Statistical Evaluation of Drug Safety in Clinical Trials

By

Jiawei Duan

M.S. Statistics, George Washington University, 2015

B.S. Mathematics, Fuzhou University, 2013

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Jo A. Wick, Ph.D., Chairperson

Byron J. Gajewski, Ph.D.

Jonathan D. Mahnken, Ph.D.

Matthew S. Mayo, Ph.D.

Scott Weir, Pharm.D., Ph.D.

Date defended: November 05, 2019

The Dissertation Committee for Jiawei Duan certifies
that this is the approved version of the following dissertation :

Statistical Evaluation of Drug Safety in Clinical Trials

Jo A. Wick, Ph.D., Chairperson

Jo A. Wick, Ph.D., Program Director

Date approved: November 18, 2019

Abstract

The evaluation of drug safety is critically important in clinical trials. The first part of this dissertation explores new statistical methods for drug safety signal detection in two-arm clinical trials. Current statistical methods for safety signal detection in two-arm clinical trials are typically based on comparing only the incidence rates of adverse events (AEs) using frequentist p values or Bayesian posterior probabilities, regardless of AE severity. To enhance the safety signal detection, chapter 2 of this dissertation describes a frequentist test for evaluating both the AE incidence rate and AE severity in two-arm clinical trials. The frequentist test is based on the Fisher's exact test for AE incidence rate and a proposed conditional test for AE severity that adjusts for potential selection bias. Moreover, in chapter 3 of this dissertation, from the Bayesian perspective, we further proposed a Bayesian three-level hierarchical non-proportional odds version of the cumulative logit model for detecting safety signal with respect to both the incidence rate and severity when all the AEs reported from a two-arm clinical trial are classified into different body system. The three-level hierarchical prior structure takes advantage of the classification of AEs and adjusts for multiplicity because information is borrowed across AEs, especially across the AEs within the same body system. The second part of this dissertation explores statistical applications for safety monitoring in two-arm clinical trials. A few statistical methods for blinded safety monitoring have been proposed. The complex nature of these methods makes the applications challenging. In chapter 4 of this dissertation, we developed two user-friendly R Shiny interactive tools to accelerate, facilitate and improve the process of blinded safety monitoring and reporting in two-arm clinical trials. The interactive tools are based on two blinded safety monitoring methods proposed by Gould & Wang (2017) and Ball (2011) respectively. The dissertation concludes with summary and future studies in chapter 5.

Acknowledgements

I'm so grateful to all of the people who have helped me and supported me during my time here at KU.

I would like to express my deepest gratitude and appreciation to my supervisor, Dr. Wick, for her guidance, understanding, and caring during my graduate studies. She has been everything a student could ask for in a mentor. Her supervision on my research and my future work has always been helpful.

In addition, I would like to thank my dissertation committee members, Dr. Gajewski, Dr. Mahnken, Dr. Mayo and Dr. Weir for their expertise and helpful advice along the way.

I would like to specially thank Dr. Mayo, the Department of Biostatistics & Data Science and the National Cancer Institute for supporting and financing my graduate education and conference travel along the way. I am also obliged to KUMC Student Professional Development Award for the financial support.

Further thanks go to Dr. Gao and Dr. Wang from Regeneron Pharmaceuticals for their collaborative work on my statistical application paper.

I'm grateful to Dr. Diaz. My work on his projects enhanced my research skills and introduced me to interesting areas and methods within biostatistics.

I also would like to thank our Graduate Education Coordinator, Mandy Rametta for her help with scheduling and arranging my dissertation defense, as well as many other aspects during my graduate studies.

It's been my privilege to meet many friends while living in Kansas City. To Guanlin, Guangyi, Richard, Peng, Bing, Shibo, Junhao, Yi, Pengcheng, Duncan, Chuanwu, Lauren, Yang, Xiaqing, Qing, Palash, Mohammed and others I'm sure I have left out: thank you.

Finally, I want to express my deepest gratitude and love for my parents Shaodong and Yuanhua, thank you for your sacrifices and your love. I could not have hoped for more supportive parents.

Contents

1 Introduction **1**

1.1 Clinical trial safety evaluation 2

1.1.1 Safety monitoring 4

1.1.2 Safety signal detection 5

1.2 Post-marketing safety evaluation 7

1.3 Research motivation and current studies 8

2 Statistical Evaluation of Causal Treatment Effect on the Incidence and Severity of Adverse Events in Clinical Trials **12**

2.1 Introduction 13

2.2 Notations and composite null 17

2.3 Combining Separate Tests for Testing the Composite Null 19

2.3.1 Testing $H_0^{(1)} : \theta_1 = \theta_2$ and $H_0^{(2)} : F_c(z) = F_t(z)$ 20

2.3.2 Methods for Combining Separate Tests 20

2.4 Test for Causal Treatment Effect on AE Severity 21

2.4.1 Biased sampling model 21

2.4.2 Hypothesis Testing of Causal Effect 27

2.4.3 Bootstrap resampling 28

2.5 Simulation study 29

2.6 Application 33

2.7 Discussion 37

3 Assessing the Incidence and Severity of Drug Adverse Events: a Bayesian Hierarchical Cumulative Logit Model **39**

3.1 Introduction 40

3.2 Bayesian hierarchical logit model: a review of Berry's method 43

3.2.1 Safety data structure . . . 44
3.2.2 Likelihood functions and priors . . . 44
3.2.3 Hypothesis test and decision rule . . . 47
3.3 Cumulative logit model for safety signal detection . . . 47
3.3.1 Data structure and notations . . . 48
3.3.2 Bayesian hierarchical non-proportional odds version of cumulative logit model . . . 50
3.3.3 Solo Bayesian non-proportional odds cumulative logit model . . . 55
3.3.4 Hypothesis test and decision rule . . . 56
3.4 Simulation study . . . 56
3.4.1 Simulation setup . . . 57
3.4.2 Simulation results . . . 60
3.5 Application . . . 62
3.6 Discussion . . . 66

4 Interactive Tools for Blinded Safety Monitoring in Clinical Trials using Bayesian Methods 68

4.1 Introduction . . . 69
4.2 Bayesian approach based on pooled rate . . . 72
4.2.1 Binary event data - Beta-binomial model . . . 72
4.2.2 Exposure adjusted incidence rate - Poisson gamma model . . . 73
4.3 Bayesian approach based on treatment effect metrics . . . 74
4.3.1 Binary event data - Beta-binomial model . . . 74
4.3.2 Exposure adjusted incidence rate - Poisson gamma model . . . 76
4.4 Interactive tools . . . 78
4.4.1 An interactive tool based on pooled rate . . . 79
4.4.2 An interactive tool based on treatment effect metrics . . . 83
4.5 Discussion . . . 89

5 Summary and Future Directions **91**

A Some Derivations for Chapter 4 **100**

A.1 Single binomial trial 100

A.2 Single Poisson trial 101

List of Figures

2.1 Plot of p-values versus the degree of selection bias for each adverse event in the safety data of isotretinoin trial 36

3.1 Overview of the Bayesian hierarchical prior structure of Berry & Berry (2004)'s model 46

3.2 Overview of the Bayesian hierarchical prior structure of the non-proportional odds version of the cumulative logit model 53

3.3 Priors of the solo Bayesian non-proportional odds version of the cumulative logit model 56

4.1 "About" tab of the first interactive tool 79

4.2 Beta distribution density plot in the "Prior Visualization" tab 80

4.3 Gamma distribution density plot in the "Prior Visualization" tab 80

4.4 "Binary event rate - beta-binomial model" tab in the first tool 81

4.5 "Incidence rate - Poisson-Gamma model" tab in the first tool 82

4.6 "About" tab of the second interactive tool 84

4.7 "Risk Ratio" sub-tab under "Binary event rate - beta-binomial model" tab in the second tool 86

4.8 "Risk Difference" sub-tab under "Binary event rate - beta-binomial model" tab in the second tool 87

4.9 "Risk Ratio" sub-tab under "Incidence rate - Poisson-Gamma model" tab in the second tool 88

4.10 "Risk Difference" sub-tab under "Incidence rate - Poisson-Gamma model" tab in the second tool 89

List of Tables

2.1 The basic principal stratum or strata to which the subjects who experience the AE belong, and the information available on $Z_i(t)$ and $Z_i(c)$ 24

