(46.71%)	84915(3.49%)	143872(53.29%)	270000	2700000

Table D.8. LBS4 Results.

From §D.1.1 we expect that as the number of adverse events in each body-system increases the DFDR will maintain its power. In Table D.9 we can see that the estimated power for the DFDR at the 5% level does remain relatively constant as

the body-system sizes increase confirming our expectations.

Body-System Size	Correct	Type-II	Raised Rates[1]	Estimated Power
10	13104	24396	37500	0.349
20	26126	48874	75000	0.348
50	65386	122114	187500	0.349
100	131069	243931	375000	0.350

Table D.9. DFDR power as body-system size increases.

[1] For each body-system size, n, the number of events with raised treatment rates is given by: $n\times$ 3 $\times$ 500 $\times$ (0 + 0.1 + 0.5 + 0.9 + 1). There are 3 different values of p_2 (Table D.2), 500 repeated simulations, and 0, 0.1, 0.5, 0.9 and 1 are the proportions of events with raised treatment rates in body-system 3.

D.2.3 Conclusions

As with the main simulation study in Chapter 5 we can see that for this data both DFDR and GBH control the Type-I error rate very tightly, with DFDR providing better overall control, and GBH being the more powerful of the two methods, with GBH using a 5% significance level comparable to DFDR at the 10% level. The DFDR maintains its power as the body-system sizes increase (Table D.9), which we expect from D.1.1.

Appendix E

Trial Adverse Event Simulation

The generation of the simulated trial data used in Chapter 7 is described in this appendix. There are two parts to the simulation, patient recruitment, and event generation.

E.1 Simulating Patient Recruitment

Patients are recruited to each arm of the trial according to a Poisson process up to a maximum number of patients. The recruitment rates used in Chapter 7 are given in Table 7.1.

E.2 Simulating Adverse Event Data

Apart from the data in §7.8.4, the adverse events are simulated by choosing an underlying overall adverse event rate for the trial (Tables 7.6, 7.68), with the adverse event rate in each particular body-systems being a random sample from a normal distribution whose mean is this overall rate. The standard deviation of the normal distribution used in the simulation is given in Table E.1. The data in §7.8.4 used a combination of normal centred samples from the rates in Tables 7.6 and 7.68 as the adverse event background rates, with the adverse events with higher rates given in Table 7.75.

Parameter	Value	Description
σ	0.0001	The standard deviation of the normal distribution used when sampling from the adverse event background rates.

Table E.1. Interim analysis trial simulation parameters.

As part of the simulation we also specify which adverse events on the treatment

arm will have increased rates compared to the control, the size of the increase, and the intervals over which the rates are increased (Tables 7.5, 7.7, 7.70, 7.77).

Once the adverse event rates have been established, the simulated event data is generated by a marked or compound inhomogeneous Poisson process. Each individual adverse event generated by the simulation process is independently assigned a severity from 1-5, corresponding to the NCI CTCAE severities (Table 1.1), based on a predefined probability of occurrence (Table 7.4). If $\lambda(t)$ is the intensity or rate function of the Poisson process we let:

$$\bar{\lambda} = \max_{t \in [0,1800]} \lambda(t)$$

and proceed as follows:

1. For each subject in the trial generate events according to a Poisson process with rate: $\bar{\lambda}$.
2. If an event is generated at time T^*, generate $u \sim U(0,1)$. Accept the event as belonging to the process if $\frac{\lambda(T^*)}{\bar{\lambda}} < u$.
3. Assign a severity to the event with the probabilities from Table 7.4.
4. If an event has severity 5 no further events are generated for that subject for any adverse event.

Where the rates are constant within the intervals, as it is in Chapter 7, $\lambda(t)$ is a step-function.

Appendix F

Table of Methods

A number of the methods reviewed in Chapter 2 and 3, representing the different modelling or error controlling approaches to safety data, are summarised in the Table F.1. The headings in the table are as follows:

1. Reference: The authors and references to the paper originating the methods.
2. Data: The type of data the method uses, e.g. adverse event count data.
3. Model/Method Type: The method may be a model (parametric, non-parametric, etc.) or a statistical procedure.
4. Error Control: The method may control error rates.
5. Body-system: The method may be suitable for use with body-systems.
6. Subgroupings: The method may allow the data (population) to be divided into subgroupings, for example by covariates.
7. MCP (Multiple comparison procedure): The method may control for multiple comparisons.
8. Censored Data: The method may be suitable for analysing censored data.