2.2 Six scenarios for simulation study 30

2.3 Type I error rate $\times 100\%$ and power $\times 100\%$ of the five tests under scenario 1 to 3 . 32

2.4 Power $\times 100\%$ of the five tests under scenario 4 to 6 34

2.5 Toxicity frequency and corresponding proportion for randomized eligible patients by study arms in L.G. Leon-Novelo & Muller (2010) 35

2.6 P values of the Fisher's exact test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$ for each AE and the corresponding adjusted p values using Hochberg procedure 35

3.1 Safety data structure with only AE incidence information 44

3.2 Safety data structure with AE incidence and AE severity information 49

3.3 Probability vectors of AEs in the control group for simulation study 58

3.4 The posterior probability that $H_1^{(bj)}$ is true and the posterior probability that $\omega_{bj,k} > 0, k = 1,2,3$ for each adverse event A_{bj} 60

3.5 FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenario 1, the global null 60

3.6 Powers and FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenarios 2 to 3 61

3.7 Powers and FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenarios 4 to 7 62

3.8 Simulated safety dataset for data analysis 64

3.9 The posterior probability that $H_1^{(bj)}$ is true and the posterior probability that $\omega_{bj,k} > 0, k = 1,2,3$ for A_{bj} based on two Bayesian cumulative logit models 65

Chapter 1

Introduction

1.1 Clinical trial safety evaluation

Clinical trials are prospective biomedical or behavioral studies designed on human participants to answer questions about new treatments and known interventions that warrant further study and comparison. In clinical trials, the evaluation of drug efficacy and drug safety are two major goals. The purpose of drug efficacy evaluation is to study whether the experimental drug is efficacious in disease treatment and it requires a well designed trial and the trial must be identified unambiguously before it starts to support the efficacy of the experimental drug (Gould, 2018).

The evaluation of safety aims at characterizing the safety profile of the experimental drug and ensuring timely alteration or termination of the trial to protect trial participants from potentially harmful treatment. Drug safety evaluation is quite different from the evaluation of efficacy. As pointed out by Wang et al. (2017), the efficacy evaluation tends to be linear, targeted and hypothesis driven while the safety evaluation tends to be iterative, holistic and dynamic.

Some defining characteristics of safety monitoring and reporting from the sponsor's perspective are described by Wang et al. (2017):

> "Safety monitoring is a process and involves a wide range of stakeholders.
>
> It assesses side effects across a spectrum of frequencies and magnitudes...
>
> It serves to lay the foundation for an integrated analysis of safety and benefit-risk analysis in regulatory submissions, such as a new drug application (NDA), or for a possible advisory committee meeting..."

Safety evaluation is critically important in drug development. Only treatments that are concluded as reasonably safe will be approved by regulatory authorities. In recent years, the US Food and Drug Administration (FDA) has issued guidance regarding safety monitoring and reporting for an investigational new drug (IND) to assist fuller development of safety profiles, as shown in the FDA guidance, see FDA (2010, 2012, 2015). The FDA guidance states that sponsors should develop a safety assessment committee and a safety surveillance plan as key elements of a systematic approach to safety surveillance. The evolving safety profile of the IND should be evaluated

based on cumulative AEs from all of the trials in the development program. In addition, the Safety Planning, Evaluation and Reporting Team (SPERT) was formed in 2006 by the Pharmaceutical Research and Manufacturers of America (PhRMA) to recommend a pharmaceutical industry standard for safety planning, data collection, evaluation and reporting, see Crowe et al. (2009) and Xia et al. (2011).

Drug safety is evaluated on the basis of adverse events (AE). For example, a new experimental drug might cause physical signs (rash), symptoms (fever), laboratory tests (serum albumin) and other relevant assessments (blood glucose). These adverse effects exist when patients' body functions, such as liver function, renal function, hematology, neurologic function, are affected by the experimental treatments. Close analysis of the safety data containing incidence and severity information of adverse events is important in order to improve the timing of identifying risks and justify the safety of the treatment that warrant a next stage clinical trial or regulatory agency approval.

Adverse events are typically classified into three tiers (Crowe et al., 2009; Gould, 2002): Tier 1 AEs are those associated with specific hypotheses being tested formally in the clinical trial and they are empty in many trials. Tier 2 AEs are those routinely collected in clinical trials but about which no specific hypotheses have been formulated in advance. The goal is to identify any unexpected deleterious effects of the drug and to quantify their rates. Tier 3 AEs are rare spontaneous reports of serious events that require specific clinical evaluation. See also Berry & Berry (2004).

Statistics play an important role in evaluating the drug's safety profile. An American Statistical Association (ASA) safety working group was established to take advantage of the leading role of the statistician and better enable the quantification in drug safety monitoring.

Statistical literatures on drug safety evaluation include two major areas of drug safety evaluation: "safety signal detection" and "safety monitoring" (Zhu et al., 2016). They are different regarding the number of AEs included in the analysis. In the following sections, we provide a comprehensive review of statistical methods of safety evaluation in clinical trials.

1.1.1 Safety monitoring

The first type of safety evaluation is the safety monitoring. It aims at monitoring an adverse event of special interest (AESI) in an ongoing trial. Some Bayesian methods for dynamically monitoring the AESI in the ongoing single-arm trials have been proposed. One feature of these methods is that the Bayesian framework allows for the inclusion of prior information and the dynamic updates of the posterior distribution of the parameters as data accrue. The decision rule is that if the posterior probability of exceeding a certain rate is greater than a threshold, it indicates a safety signal with respect to the AESI. For example, Yao et al. (2013) adapted a Bayesian decision criterion for monitoring AESIs. This Bayesian decision criterion was first proposed by Thall & Simon (1994) for the evaluation of drug efficacy in a Phase IIb single-arm oncology study. Chen & Chaloner (2006) proposed another Bayesian approach for continuous monitoring the AESI. Unlike Thall & Simon (1994), they only consider the distribution of treatment group event rate. For safety monitoring of an AESI in a two-arm trial, see Bayesian Beta-Binomial model and Poisson-Gamma model in Yao et al. (2013) and Zhu et al. (2016).

In addition, frequentist monitoring of clinical safety could be performed under the group sequential paradigm. For example, sequential probability ratio test (SPRT) (Wald, 1945) can be used for monitoring the AESI in a single-arm trial. However, A limitation of the test is the need to specify the single alternative hypothesis (Li & Kulldorff, 2010). Several other sequential tests were proposed. See Kulldorff et al. (2011), Goldman & Hannan (2001), Bolland & Whitehead (2000), Li & Kulldorff (2010), Shih et al. (2010).

The above methods were developed for unblinded safety monitoring, where treatment information is known. Unblinded safety monitoring is often conducted by an external data monitoring committee (DMC) periodically. As sponsors develop novel drugs and biological agents in new diseases or therapeutic areas, there is an increasing need for sponsors to monitor patient safety to detect potential safety signals as soon as possible, regardless of DMC schedule, while maintaining study blinding to minimize bias and protect the integrity of the blinded studies. The trial sponsors review blinded reports and listings of safety data on a regular basis and make determinations on

whether there is a change in the risk profile of the drug based on clinical judgment without preset decision rules or criteria. Although blinded data analysis is less informative and does not provide a definitive treatment effect estimate using blinded data alone, blinded safety data monitoring could identify potential safety issues before the data is unblinded and prompt decisions regarding an unblinded analysis. There are a few methods in the literature about blinded safety data monitoring. Ball (2011) described an approach using adverse event rate as an example based on pooled event rate from a two-arm randomized trial. The decision rule is that if there is a high posterior probability given cumulative data that the observed pooled rate is higher than control or background rate in the literature, then there is a potential safety signal with respect to the AE. Gould & Wang (2017) proposed a Bayesian approach for safety monitoring of the two-arm randomized trial where the potential adverse event risk levels can be estimated with different treatment effect metrics such as relative risk, absolute risk difference, or odds ratio. Schnell & Ball (2016) introduced a safety monitoring procedure for two-arm blinded clinical trials that incorporates a Bayesian hierarchical exposure-time model to make inferences on the rate of AESI in the test treatment arm. Mukhopadhyay et al. (2018) proposed a two-step Bayesian method for monitoring and detecting safety signals from blinded safety data for AESI. Lin et al. (2019) proposed a two-stage framework incorporating periodic analyses of blinded safety data to detect AEs that may have potential treatment effect on the treatment group incidence rate, as well as planned unblinded analyses to quantify associations between the drug and AEs.

1.1.2 Safety signal detection

More often, the analysis of safety data involves many types of tier 2 AEs that may not be anticipated (Mehrotra & Heyse, 2004). The second type of safety evaluation in clinical trials is safety signal detection, where all AEs instead of just an AESI are included in the analysis. In a two-arm clinical trial, the aim of the safety signal detection is to compare the incidence rates of all the AEs between two groups. If the incidence rates of some of the AEs in the treatment group are significantly greater than those in the control group (or vice versa), these AEs will be flagged

and further investigation is needed about the safety of the drug. Simultaneously comparing the incidence rates of many AEs causes multiplicity issues. This is a common challenge that faces statisticians. Bayesians and frequentists alike. Failure to adjust for multiplicity will give excess false positive findings, thus needlessly complicating the interpretation of the safety profile of the experimental drug.

One frequentist adjustment for multiplicities is to control for overall type I error rate by using a lower nominal significance level. This is the idea of Bonferroni adjustment. However, such adjustments are conservative as the power of the test will be very low. Other more suitable frequentist adjustments are preferred. In the clinical adverse event context, Mehrotra & Heyse (2004) proposed a double false discover rate procedure (DFDR), a novel method for addressing multiplicity by controlling the false discovery rate (FDR) to a desired level in the evaluation of adverse event data. The method was a two-step application of adjusted p values based on the false discovery rate procedure of Benjamini & Hochberg (1995). Furthermore, Mehrotra & Adewale (2012) proposed another DFDR procedure, which is an enhancement of the DFDR, that significantly lowers the FDR without materially compromising the power for detecting true signals. The DFDR procedure and the new DFDR procedure can be used to adjust p values of any individual tests. The p value of individual test can be for testing the inequality of incidence rate, inequality of exposure adjusted incidence rate, or inequality of severity between the control group and the treatment group. Diao et al. (2019) developed Monte-Carlo-based methods for the safety signal detection in clinical trials. The proposed methods account for the rare events and arbitrary correlation structures among AEs within and/or between body system.

Moreover, some Bayesian methods have been proposed for safety signal detection and adjustment of multiplicity issues. The Bayesian approach is less tied to type I error. It focuses on assessing the probability that the treatment causes an adverse event on the basis of all available information (Berry & Berry, 2004). McEvoy et al. (2013) introduced a Bayesian approach for modeling the risk differentials of the AEs between the treatment and comparator arms. The proposed procedure uses an Ising prior to unite medically related AEs. Berry & Berry (2004) proposed

a three-level Bayesian hierarchical model to monitor the incidence rate differences of many adverse events between two groups. The hierarchical model accounts for multiplicities in adverse event assessment. It provides an explicit method for borrowing information across types of adverse events. The posterior probability that the odds ratio of the the adverse event incidence rate between the control group and treatment group is obtained for each adverse event. Xia et al. (2011) expanded Berry & Berry (2004)'s method into a hierarchical Poisson mixture model which accounts for the length of the observation of subjects and improves the characteristics of the analysis for rare events.

Gould (2008), Gould (2013), Gould (2018) proposed an alternative Bayesian screening approach to detect potential safety issues when event counts arise from binomial or Poisson distributions. The method assumes that the adverse event incidences are realizations from a mixture of distributions and seeks to identify the element of the mixture corresponding to each adverse event.

DuMouchel (2012) described a multivariate Bayesian logistic regression (MBLR) method for model-based analysis of safety data when there are rare events and sparse data from a pool of clinical trials. As with Berry's method, MBLR allows information from the different AEs to "borrow strength" from each other. The logistic regression model also examines the relationship between AE frequencies to multiple covariates and to treatment by covariate interactions, which enables a search for vulnerable subgroups.

Safety signal detection is usually conducted after the clinical trial is finished. When it is conducted in the ongoing clinical trial, in addition to the multiplicity issue of multiple events at each look, there is a second dimension of multiplicity issue: multiple looks during the course of a clinical trial. Chen et al. (2013) applied a Bayesian hierarchical model in a group sequential manner for multiple interim analyses of safety events. A decision-theoretic approach is employed to determine threshold values in the process of safety signal detection.

1.2 Post-marketing safety evaluation

Safety evaluation continues to play an important role in the post-marketing phase when the drug gets approved by regulatory authority. The goal of safety evaluation in the post-marketing phase is

to study the association between the specific drug and AE and identify the AEs with high reporting rates compared to other AEs associated with a particular drug or identify drugs associated with high reporting rate of an AE compared to the other drugs (Huang et al., 2011).

Compare to the safety data collected in clinical trials, more data sources become available as a drug moves into the post-approval phase and questions that require larger exposures or longer treatment duration can be answered. Therefore, safety data in the post-approval phase fill an information gap of clinical trials. Safety data in the post-marketing phase is collected through spontaneous reporting (SR). In spontaneous reporting, health care professionals or patients report suspected AEs from a drug to the local or national drug administration. SR data is usually stored in some databases. For example, the Adverse Event Reporting System (AERS) was established in 1968 by US Food and Drug Administration to collect AEs associated with drugs on the market through a spontaneous reporting system. The data in AERS is made available quarterly online.

Various methods have been developed for safety signal detection in SR data. For example, proportional reporting ratios (Evans et al., 2001), the likelihood ratio tests (Huang et al., 2011, 2013, 2014), Bayesian methods (DuMouchel, 1999; Bate et al., 1998; Hu et al., 2015).

1.3 Research motivation and current studies

The statistical methods for safety signal detection discussed in the last section are typically based on comparing only the incidence of AEs between two groups, regardless of the severity. It is possible that two treatments might have same incidence rate for an AE but the severity of the AE may be greater for one treatment versus the other. For example, suppose the severity of an AE has three levels: mild, moderate or severe. The probabilities that the severity of the AE is moderate or severe are both higher for one treatment versus the other even if the incidence rates of the AE are the same in both groups. In this case, it would be unappealing for the AE not to be flagged. This motivates us to develop new statistical methods to enhance the safety signal detection in clinical trials.

In the first part of this dissertation, we have developed statistical methods to enhance safety

signal detection in two-arm clinical trials by comparing both the incidence and severity of the AEs. Thus, AEs with higher incidence rate or greater severity in the treatment group will be flagged. Interest lies in testing the following composite null hypothesis for each of the AEs

$$H_0 : \theta_1 = \theta_2 \quad \text{and} \quad \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 \tag{1.1}$$

versus the alternative hypothesis that $\theta_1 < \theta_2$ and/or treatment group has greater severity of the AE than the control group. θ_1, θ_2 are the incidence rate of the AE in the control group and treatment group respectively. $\boldsymbol{\pi}_1 = (\pi_{11}, \pi_{12}, \pi_{13})^T, \boldsymbol{\pi}_2 = (\pi_{21}, \pi_{22}, \pi_{23})^T$ are two vectors of probabilities of three different severity levels of the AE in control and treatment group respectively if we categorize the severity of an adverse event as mild, moderate and severe.

Greater severity of an adverse event in treatment group is defined as that the AE severity score in the treatment group is stochastically greater than that in the control group, i.e.,

$$\pi_{13} \leq \pi_{23}$$

$$\pi_{12} + \pi_{13} \leq \pi_{22} + \pi_{23}$$

where at least one inequality is strict. We use this to define greater severity of an AE because among several formally defined notions, the least stringent is stochastic order. See Cohen & Sackrowitz (2000) and Cohen et al. (2000).

We first developed a method from the frequentist point of view in Chapter 2. The composite null hypothesis in (1.1) is an intersection of two null hypotheses: $H_0 : H_0^{(1)} \cap H_0^{(2)}$ where $H_0^{(1)} : \theta_1 = \theta_2$ and $H_0^{(2)} : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$. Thus testing the composite null hypothesis in (1.1) is equivalent to testing the individual hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ and then combine the p-values of the individual test.

Denote p_1, p_2 as the p values for one-tailed tests of $H_0^{(1)}$ versus $H_1^{(1)} : \theta_1 < \theta_2$, and $H_0^{(2)}$ versus $H_1^{(2)} : \boldsymbol{\pi}_2$ is stochastically larger than $\boldsymbol{\pi}_1$, respectively. p_1 is the p value of the Fisher's exact test for the equality of AE incidence rate. Since the test for the AE severity is based on the

subjects selected who are selected based on a post-randomization event (AE occurrence), it does not assess a causal treatment effect. To adjust for potential selection bias, we further propose a biased sampling model, which is an extension of the work by Gilbert et al. (2003) that is based on the principal stratification framework developed by Frangakis & Rubin (2002), and a procedure for testing causal treatment effect on the AE severity. p_2 is the p value of the proposed test for AE severity. Simes' test (Simes, 1986) and Fisher's test (Fisher, 1932) are introduced to combine the p-values p_1 and p_2. Once we obtained the combined p value for testing the composite null hypothesis for each of the AEs, multiple testing procedure such as DFDR (Mehrotra & Adewale, 2012) can be applied in order to adjust for multiplicity.