Reference	Data	Model/Method Type(s)	Error Control	Body-System	Sub-groupings	MCP	Censored Data
Siddiqui (2009) [4]	Event counts and timings.	Non-parametric (MCF)	No	Yes	Yes	No	Random
Cook et al. (1997) [92]	Event counts and timings	Parametric and semi-Parametric.	No	Yes	Yes	No	Dependent (terminal events)
O'Neill (1995)[77]	Marginal adverse event counts	Parametric	No	No	Yes	No	Yes
Rosenkranz (2006)[94]	Event counts and timings	Parametric, semi-parametric	No	No	No	No	Dependent (terminal time)
Wang et al. (2001) [93]	Event counts and timings	Parametric, semi-parametric, latent variable, multiplicative intensity model.	No	No	Yes	No.	Dependent (terminal time)
Wang et al. (2012) [104]	Event counts and timings	Non-parametric (MCD). Extends [102]	No	Yes	Yes	No	Random and dependent
Wang et al. (2013) [105]	Event counts and timings	Semi-parametric approach to[104]	No	Yes	Yes	No	Random and dependent
Zhao, Zhou (2012) [106]	Event counts and timings	Parametric, additive intensity model.	No	No	Yes	No	Dependent
Cook et al. (2009) [98]	Event counts and timings	Parametric	No	No	No	No	Dependent
Mehrotra, Adewale [3]	-	Error controlling procedure	Yes (FDR)	Yes	N/A	Yes	N/A

Benjamini, Hochberg [31]	-	Error controlling procedure	Yes (FDR)	N/A	N/A	Yes	N/A
Benjamini, Liu [42]	-	Error controlling procedure	Yes (FDR)	N/A	N/A	Yes	N/A
Benjamini, Yekutieli [32]	-	Error controlling procedure	Yes (FDR)	N/A	N/A	Yes	N/A
Benjamini, Krieger, Yekutieli [43]	-	Error controlling procedure	Yes (FDR)	No	N/A	Yes	N/A
Yekutieli [57]	-	Error controlling procedure	Yes (FDR)	Yes	N/A	Yes	N/A
Hu, Zhao,Zhao [30]	-	Error controlling procedure	Yes (FDR)	Yes	N/A	Yes	N/A
Storey, Taylor, Siegmund [48]	-	Error controlling procedure	Yes (FDR)	No	No	Yes	N/A
Genovese, Roeder, Wasserman [56]	-	Error controlling procedure	Yes (FDR)	No	No	Yes	N/A
Muller et al. [54]	-	Error controlling procedure (Bayesian)	Yes (FDR)	No	No.	Yes	N/A
Berry, Berry (2004) [5]	Marginal adverse event counts	Bayesian (parametric)	No	Yes	No	Yes (priors)	N/A

Xia et al [60]	Marginal adverse event counts, Timings of adverse event occurrence	Bayesian (parametric)	No	Yes	No	Yes (priors)	N/A
Agresti, Klingenberg [131]	Count data	Parametric	No	Yes	No	Yes	N/A
Chuang-Stein, Mohberg, Musselman [29]	Adverse event incidence	Non-parametric	No	Yes	No	Yes	N/A
Chen et al. [125]	Marginal adverse event counts, interim counts	Bayesian (parametric), sequential update of the posteriors.	No	Yes	No	Yes (priors)	N/A
Crooks et al. ([130])	Adverse event counts	Parametric, Bayesian (parametric)	No	Yes	No	Yes (priors)	N/A
Goldberg-Alberts, Page [132]	Adverse event counts	parametric, log-linear model.	No	Yes	No	Yes	N/A
Gould [133]	Adverse event counts	Bayesian (parametric).	No	Yes	No	Yes (priors)	N/A
Kim et al. [134]	Event counts and timings	Parametric, log-linear model	No	Yes	Yes	Yes	N/A
Rosenkranz ([135])	Event counts and timings	Bayesian(parametric), loglinear model	No	Yes	Yes	Yes (prior)	N/A
Schildcrout et al. (2008)[119]	Event counts and timings	Non-parametric, time dependent linear model	No	No	Yes	No	Yes