In addition to the frequentist method, we can also evaluate the incidence rate and severity of AEs from the Bayesian perspective. As reviewed in the last section, Berry & Berry (2004) proposed a three-level Bayesian hierarchical model to detect safety signal with respect to the incidence rate of AEs. This method accounts for multiplicities in adverse event assessment. It provides an explicit method for borrowing information across types of adverse events. The posterior probability that the odds ratio of the the adverse event incidence rate between the control group and treatment group is obtained for each AE.

In Chapter 3, we extend the hierarchical model of Berry & Berry (2004) for comparing both the incidence rate and severity of all AEs between the control group and treatment group. We proposed a three-level Bayesian hierarchical non-proportional version of the cumulative logit model. The primary drug safety outcome is a four level ordinal categorical variable, representing four AE severity levels from none, mild, moderate to severe. Our model allows for testing the equivalence of incidence rate and severity for all the AEs simultaneously between the control group and the treatment group. Similar to the feature of the model in Berry & Berry (2004), the method self-adjusts for the multiplicity issue with the help of the hierarchical nature as it borrows information across types of AEs, especially across the AEs within the same body system. We conduct simulation study to investigate the power and false discovery rate of the proposed hierarchical model and compare it to a non-hierarchical cumulative logit model with only the first level of prior structure.

The simulation results show that, in general, the proposed method not only controls for false discovery rate but also performs well in detecting safety signals when either the incidence rate or the severity is greater in the treatment group.

Furthermore, as can be seen in literature introduced in the last section, the complex nature of the current methods make the blinded safety monitoring and reporting challenging. In order to accelerate, facilitate and improve the process of safety monitoring and reporting, it is necessary to develop ready-to-use tools based on the existing safety monitoring and reporting methods. In Chapter 4, we develop two R Shiny interactive tools for blinded safety data monitoring. The blinded data monitoring function of the Shiny interactive tools are based on the method proposed by Gould & Wang (2017) and Ball (2011) respectively. The tool can be used to perform computations for posterior probability of the pooled event rate or treatment effect metrics and dynamically output the monitoring result about whether the data suggest a safety signal.

The rest of this dissertation is organized as follows. In chapter 2, we introduce the frequentist method for evaluating the causal treatment effect on the incidence and severity of adverse events in clinical trials. In chapter 3, we describe the three-level Bayesian hierarchical non-proportional version of the cumulative logit model for assessing the incidence and severity of AEs. In chapter 4, two R-Shiny interactive tools for blinded safety monitoring of AESI are developed. We conclude this dissertation with summary and future research in chapter 5.

Chapter 2

Statistical Evaluation of Causal Treatment Effect on the Incidence and Severity of Adverse Events in Clinical Trials

Abstract

Clinical safety data in two-arm trials are routinely evaluated using between group p values for every reported adverse event (AE), with multiple testing procedure applied to the p values to adjust for multiplicity. However, the p value generated for each AE is often based on comparing only the AE incidence rate between two randomized groups, regardless of AE severity. To enhance the evaluation of drug safety, for each AE, we propose to use AE occurrence and severity as co-primary endpoints and to perform a statistical test of the composite null hypothesis that both the incidence rate and severity are equivalent in two groups. The p-value of the test for the composite null hypothesis is obtained by combining the p-values of the Fisher's exact test for AE incidence and a proposed test for AE severity, respectively. The proposed test for AE severity is based on an extension of a biased sampling model initially developed by Gilbert et al. (2003, Biometrics 59, 531-541) for continuous outcome. We conduct a simulation study to investigate the power and type I error rate of the proposed test compared to the usual test for AE incidence. The simulation results show that, with large enough sample size, the proposed method performs as well or better than the test for AE incidence in detecting a safety signal. The proposed method is demonstrated via an application in cancer.

Key words: Adverse events; Causal inference; Composite null hypothesis; Posttreatment selection bias; Principal stratification; Safety signal detection; Severity.

2.1 Introduction

Drug safety evaluation is critically important in clinical trials. In recent years, the US Food and Drug Administration (FDA) has issued guidance regarding safety monitoring and reporting for an investigational new drug (IND) to assist fuller development of safety profiles, as shown in the US FDA guidance (FDA, 2010, 2012, 2015). Drug safety is evaluated on the basis of adverse events (AE) reported in the clinical trials. AEs are typically classified into body systems. Each body system contains AEs that are biologically related. Close analysis of the safety data containing

incidence and severity information of AEs characterizes the safety profile of the experimental drug to determine if it warrants a next stage trial or regulatory agency approval.

Drug safety evaluation includes two major areas: "safety monitoring" and "safety signal detection" (Zhu et al., 2016). Safety monitoring aims at monitoring an adverse event of special interest (AESI) in an ongoing trial, while in safety signal detection, all AEs instead of just an AESI are included in the analysis. The goal of safety signal detection in a two-arm clinical trial is to compare the incidence rates of all AEs between a control group and a treatment group. If the incidence rates of some of the AEs in the treatment group are significantly larger than those in the control group (or vice versa), these AEs will be "flagged" and further investigation is needed about the safety of the drug. Simultaneously comparing the incidence rates of many AEs leads to multiplicity issues. This is a common challenge that faces statisticians. Bayesians and frequentists alike. Ignoring multiplicities will give excess false positive findings, thus needlessly complicating the interpretation of the safety profile of the experimental drug.

From the frequentist perspective, to assess the equality of incidence rate of every AE encountered in the clinical trial and detect safety signal while adjusting for multiplicity issues, p value for testing the equality of incidence rate is generated for every AE and is adjusted and evaluated by the multiple testing procedure. Among several multiplicity adjustment methods, a double false discover rate (DFDR) procedure proposed by Mehrotra & Heyse (2004) is a novel method for controlling the false discovery rate (FDR) to a desired level. It is a two-step application of the false discovery rate procedure proposed by Benjamini & Hochberg (1995). Mehrotra & Adewale (2012) improved DFDR procedure that significantly lowers the FDR without materially compromising the power for detecting true signals.

In addition, some Bayesian methods have also been proposed for safety signal detection and adjustment of multiplicity issues. Berry & Berry (2004) proposed a three-level Bayesian hierarchical model to account for multiplicities in adverse event assessment. The hierarchical model provides an explicit method for borrowing information across types of adverse events. Xia et al. (2011) expanded Berry's method into a hierarchical Poisson mixture model which accounts for

the length of the observation of subjects and improves the characteristics of the analysis for rare events. DuMouchel (2012) described a multivariate Bayesian logistic regression (MBLR) method for model-based analysis of safety data when there are rare events and sparse data from a pool of clinical trials. The logistic regression model examines the relationship between AE frequencies to multiple covariates and to treatment by covariate interactions, which enables a search for vulnerable subgroups. Gould (2008, 2013, 2018) proposed an alternative Bayesian screening approach to detect potential safety signals when event count follows binomial or Poisson distribution.

As we can see from the literature, safety signal detection is often based on comparing only the incidence of adverse events between two groups, regardless of the AE severity. It is possible that for some AEs, the incidence rate might be the same in both groups but the severity is "greater" for one group versus the other. In this case, it would be unappealing for the AE not to be flagged.

To fully capture the presence and severity of adverse events, it is important to incorporate an endpoint that describes the severity of each AE. Klingenberg et al. (2009) proposed a method for investigating the toxicity effect of a chemical compound on animals in an environmental study. They introduced a single primary endpoint to represent the presence and severity of every type of toxicity effect of the chemical compound. They used permutation test and a bootstrap method for testing the simultaneous marginal homogeneity for all the toxicity effect of the chemical compound and adjusted the p values to control for family wise error rate (FWER). The method can be readily carried over to safety analysis in clinical trials. However, power of the test based upon single endpoint for each type of AE is low for detecting certain alternatives of interest, for instance, when the AE incidence rate is the same but the severity is different. Recently, Duan et al. (2019b) proposed a three-level Bayesian hierarchical non-proportional version of the cumulative logit model for assessing the incidence and severity of drug AEs in two-arm clinical trials. Their method not only controls for false discovery rate but also performs well in detecting safety signals when either the incidence rate or the severity is greater in the treatment group.

In this article, we seek to enhance the p value for evaluating each AE. We propose to use AE occurrence and severity as co-primary endpoints and to perform a statistical test of the compos-

ite null hypothesis that the incidence rate and severity are equivalent between groups. The first endpoint, AE occurrence, is a binary variable. The second endpoint, AE severity, is a 3-level ordinal categorical variable. For more information about the severity level used in clinical trial, see "Common Terminology Criteria for Adverse Events" published in the National Cancer Institute. The p value of the test of the composite null hypothesis is obtained by combining the p value of the Fisher?s exact test for AE incidence and the p value of the test for AE severity using Simes' method (Simes, 1986) and Fisher's method (Fisher, 1932). See Shih & Quan (1997) and Mehrotra et al. (2006) for a discussion of the statistical testing of the composite hypothesis.

The test for AE severity is restricted to subjects who are selected based on a post-randomization event (AE occurrence). This poses a major challenge to making an unbiased inference of the treatment effect on AE severity. Gilbert et al. (2003) proposed methods for adjusting post-randomization selection bias in the context of HIV vaccine trials. Their methods are based on the principal stratification framework developed by Frangakis & Rubin (2002). However, the second endpoint they considered is the viral load set point of a subject infected by HIV, which is a continuous variable but the second endpoint in our problem is an ordinal categorical variable. We extend the method of Gilbert et al. (2003) to adjust for selection bias. Simulation studies are conducted to investigate the power and type I error rate of different tests and to investigate the power of the combined tests after adjusting for potential selection bias.

The rest of the paper is organized as follows. In Section 2.2, we introduce notation and define the composite null hypothesis. In Section 2.3, we describe the combined test for testing the composite null hypothesis. In Section 2.4, we introduce a proposed method for adjusting for selection bias. In Section 2.5, we compare the power of different tests in a comprehensive simulation study and then in Section 2.6 we apply the proposed method in a clinical trial. We conclude the article in Section 2.7.

2.2 Notations and composite null

Suppose the drug safety evaluation is performed in a two-group clinical trial that consists of a control group and a treatment group. Our goal is to detect AEs with safety signals. An AE has a safety signal if it has greater incidence rate or greater severity in the treatment group.

We proposed to use two co-primary endpoints for each AE: AE occurrence and AE severity. AE occurrence is a binary outcome, indicating whether the AE occurs or not. AE severity is an ordinal categorical outcome, representing the severity of the AE. Without loss of generality, we assume there are three AE severity levels: mild, moderate and severe (or 1,2 and 3). This can be easily extended to more severity levels, for example, grade 1 to 4 severity level. See "Common Terminology Criteria for Adverse Events" published by the National Cancer Institute.

Suppose there are a total of N subjects in two groups and the number of subjects in the control group and treatment group are N_1 and N_2 respectively. For a specific AE, let $y_{1i} = 1$ if the i^{th} subject in the control group experiences the AE and 0 if he or she does not experience the AE. $i = 1, ..., N_1$. Let $y_{2i} = 1$ if the i^{th} subject in the treatment group experiences the AE and 0 if he or she does not experience the AE. Denote $\theta_1 = P(y_{1i} = 1), \theta_2 = P(y_{2i} = 1)$ as the incidence rates of the AE in control and treatment group respectively. Denote $x_1 = \sum_{i=1}^{N_1} y_{1i}, x_2 = \sum_{i=1}^{N_2} y_{2i}$ as the number of subjects with the AE in control and treatment group respectively. Thus $x_1 \sim Bin(N_1, \theta_1), x_2 \sim Bin(N_2, \theta_2)$.

Let z_{1i} and z_{2i} be the severity score (1,2 or 3) of the i^{th} subject in the control group and treatment group respectively. Of course $z_{1i}(z_{2i})$ exists only if $y_{1i} = 1 (y_{2i} = 1)$. Denote $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ as two vectors of severity scores of the subjects with the AE in the control and treatment group respectively. For subjects with the AE, let $\boldsymbol{n}_1 = (n_{11}, n_{12}, n_{13})$ be a vector of the number of subjects whose AE severity level is $1, 2, 3$ respectively in the control arm, where

$$n_{11} = \sum_{i=1}^{N_1} I\{z_{1i} = 1\}, n_{12} = \sum_{i=1}^{N_1} I\{z_{1i} = 2\}, n_{13} = \sum_{i=1}^{N_1} I\{z_{1i} = 3\}$$

And let $\boldsymbol{n}_2 = (n_{21}, n_{22}, n_{23})$ be a vector of the number of subjects whose AE severity level is $1, 2, 3$

respectively in the treatment arm, where

$$n_{21} = \sum_{i=1}^{N_2} I\{z_{2i} = 1\}, n_{22} = \sum_{i=1}^{N_2} I\{z_{2i} = 2\}, n_{23} = \sum_{i=1}^{N_2} I\{z_{2i} = 3\}$$

Let $\boldsymbol{\pi}_1 = (\pi_{11}, \pi_{12}, \pi_{13})^T, \boldsymbol{\pi}_2 = (\pi_{21}, \pi_{22}, \pi_{23})^T$ be two vectors of probabilities that the AE severity is 1, 2 or 3 respectively in control and treatment groups, where

$$\pi_{11} = P(z_{1i} = 1 | y_{1i} = 1), \pi_{12} = P(z_{1i} = 2 | y_{1i} = 1), \pi_{13} = P(z_{1i} = 3 | y_{1i} = 1)$$

$$\pi_{21} = P(z_{2i} = 1 | y_{2i} = 1), \pi_{22} = P(z_{2i} = 2 | y_{2i} = 1), \pi_{23} = P(z_{2i} = 3 | y_{2i} = 1)$$

We assume that $\boldsymbol{n}_1 \sim multi(x_1, \boldsymbol{\pi}_1), \boldsymbol{n}_2 \sim multi(x_2, \boldsymbol{\pi}_2)$.

In addition to the use of two co-primary endpoints, it is also reasonable to use one primary endpoint. Let w_{1i} and w_{2i} be the severity of the i^{th} subject in the control group and treatment group respectively. The value of w_{1i} and w_{2i} can be $0, 1, 2, 3$. $w_{1i} = 0 (w_{2i} = 0)$ if subject i in the control group (treatment group) does not experience the AE. $w_{1i} = 1 (w_{1i} = 1)$ if subject i in the control group (treatment group) experience the AE and the severity score is 1, etc. Also denote $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ as two vectors of severity of the subjects who experience the AE in the control and treatment group respectively.

Let $\boldsymbol{m}_1 = (m_{11}, m_{12}, m_{13}, m_{14})$ be a vector of the number of subjects whose AE severity is $0, 1, 2, 3$ respectively in the control arm, where

$$m_{11} = \sum_{i=1}^{N_1} I\{w_{1i} = 0\}, m_{12} = \sum_{i=1}^{N_1} I\{w_{1i} = 1\}, m_{13} = \sum_{i=1}^{N_1} I\{w_{1i} = 2\}, m_{14} = \sum_{i=1}^{N_1} I\{w_{1i} = 3\}$$

And let $\boldsymbol{m}_2 = (m_{21}, m_{22}, m_{23}, m_{24})$ be a vector of the number of patients whose AE severity is $0, 1, 2, 3$ respectively in the treatment arm, where

$$m_{21} = \sum_{i=1}^{N_2} I\{w_{2i} = 0\}, m_{22} = \sum_{i=1}^{N_2} I\{w_{2i} = 1\}, m_{23} = \sum_{i=1}^{N_2} I\{w_{2i} = 2\}, m_{24} = \sum_{i=1}^{N_2} I\{w_{2i} = 3\}$$

Let $\boldsymbol{\phi}_1 = (\phi_{11}, \phi_{12}, \phi_{13}, \phi_{14})^T, \boldsymbol{\phi}_2 = (\phi_{21}, \phi_{22}, \phi_{23}, \phi_{14})^T$ be two vectors of probabilities that the AE severity is 0, 1, 2 or 3 respectively in control and treatment groups. We assume $\boldsymbol{m}_1 \sim multi(N_1, \boldsymbol{\phi}_1), \boldsymbol{m}_2 \sim multi(N_2, \boldsymbol{\phi}_2)$.

Denote $F_c(z)$ and $F_t(z)$ $(z = 1, 2, 3)$ as the cumulative density functions of the severity score of the subjects who experience the AE in the control group and treatment group respectively.