Simo [136]	Event counts	Bayesian (parametric), regression model	No	Yes	No	Yes (priors)	N/A
South-worth, O'Connell [138]	Event counts	Parametric, machine-learning, Bayesian	No	Yes	Yes	Yes	N/A
DuMouchel[6]	Marginal adverse event counts only	Bayesian (parametric)	No	Single body-system	Yes	Yes (priors)	N/A
DuMouchel[9]	Adverse event counts	Bayesian (parametric)	No	No	Yes	Yes (prior)	N/A
Gould[8]	Adverse event counts	Bayesian (parametric)	No	No	No.	Yes (prior)	N/A
Bate et al(1998) [10]	Adverse event counts	Bayesian (parametric), neural network	No	No	No	Yes (prior)	N/A
Shaddox et al(2016) [129]	Adverse event counts	Bayesian (parametric)	No	No but grouping used	No	Yes (prior)	N/A

Table F.1. Table of methods.

References

[1] L.M. Friedman, C.D. Furberg, and D.L. DeMets. *Fundamentals of Clinical Trials*. Springer, 2010.

[2] D.L. DeMets, C.D. Furberg, and L.M. Friedman. *Data Monitoring in Clinical Trials: A Case Studies Approach*. Springer, 2006.

[3] D. V. Mehrotra and A. J. Adewale. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Statistics in Medicine*, 31(18):1918–30, 2012.

[4] Ohidul Siddiqui. Statistical methods to analyze adverse events data of randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 19(5):889–899, 2009.

[5] Scott M. Berry and Donald A. Berry. Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60(2):418–426, 2004.

[6] William DuMouchel. Multivariate bayesian logistic regression for analysis of clinical study safety issues. *Statistical Science*, 27(3):319–339, 2012.

[7] S. J. Pocock. *Clinical trials: a practical approach*. Wiley medical publication. Wiley, 1983.

[8] A. Lawrence Gould. Accounting for multiplicity in the evaluation of "signals" obtained by data mining from spontaneous report adverse event databases. *Biometrical Journal*, 49(1):151–165, 2007.

[9] William DuMouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999.

[10] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998.

[11] D. L. DeMets and K. K. Lan. Interim analysis: the alpha spending function approach. *Statistics in Medicine*, 13(13-14):1341–52; discussion 1353–6, 1994.

[12] John N. S. Matthews. *Introduction to Randomized Controlled Clinical Trials, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 2006. doi:10.1201/9781420011302.fmatt.

[13] Brenda J. Crowe, H. Amy Xia, Jesse A. Berlin, Douglas J. Watson, Hongliang Shi, Stephen L. Lin, Juergen Kuebler, Robert C. Schriver, Nancy C. Santanello, George Rochester, Jane B. Porter, Manfred Oster, Devan V. Mehrotra, Zhengqing Li, Eileen C. King, Ernest S. Harpur, and David B. Hall. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. *Clinical Trials*, 6(5):430–440, 2009.

[14] Q. Jiang and H. A. Xia. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, 2014.

[15] D. V. Mehrotra and J. F. Heyse. Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13(3):227–38, 2004.

[16] Funda Meric, Mien-Chie Hung, Gabriel N. Hortobagyi, and Kelly K. Hunt. Her2/neu in the management of invasive breast cancer. *Journal of the American College of Surgeons*, 194(4):488–501, 2002.

[17] Hanfang Jiang and Hope S. Rugo. Human epidermal growth factor receptor 2 positive (her2+) metastatic breast cancer: how the latest results are improving therapeutic options. *Therapeutic Advances in Medical Oncology*, 7(6):321–339, 2015.

[18] Andrew Seidman, Clifford Hudis, Mary Kathryn Pierri, Steven Shak, Virginia Paton, Mark Ashby, Maureen Murphy, Stanford J. Stewart, and Deborah Keefe. Cardiac dysfunction in the trastuzumab clinical trials experience. *Journal of Clinical Oncology*, 20(5):1215–1221, 2002.

[19] Timothy Kute, Christopher M. Lack, Mark Willingham, Bimjhana Bishwokama, Holly Williams, Kathy Barrett, Tanita Mitchell, and James P. Vaughn. Development of herceptin resistance in breast cancer cells. *Cytometry Part A*, 57A(2):86–93, 2004.