The research goal is to test the composite null hypothesis

$$H_0 : H_0^{(1)} \cap H_0^{(2)}$$

where $H_0^{(1)} : \theta_1 = \theta_2$ and $H_0^{(2)} : F_c(z) = F_t(z)$ versus the one-sided composite alternative hypothesis:

$$H_1 : H_1^{(1)} \cup H_1^{(2)}$$

where $H_1^{(1)}$ is $\theta_1 < \theta_2$, and $H_1^{(2)}$ is the AE severity is greater in the treatment group than that in the control group.

The AE severity is greater in the treatment group than that in the control group if $F_c(z) > F_t(z), z = 1, 2$. This means that

$$\pi_{13} \leq \pi_{23}, \pi_{12} + \pi_{13} \leq \pi_{22} + \pi_{23}$$

where at least one inequality is strict. We use this definition as the greater severity of an adverse event because among several formally defined notions, the least stringent is stochastic order. See Cohen & Sackrowitz (2000) and Cohen et al. (2000). Thus $H_1^{(2)}$ is $\pi_{13} \leq \pi_{23}, \pi_{12} + \pi_{13} \leq \pi_{22} + \pi_{23}$ where at least one inequality is strict.

2.3 Combining Separate Tests for Testing the Composite Null

To test the composite null hypothesis, we conduct the individual test of $H_0^{(1)}$ and $H_0^{(2)}$ separately and combine the p values of the tests using Simes method or Fisher's method. In this section we

introduce the methods for testing $H_0^{(1)}$ and $H_0^{(2)}$ and then introduce methods for combining the p values.

2.3.1 Testing $H_0^{(1)} : \theta_1 = \theta_2$ and $H_0^{(2)} : F_c(z) = F_t(z)$

We use one-sided Fisher's exact test for testing $H_0^{(1)}$. Denote p_1 as the p value for the test of

$$H_0^{(1)} \quad \text{versus} \quad H_1^{(1)}$$

The test for the second individual hypothesis $H_0^{(2)} : F_c(z) = F_t(z)$ is restricted to subjects who are selected based on a post-randomization event (AE occurrence) and it is possible that the severity outcomes of subjects in control group and the treatment group are not from a completely randomization procedure and thus may not be comparable.

To test $H_0^{(2)}$, we propose a test for comparing the severity score z_1, z_2 of the subjects who experience the AE. The test is based on an extension of a biased sampling model proposed by Gilbert et al. (2003). Details about the test are introduced in Section 2.4. Denote p_2 as the p value of the test for:

$$H_0^{(2)} \quad \text{versus} \quad H_1^{(2)}$$

2.3.2 Methods for Combining Separate Tests

We consider the following two combination methods for testing the composite null hypothesis of an AE at level α. Note that p_1 and p_2 derived from Fisher' exact test for AE incidence and the proposed test for AE severity respectively are stochastically independent under H_0. This result has been proved by Shih & Quan (1997) in an unrelated context and it establishes the validity of the combination tests.

1. Simes' method Simes (1986): Reject H_0 if $\max(p_1, p_2) < \alpha$ or $\min(p_1, p_2) < \alpha/2$

2. Fisher's method Fisher (1932): Reject H_0 if $p < \alpha$ where $p = P(\chi_4^2 > -4\log(\sqrt{p_1 p_2}))$

The performances of the above two methods have been studied by Shih & Quan (1997). No method is uniformly superior to the other. The choice between Simes' method and Fisher's method requires prior knowledge of the alternative hypothesis H_1. As Shih & Quan (1997) pointed out, unless the AE severity is stochastically greater in the treatment group but the AE incidence rate is similar in both groups (or the opposite), we would expect the Fisher's test to be superior than Simes test.

2.4 Test for Causal Treatment Effect on AE Severity

In this section we introduce a biased sampling model and how we use it to test the causal treatment effect on the severity of an AE, i.e., to test $H_0^{(2)} : F_c(z) = F_t(z)$ unbiasedly. The biased sampling model originally proposed by Gilbert et al. (2003) is based on the principal stratification framework developed by Frangakis & Rubin (2002) for causal inference.

2.4.1 Biased sampling model

Following Gilbert et al. (2003), theoretically, each subject has two potential outcomes of adverse event occurrence: one under the assignment to the control group $Y_i(c)$ and one under assignment to the treatment group $Y_i(t)$. $Y_i(c) = 1(Y_i(t) = 1)$ if the subject has the adverse event under the assignment to the control group (treatment group) and $Y_i(c) = 0(Y_i(t) = 0)$ if the subject does not have the adverse event under the assignment to the control group (treatment group). In addition, each subject with the adverse event under assignment to control group has a potential severity outcome $Z_i(c)$ and under assignment to treatment group has a potential severity outcome $Z_i(t)$. For each subject, only one of $Y_i(c)$ or $Y_i(v)$ is observed and $Z_i(c)(Z_i(v))$ is defined only if $Y_i(c) = 1(Y_i(v) = 1)$.

By property 2 of Frangakis & Rubin (2002), a causal treatment effect on the severity of the adverse event can be defined based on the comparison between the sets $\{Z_i(c) : Y_i(c) = Y_i(t) = 1\}$ and $\{Z_i(t) : Y_i(c) = Y_i(t) = 1\}$ because the comparison is made within the principal stratum of subjects who would always experience the AE regardless of randomization to control or treatment drug.

For subjects in the set $\{Z_i(c) : Y_i(c) = Y_i(t) = 1\}$, suppose $Z_i(c)$ are identically distributed as $F^{alw.}_{(c)}(z)$ and for subjects in the set $\{Z_i(t) : Y_i(c) = Y_i(t) = 1\}$, suppose $Z_i(t)$ are identically distributed as $F^{alw.}_{(t)}(z)$, also denote $f^{alw.}_{(c)}(z)$ and $f^{alw.}_{(t)}(z)$ as the probability mass function that corresponds to $F^{alw.}_{(c)}(z)$ and $F^{alw.}_{(t)}(z)$ respectively. Then any functional that measures a contrast of the distributions

$$F^{alw.}_{(c)}(z) = Pr(Z_i(c) \leq z | Y_i(c) = Y_i(t) = 1) \quad \text{and}$$
$$F^{alw.}_{(t)}(z) = Pr(Z_i(t) \leq z | Y_i(c) = Y_i(t) = 1)$$

is a causal estimand (Gilbert et al., 2003). Thus to test the second null hypothesis that there is no causal treatment effect on the severity of adverse event $(H^{(2)}_0 : F_c(z) = F_t(z))$, we compare $F^{alw.}_{(c)}(z)$ and $F^{alw.}_{(t)}(z)$. Or equivalently, to compare $f^{alw.}_{(c)}(z)$ and $f^{alw.}_{(t)}(z)$. The second null hypothesis can thus be rewritten as $H^{(2)}_0 : F^{alw.}_{(c)}(z) = F^{alw.}_{(t)}(z)$.

Unfortunately, because neither distribution in is readily identifiable for us to make comparisons (because $Y_i(c)$ and $Y_i(t)$ are not both observed). To test the causal effect of treatment on severity of AE, we need to make the following assumptions:

1. The potential AE occurrence outcomes for each subject are independent of the treatment assignments of other subjects

2. The treatment assignment for each patient is independent of his or her potential outcomes

3. The intervention used in the control group does not increase the risk of experiencing the AE compare to the treatment group, or the experimental treatment does not purposely cure the AE. Thus the incidence rate of the AE in the control group is less than or equal to that in the treatment group

Assumption 1 is actually implied by Rubin's (1978) stable unit treatment value assumption (SUTVA) (Gilbert et al., 2003). With this assumption, the potential AE occurrence outcome of a subject can be written as a function of the treatment assignment of the subject instead of being

written as a function of the treatment assignment of the subject and all other subjects, i.e., it can be written as $Y_i(c)$ and $Y_i(t)$. Assumption 2 holds due to randomization and blinding of the clinical trial.

Assumption 3 means that for a subject, if he/she experience the AE after being administered the intervention of the control group, he/she will experience the AE after being administered the intervention of the treatment group, given all the other experimental conditions are the same. The assumption is reasonable as the control group is usually a group of subjects who are administered the lower dose of the treatment (or placebo) and the treatment group is usually a group of subjects who are administered the higher dose of the treatment. The incidence rate of the adverse event in the group with lower dose is likely to be less than that in the group with higher dose. Assumption 3 can be checked by testing if the AE incidence rate is higher in control group than treatment group recipients for any participant subgroup.

These three assumptions are very important because only based on these assumptions are we able to make the following statistical inferences.

Denote $f_{(c)}(z)$ and $f_{(t)}(z)$ as the probability mass function (pmf) of the AE severity level in subjects with AE under randomization to control group and the pmf of the AE severity level in subjects with adverse event under randomization to treatment group, respectively. $F_{(c)}(z)$ and $F_{(t)}(z)$ are the corresponding cumulative density function. Under assumption 2, $f_{(c)}(z)$ and $f_{(t)}(z)$ are also the pmf of the AE severity level outcome of subjects with AE from control group and treatment group respectively.

Table 2.1 shows the principal stratum or strata to which a subject with AE must belong, and lists the information available on potential severity level outcome. The tables makes clear that the set of subjects $\{Y_i(c) = 1, Y_i(t) = 1\}$ is the natural subpopulation for causal inference on severity level since it is the only stratum in which severity level outcome is observable from the data.

From Table 2.1 we know $F^{alw.}_{(c)}(z) = F_{(c)}(z)$, or equivalently, $f^{alw.}_{(c)}(z) = f_{(c)}(z)$. Thus $F^{alw.}_{(c)}(z)$ is identified from the observed data. $F^{alw.}_{(t)}(z)$ cannot be identified by the above assumptions. However, from Table 2.1 we know the subjects who have the AE in the treatment group consists of the

Table 2.1: The basic principal stratum or strata to which the subjects who experience the AE belong, and the information available on $Z_i(t)$ and $Z_i(c)$

Randomized assignment	Is AE present	Principal Stratum $\{Y_i(c), Y_i(t)\}$	
Control group	Yes	$\{Y_i(c)=1, Y_i(t)=0\}$ (empty set by assumption 3)	$\{Y_i(c)=1, Y_i(t)=1\}$ $Z_i(c)$ observed, $Z_i(t)$ unobserved
Treatment group	Yes	$\{Y_i(c)=0, Y_i(t)=1\}$ $Z_i(c)$ undefined, $Z_i(t)$ observed	$\{Y_i(c)=1, Y_i(t)=1\}$ $Z_i(t)$ observed, $Z_i(c)$ unobserved

subjects who will always have the AE regardless of randomization to control or treatment group and the subjects who will not have the AE if he/she is administered control group treatment. For subjects in the set $\{Z_i(c) : Y_i(c)=1, Y_i(t)=0\}$, denote $f_{(t)}^{prot.}(z)$ as the pmf of $Z_i(c)$. Thus $f_{(t)}(z)$ can be written as a mixture of $f_{(t)}^{prot.}(z)$ and $f_{(t)}^{alw.}(z)$ (Gilbert et al., 2003):

$$f_{(t)}(z) = P(Y_i(c)=0|Y_i(t)=1) f_{(t)}^{prot.}(z) + P(Y_i(c)=1|Y_i(t)=1) f_{(t)}^{alw.}(z)$$

Next we prove that $P(Y_i(c)=0|Y_i(t)=1) = 1 - RR^{-1}$ so that

$$f_{(t)}(z) = (1 - RR^{-1}) f_{(t)}^{prot.}(z) + RR^{-1} f_{(t)}^{alw.}(z)$$

where $RR = \frac{\theta_2}{\theta_1} = \frac{Y_i(t)=1}{Y_i(c)=1}$ is the relative risk of of the AE between treatment group and control

group. Note that

$$
\begin{aligned}
&1 - RR^{-1} \\
&= 1 - \frac{Pr(Y_i(c)=1)}{Pr(Y_i(t)=1)} \\
&= \frac{Pr(Y_i(t)=1) - Pr(Y_i(c)=1)}{Pr(Y_i(t)=1)} \\
&= \frac{Pr(Y_i(t)=1, Y_i(c)=1) + Pr(Y_i(t)=1, Y_i(c)=0) - Pr(Y_i(c)=1, Y_i(t)=1) - Pr(Y_i(c)=1, Y_i(t)=0)}{Pr(Y_i(t)=1)} \\
&= \frac{Pr(Y_i(t)=1, Y_i(c)=0) - Pr(Y_i(c)=1, Y_i(t)=0)}{Pr(Y_i(t)=1)} \\
&= \frac{Pr(Y_i(t)=1, Y_i(c)=0))}{Pr(Y_i(t)=1)} \\
&= P(Y_i(c)=0 | Y_i(t)=1)
\end{aligned}
$$

Thus

$$
\begin{aligned}
f_{(t)}(z) &= P(Y_i(c)=0|Y_i(t)=1) f_{(t)}^{prot.}(z) + P(Y_i(c)=1|Y_i(t)=1) f_{(t)}^{alw.}(z) \\
&= (1 - RR^{-1}) f_{(t)}^{prot.}(z) + RR^{-1} f_{(t)}^{alw.}(z)
\end{aligned}
$$

With some calculations , the above mixture can be re-expressed as a biased sampling model (Gilbert et al., 2003):

$$
f_{(t)}^{alw.}(z) = W^{-1} w(z) f_{(t)}(z)
$$

where $w(z) = Pr(Y_i(c)=1|Z_i(t)=z, Y_i(t)=1)$ and $W^{-1} = (\sum_{z=1}^{3} w(z) f_{(t)}(z))^{-1}$ is a normalizing constant equal to RR. The weight function $w(z)$ is the probability that a subject who is randomized to treatment group and has the adverse event with severity level z would have the adverse event if randomized to control group.

If $w(z)$ were known then $f_{(t)}^{alw.}(z)$ would be identified. However, $w(z)$ is unknown and it is not possible to test whether a particular w is correctly specified. The approach to this problem

by Gilbert et al. (2003) is to assume $w()$ is known. They proposed a logistic function for $w(z)$. In their context, the response variable is a continuous variable. However, the severity endpoint in our context is an ordinal categorical variable. Thus, the logistic function may not be used here. Instead, we set value for each $w(z), z = 1,2,3$, guided by our beliefs about plausible degrees of selection bias. We propose the following measure of weight: $w(1) = w(1|\gamma,r) = \gamma, w(2) = w(2|\gamma,r) = r\gamma, w(3) = w(3|\gamma,r) = r^2\gamma$, $r(>0)$ is the relative risk of the occurrence of adverse event under randomization to control group given the occurrence of adverse event under randomization to treatment group with severity level z versus with severity level $z-1$, $z = 2,3$. In this way, the unidentified sensitivity function $w()$ is interpretable, which makes the approach fruitful and is important (Gilbert et al., 2003). Thus,

$$f_{(t)}^{alw.}(z) = RR \times w(z|\gamma,r) f_{(t)}(z) = f_{(t)}(z|r)$$
$$F_{(t)}^{alw.}(z) = \sum_{d=1}^{z} RR \times w(d|\gamma,r) f_{(t)}(d) = F_{(t)}(z|r)$$

Given fixed r, γ is determined as the solution to the equation $F_{(t)}(3|r) = 1$.

If $RR = 1$, i.e., $W = 1$ and thus $w(z) = Pr(Y_i(c) = 1|Z_i(t) = z, Y_i(t) = 1) = 1$, then there is no selection bias and $f_{(t)}^{alw.}(z) = f_{(t)}(z)$. If $RR > 1$, then whether there is selection bias depends on the value of $w(z)$ and thus depends on r.

Fixing $r = 1$ specifies a constant weight, i.e., $\gamma = RR^{-1}$ and the weights will be $w(1) = w(2) = w(3) = RR^{-1}$ and reflects an assumption of no selection bias. Thus when $RR = 1$ and/or we fix $r = 1$, there will be no selection bias and the second null hypothesis $H_0^{(2)} : F_{(c)}^{alw.}(z) = F_{(t)}^{alw.}(z)$ can be tested by simply comparing the severity of the subjects with the adverse event in both groups.

Fixing $r > 1$ makes $w(z|\gamma,r)$ an increasing function of z and it means some factors other than treatment make the severity levels of the subjects in treatment group small, then to be fair for control group, we should adjust the distribution of the treatment group so that its severity is stochastically larger. The larger r is from 1, the higher degree of bias we believe. Similarly, $r < 1$ makes $w(z|\gamma,r)$ an decreasing function of z and it means some factors other than control group treatment

make the severity levels of the subjects in control group small, then to be fair for treatment group, we should adjust the distribution of the treatment group so that its severity is stochastically smaller. The smaller r is from 1, the higher degree of bias we believe.