[20] Paula R. Pohlmann, Ingrid A. Mayer, and Ray Mernaugh. Resistance to trastuzumab in breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(24):7479–7491, 2009.

[21] David Cameron, Michelle Casey, Cristina Oliva, Beth Newstat, Bradley Imwalle, and Charles E. Geyer. Lapatinib plus capecitabine in women with her-2–positive advanced breast cancer: Final survival analysis of a phase iii randomized trial. *The Oncologist*, 15(9):924–934, 2010.

[22] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979.

[23] Arlene Chan. Lapatinib - overview and current role in metastatic breast cancer. *Cancer Research and Treatment : Official Journal of Korean Cancer Association*, 38(4):198–200, 2006.

[24] David Cameron, Michelle Casey, Michael Press, Deborah Lindquist, Tadeusz Pienkowski, C. Gilles Romieu, Stephen Chan, Agnieszka Jagiello-Gruszfeld, Bella Kaufman, John Crown, Arlene Chan, Mario Campone, Patrice Viens, Neville Davidson, Vera Gorbounova, Johannes Isaac Raats, Dimosthenis Skarlos, Beth Newstat, Debasish Roychowdhury, Paolo Paoletti, Cristina Oliva, Stephen Rubin, Steven Stein, and Charles E. Geyer. A phase iii randomized comparison of lapatinib plus capecitabine versus capecitabine alone in women with advanced breast cancer that has progressed on trastuzumab: updated efficacy and biomarker analyses. *Breast Cancer Research and Treatment*, 112(3):533–543, 2008.

[25] Xiaolei Zhou, David Cella, David Cameron, Mayur M. Amonkar, Anthony Segreti, Steven Stein, Mel Walker, and Charles E. Geyer. Lapatinib plus capecitabine versus capecitabine alone for her2+ (erbb2+) metastatic breast cancer: quality-of-life assessment. *Breast Cancer Research and Treatment*, 117(3):577–589, 2009.

[26] Charles E. Geyer, John Forster, Deborah Lindquist, Stephen Chan, C. Gilles Romieu, Tadeusz Pienkowski, Agnieszka Jagiello-Gruszfeld, John Crown, Arlene Chan, Bella Kaufman, Dimosthenis Skarlos, Mario Campone, Neville Davidson, Mark Berger, Cristina Oliva, Stephen D. Rubin, Steven Stein, and David Cameron. Lapatinib plus capecitabine for her2-positive advanced breast cancer. *New England Journal of Medicine*, 355(26):2733–2743, 2006.

[27] Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[28] B. Dean. Adverse drug events: what's the truth? *Quality & safety in health care*, 12(3):165–166, 2003.

[29] Christy Chuang-Stein, Noel R. Mohberg, and David M. Musselman. Organization and analysis of safety data using a multivariate approach. *Statistics in Medicine*, 11(8):1075–1089, 1992.

[30] J. X. Hu, H. Zhao, and H. H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010.

[31] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[32] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. Times Cited: 1314 Benjamini, Y Yekutieli, D.

[33] Thomas R. Fleming, Judith R. O'Fallon, Peter C. O'Brien, and David P. Harrington. Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625, 1980.

[34] N. Balakrishnan and C. R. Rao. *Handbook of Statistics: Advances in Survival Analysis*. Handbook of Statistics. Elsevier Science, 2004.

[35] G. F. Liu, J. Y. Wang, K. Liu, and D. B. Snavely. Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in Medicine*, 25(8):1275–1286, 2006.

[36] Zbynek Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[37] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[38] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[39] H. Scheffé. *The Analysis of Variance*. A Wiley publication in mathematical statistics. Wiley, 1999.

[40] R. D'Agostino, S. P. Ralph B. D'agostino, L. M. Sullivan, and J. Massaro. *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons Inc, 2008.

[41] Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[42] Yoav Benjamini and Wei Liu. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82(1–2):163–170, 1999.

[43] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[44] John D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[45] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[46] J. D. Storey and R. Tibshirani. Estimating false discovery rates under dependence, with applications to dna microarrays. *Unpublished*, 2001. http://genomics.princeton.edu/storeylab/papers/dep.pdf.

[47] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[48] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[49] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[50] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[51] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[52] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[53] C. Genovese and L. Wasserman. Bayesian and frequentist multiple testsing. In J. M. Bernardo, editor, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 145–161. Oxford University Press, 2003.

[54] Peter Muller, Giovanni Parmigiani, and Kenneth Rice. FDR and Bayesian Multiple Comparison Rules. In J.O. Berger A.a.P Dawid D. Heckerman A.F.M. Smith J.M. Bernardo, M.J. Bayarri and M. West, editors, *Bayesian Statistics 8*, pages 349–370. Oxford University Press, 2007.

[55] Luis G. León-Novelo, Peter Müller, Wahid Arap, Jessica Sun, Renata Pasqualini, and Kim-Anh Do. Bayesian decision theoretic multiple comparison procedures: An application to phage display data. *Biometrical Journal*, 55(3):478–89, 2013.

[56] Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[57] Daniel Yekutieli. False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference*, 138(2):405–415, 2008.

[58] International Conference on Harmonisation E9 Expert Working Group. Statistical principles for clinical trials. ICH Harmonised Tripartite Guideline. *Statistics in Medicine*, 18(15):1905–42, 1999.

[59] A. D. Lunn and S. J. Davies. A note on generating correlated binary variables. *Biometrika*, 85(2):487–490, 1998. Times Cited: 36 Lunn, AD Davies, SJ.

[60] H. Amy Xia, Haijun Ma, and Bradley P. Carlin. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics*, 21(5):1006–1029, 2011.

[61] V. Calian, D. M. Li, and J. C. Hsu. Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical Journal*, 50(5):756–766, 2008.

[62] Valerie S. L. Williams, Lyle V. Jones, and John W. Tukey. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69, 1999.

[63] R. Heller, Y. Golland, R. Malach, and Y. Benjamini. Conjunction group analysis: An alternative to mixed/random effect analysis. *Neuroimage*, 37(4):1178–1185, 2007.

[64] Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008.

[65] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

[66] David B. Dunson, Amy H. Herring, and Stephanie M. Engel. Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, 103(482):534–546, 2008.

[67] Scott M. Berry, Bradley P. Carlin, J. Jack Lee, and Peter Müller. *Bayesian adaptive methods for clinical trials*, volume 38. CRC Press, 2010.

[68] Stephen Evans. An answer to multiple problems with analysis of data on harms? *Statistical Science*, 27(3):346–347, 2012.

[69] Don Berry. Discussion of "Multivariate bayesian logistic regression for analysis of clinical trial safety issues" by W. DuMouchel. *Statistical Science*, 27(3):344–345, 08 2012.

[70] B. W. McEvoy and R. C. Tiwari. Discussion of "Multivariate bayesian logistic regression for analysis of clinical trial safety issues" by W. DuMouchel. *Statistical Science*, 27(3):340–343, 2012.

[71] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[72] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer New York, 2005.

[73] D. Collett. *Modelling Survival Data in Medical Research, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2015.

[74] R. J. Cook and J. F. Lawless. *The Statistical Analysis of Recurrent Events*. Statistics for Biology and Health. Springer, 2007.

[75] O. Aalen, O. Borgan, and H. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer New York, 2008.

[76] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford, 2001.

[77] Robert T. O'Neill. Statistical concepts in the planning and evaluation of drug safety from clinical trials in drug development: Issues of international harmonization. *Statistics in Medicine*, 14(9):1117–1127, 1995.

[78] John D. Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):214–221, 1978.

[79] J. Burridge. Empirical bayes analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):65–75, 1981.

[80] David G. Clayton. A monte carlo method for bayesian inference in frailty models. *Biometrics*, 47(2):467–485, 1991.

[81] Debajyoti Sinha. Semiparametric bayesian analysis of multiple event time data. *Journal of the American Statistical Association*, 88(423):979–983, 1993.

[82] J. G. Ibrahim, M. H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Partially Ordered Systems. Springer, 2001.

[83] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley, 2002.

[84] L. Duchateau and P. Janssen. *The Frailty Model*. Statistics for Biology and Health. Springer Verlag, 2007.

[85] David B. Dunson and Amy H. Herring. Bayesian model selection and averaging in additive and proportional hazards models. *Lifetime Data Analysis*, 11(2):213–232, 2005.

[86] Shaban Shaban and Ayman Mostafa. Shared frailty survival analysis using semiparametric bayesian method. *Interstat*, 2005. `https://ssrn.com/abstract=2556733`.

[87] Athanasios Kottas. Nonparametric bayesian survival analysis using mixtures of weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578–596, 2006.

[88] Terry M. Therneau. Extending the cox model. Report 58, Mayo Clinic, Rochester, Minnesota, 1996.

[89] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, 2000.

[90] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

[91] P. E. R. K. Andersen, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York, 1996.

[92] Richard J. Cook and Jerald F. Lawless. Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16(8):911–924, 1997.

[93] Mei-Cheng Wang, Jing Qin, and Chin-Tsang Chiang. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065, 2001.

[94] Gerd Rosenkranz. Analysis of adverse events in the presence of discontinuations. *Drug Information Journal*, 40(1):79–87, 2006.

[95] Daniel F. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49(4):1099–1109, 1993.

[96] D. Y. Lin, J. M. Robins, and L. J. Wei. Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*, 83(2):381–393, 1996.

[97] Hongyu Jiang, Rick Chappell, and Jason P. Fine. Estimating the distribution of nonterminal event time in the presence of mortality or informative dropout. *Controlled Clinical Trials*, 24(2):135–146, 2003.

[98] Richard J. Cook, Jerald F. Lawless, Lajmi Lakhal-Chaieb, and Ker-Ai Lee. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association*, 104(485):60–75, 2009.

[99] Lei Liu, Robert A. Wolfe, and Xuelin Huang. Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3):747–756, 2004.

[100] Debashis Ghosh and D. Y. Lin. Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*, 59(4):877–885, 2003.

[101] Yu Zhangsheng and Lei Liu. A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statistics in Medicine*, 30(22):2683–2695, 2011.

[102] J. F. Lawless and C. Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168, 1995.

[103] W. B. Nelson. *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, 2003.

[104] Jixian Wang and George Quartey. Nonparametric estimation for cumulative duration of adverse events. *Biometrical Journal*, 54(1):61–74, 2012.

[105] Jixian Wang and George Quartey. A semi-parametric approach to analysis of event duration and prevalence. *Computational Statistics & Data Analysis*, 67(0):248–257, 2013.

[106] Xiaobing Zhao and Xian Zhou. Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data Analysis*, 56(2):370–383, 2012.

[107] Joel A. Dubin and Stephanie S. O'Malley. Event charts for the analysis of adverse events in longitudinal studies: An example from a smoking cessation pharmacotherapy trial. *Open Epidemiology Journal*, 3:34–41, 2010.

[108] J. Heinrich. Drug safety: most drugs withdrawn in recent years had greater health risks for women. *United States General Accounting Office*, 2001. `http://www.gao.gov/products/GAO-01-286R`.

[109] Donald A. Berry. *Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective*. University of Minnesota, School of Statistics, 1987.

[110] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, 2000.

[111] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[112] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2017. R package version 3.1-131.

[113] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[114] Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2006.

[115] Terry M. Therneau. *coxme: Mixed Effects Cox Models*, 2015. R package version 2.2-5.

[116] P. Diggle, P. Heagerty, K. Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. OUP Oxford, 2002.

[117] Scott L. Zeger and Peter J. Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50(3):689–699, 1994.

[118] Nicholas Lange, Bradley P. Carlin, and Alan E. Gelfand. Hierarchical bayes models for the progression of hiv infection using longitudinal cd4 t-cell numbers. *Journal of the American Statistical Association*, 87(419):615–626, 1992.

[119] J. S. Schildcrout, C. A. Jenkins, J. H. Ostroff, D. L. Gillen, F. E. Harrell, and D. C. Trost. Analysis of longitudinal laboratory data in the presence of common selection mechanisms: a view toward greater emphasis on pre-marketing pharmaceutical safety. *Stat Med*, 27(12):2248–66, 2008. Schildcrout, Jonathan S Jenkins, Cathy A Ostroff, Jack H Gillen, Daniel L Harrell, Frank E Trost, Donald C Research Support, Non-U.S. Gov't England Stat Med. 2008 May 30;27(12):2248-66.

[120] S. Berry. Meta-analysis versus large trials: Resolving the controversy. In Dalene Stangl and Donald A. Berry, editors, *Meta-Analysis in Medicine and Health Policy*, pages 65–82. CRC Press, 2011.

[121] James G. Scott and James O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

[122] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 2004.

[123] C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, 2007.

[124] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer New York, 1999.

[125] Wenfeng Chen, Naiqing Zhao, Guoyou Qin, and Jie Chen. A bayesian group sequential approach to safety signal detection. *Journal of Biopharmaceutical Statistics*, 23(1):213–230, 2013.

[126] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012.

[127] Stephen J W Evans, David Prieto-merino, David J Spiegelhalter, John Whittaker, and Liam Smeeth. Implications of using different types of priors in a 3-level hierarchical bayesian model for the analysis of drug adverse events. International Biometric Conference, 2008.

[128] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[129] Trevor R. Shaddox, Patrick B. Ryan, Martijn J. Schuemie, David Madigan, and Marc A. Suchard. Hierarchical models for multiple, rare outcomes using massive observational healthcare databases. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(4):260–268, 2016.

[130] ColinJohn Crooks, David Prieto-Merino, and StephenJ W. Evans. Identifying adverse events of vaccines using a bayesian method of medically guided information sharing. *Drug Safety*, 35(1):61–78, 2012.

[131] Alan Agresti and Bernhard Klingenberg. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):691–706, 2005.

[132] Robert Goldberg-Alberts and Sam Page. Multivariate analysis of adverse events. *Drug Information Journal*, 40(1):99–110, 2006.

[133] A. L. Gould. Detecting potential safety issues in clinical trials by bayesian screening. *Biometrical Journal*, 50(5):837–51, 2008.

[134] Hanjoo Kim, Justine Shults, Scott Patterson, and Robert Goldberg-Alberts. Analysis of adverse events in drug safety: A multivariate approach using stratified quasi-least squares. *UPenn Biostatistics Working Papers*, 29, 2008.

[135] Gerd K. Rosenkranz. An approach to integrated safety analyses from clinical studies. *Drug Information Journal*, 44(6):649–657, 2010.

[136] Annick Joëlle Nembot Simo. Approximation de la distribution a posteriori d'un modèle gamma-poisson hiérarchique à effets mixtes. Master's thesis, Université de Montréal, 2011.

[137] Cindy L. Christiansen and Carl N. Morris. Hierarchical poisson regression modeling. *Journal of the American Statistical Association*, 92(438):618–632, 1997.

[138] H. Southworth and M. O'Connell. Data mining and statistically guided clinical review of adverse event data in clinical trials. *Journal of Biopharmaceutical Statistics*, 19(5):803–817, 2009.

[139] Raymond Carragher. *c212: Methods for Detecting Safety Signals in Clinical Trials Using Body-Systems (System Organ Classes)*, 2017. R package version 0.93. `https://CRAN.R-project.org/package=c212`.

[140] Ralph B. D'Agostino, Warren Chase, and Albert Belanger. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician*, 42(3):198–202, 1988.

[141] Roderick J. A. Little. Testing the equality of two independent binomial proportions. *The American Statistician*, 43(4):283–288, 1989.

[142] J. M. Bernardo. The concept of exchangeability and its applications. 4:111–121, 1996.

[143] Elizabeth J. Atkinson, Cynthia S. Crowson, Rachael A. Pedersen, and Terry M. Therneau. Poisson models for person-years and expected rates. Report 81, Mayo Clinic, Rochester, Minnesota, 2008.

[144] Michael Friedman. Piecewise exponential models for survival data with covariates. pages 101–113, 1982.

[145] Nan Laird and Donald Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240, 1981.

[146] Theodore R. Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2):299–305, 1980.

[147] John Whitehead. Fitting cox's regression model to survival data using glim. *Journal of the Royal Statistical Society. Series C (Applied Statistics*, 29(3):268–275, 1980.

[148] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[149] Chunlin Ji and Scott C. Schmidler. Adaptive markov chain monte carlo for bayesian variable selection. *Journal of Computational and Graphical Statistics*, 22(3):708–728, 2013.

[150] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[151] Mei-Chen Hu, Martina Pavlicova, and Edward V. Nunes. Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375, 2011.

[152] Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003.

[153] Shirley Kenneth. *Inference from Simulations and Monitoring Convergence*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, 2011.

[154] C. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Use R! Springer, 2010.

[155] William A. Link and Mitchell J. Eaton. On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115, 2012.