We estimate RR with $\hat{RR} = x_2/x_1$. We estimate $f_{(c)}(z), F_{(c)}(z)$ and $f_{(t)}(z), F_{(t)}(z)$ with the maximum likelihood estimator.

$$\hat{f}_{(c)}(1) = n_{11}/x_1, \hat{f}_{(c)}(2) = n_{12}/x_1, \hat{f}_{(c)}(3) = n_{13}/x_1, \hat{\boldsymbol{\pi}}_1 = (n_{11}/x_1, n_{12}/x_1, n_{13}/x_1)$$

$$\hat{F}_{(c)}(z) = \sum_{d=1}^{z} \hat{f}_{(c)}(d), z = 1,2,3$$

$$\hat{f}_{(t)}(1) = n_{21}/x_2, \hat{f}_{(t)}(2) = n_{22}/x_2, \hat{f}_{(t)}(3) = n_{23}/x_2, \hat{\boldsymbol{\pi}}_1 = (n_{21}/x_2, n_{22}/x_2, n_{23}/x_2)$$

$$\hat{F}_{(t)}(z) = \sum_{d=1}^{z} \hat{f}_{(t)}(d), z = 1,2,3$$

Thus the estimator of $f_{(t)}^{alw.}(z)$ and $F_{(t)}^{alw.}(z)$ are

$$\hat{f}_{(t)}^{alw.}(z) = \hat{f}_{(t)}(z|r) = \hat{RR} \times w(z|\gamma, r)\hat{f}_{(t)}(z)$$

$$\hat{F}_{(t)}^{alw.}(z) = \hat{F}_{(t)}(z|r) = \sum_{d=1}^{z} \hat{RR} \times w(d|\gamma, r)\hat{f}_{(t)}(d), z = 1,2,3$$

Given fixed value of r, γ in $w(z|\gamma, r)$ can be obtained by solving $\hat{F}_{(t)}(3|r) = 1$.

2.4.2 Hypothesis Testing of Causal Effect

If selection bias is presumed to follow the selection bias model, then the causal null hypothesis of interest for the severity of adverse event is $H_0^{(2)} : F_{(c)}^{alw.}(z) = F_{(t)}^{alw.}(z)$, the corresponding alternative hypothesis of interest is: $H_0^{(2)} : F_{(c)}^{alw.}(z) > F_{(t)}^{alw.}(z)$. This means that the severity endpoint of the subjects in the treatment group is stochastically larger than that in the control group.

To test the second null hypothesis $H_0^{(2)} : F_{(c)}^{alw.}(z) = F_{(t)}^{alw.}(z)$, we propose a test statistic, denote as T_r, that is the Wilcoxon rank sum test statistic calculated using the adjusted and observed AE severity of subjects in the control and treatment groups, respectively. $(\boldsymbol{z}_1, \boldsymbol{z}_{2,r})$. The adjustment of

the AE severity of the subjects in the treatment group is: we replace the vector $\boldsymbol{n}_2$ with

$$\boldsymbol{n}_{2,r} = (x_2\hat{f}_{(t)}(1|r), x_2\hat{f}_{(t)}(2|r), x_2\hat{f}_{(t)}(3|r))$$

which is the mean vector of the estimated distribution $F^{alw.}_{(t)}(z)$. $(\boldsymbol{n}_{2,r})$ thus indicates the number of subjects with severity level $1,2,3$ in $\boldsymbol{z}_{2,r}$. We reject the null if the p value is less than the significance level α.

2.4.3 Bootstrap resampling

Because the data we obtained $(\boldsymbol{z}_1, \boldsymbol{z}_{2,r})$ are not exactly from the distributions $F^{alw.}_{(c)}(z)$ and $F^{alw.}_{(t)}(z)$ ($\boldsymbol{z}_{2,r}$ is the estimated data from $F^{alw.}_{(t)}(z)$), we cannot use the p value we obtained from the usual Wilcoxon rank sum test. Thus the null distribution of T_r is intractable under $H_0^{(2)} : F^{alw.}_{(c)}(z) = F^{alw.}_{(t)}(z)$. The p-value based on T_r, denoted by p_2 is obtained using the following modification of the parametric bootstrap procedure developed by Hudgens et al. (2003).

Suppose $N_1 = N_2$ (that is, there are an equal number of trial participants in each arm) and we estimate RR with $\hat{RR} = x_2/x_1$ if $x_1 < x_2$ and we estimate RR with 1 if $x_1 \geq x_2$. Then for $\hat{RR} > 1$, generate bootstrap sample $\boldsymbol{n}_2^*$ from multinomial distribution with parameter x_2 and $\hat{\boldsymbol{\pi}}_2$. Generate bootstrap sample $\boldsymbol{n}_1^*$ from multinomial distribution with parameter x_1 and $(\hat{f}_{(t)}(1|r), \hat{f}_{(t)}(2|r), \hat{f}_{(t)}(3|r))$. For $\hat{RR} = 1$, generate bootstrap sample $\boldsymbol{n}_1^*$ from multinomial distribution with parameter $x_1, \hat{\boldsymbol{\pi}}$ and $\boldsymbol{n}_2^*$ from multinomial distribution with parameter $x_2, \hat{\boldsymbol{\pi}}$, where $\hat{\boldsymbol{\pi}} = (\boldsymbol{n}_1 + \boldsymbol{n}_2)/(x_1 + x_2)$ is the estimated probabilities of three severity levels.

The bootstrap test statistic T_r^* is the Wilcoxon rank sum test statistic calculated using the bootstrap sample and adjusted bootstrap sample in the control and treatment groups, respectively $(\boldsymbol{z}_1^*, \boldsymbol{z}_{2,r}^*)$. The adjustment of bootstrap sample in the treatment group is the same as that in Section 2.4.2. We generate 500 bootstrap test statistic T_r^* and the p-value is obtained by calculating proportion of the 500 bootstrap test statistic that is smaller than the observed test statistic T_r.

2.5 Simulation study

We conduct simulation study to compare the empirical power and type I error rate of different tests, including traditional Fisher's exact test for AE incidence rate (FET), Wilcoxon rank sum test for stochastic order of 4 level AE toxicity endpoint (WT), proposed test for AE severity (SEV), proposed test for the composite null using Simes' method (PS) and Fisher's method (PF). Note that the null hypotheses that correspond to FET and SEV are the equality of AE incidence rate and the equality of AE severity respectively. The null hypothesis that corresponds to PS, PF or WT is the composite null hypothesis. WT is actually the Wilcoxon rank sum test applied to the 4 level severity score $\boldsymbol{w}_1, \boldsymbol{w}_2$.

We assume equal sample size in both the control group and treatment group. Data is generated in two steps: in the first step, the number of subjects who experience the AE in each group (x_1, x_2) is generated and then the number of subjects who experience the AE with severity outcomes classified into each severity level in each group $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is generated. In the first step, given the incidence rate of the AE in the control group and treatment group respectively (θ_1, θ_2), we generate a random variable from Bernoulli distribution with parameter θ_1 for the control group, and then we generate a random variable from Bernoulli distribution with parameter θ_2 for the treatment group. We continue generating Bernoulli random variable like this for each group until the summation of the Bernoulli random variables generated in the control group and the treatment group is at least x. x is given in advance. The reason we fix x is to investigate how the power changes as we increase x, which can be directly observed from the data. In the second step, $\boldsymbol{n}_1$ is generated from a multinomial distribution with parameter x_1 and $(f_{(c)}(1), f_{(c)}(2), f_{(c)}(3))$ and $\boldsymbol{n}_2$ is generated from a multinomial distribution with parameter x_2 and $(f_{(t)}(1), f_{(t)}(2), f_{(t)}(3))$. Note that $f_{(c)}(z)$ and $f_{(t)}(z)$ are determined by $f_{(c)}^{alw.}(z)$ and $f_{(t)}^{alw.}(z)$ and the true degree of selection bias r_{true}. Thus we set true values for $f_{(c)}^{alw.}(z)$ and $f_{(t)}^{alw.}(z)$ respectively and obtain the true values of $f_{(c)}(z)$ and $f_{(t)}(z)$ by transforming $f_{(c)}^{alw.}$ and $f_{(t)}^{alw.}$ according to the equations introduced in Section 2.4 as follows, $f_{(t)}(z) = \frac{f_{(t)}^{alw.}(z)}{RR \times w(z|\gamma, r_{true})}, z = 1,2,3$. $RR = \theta_2/\theta_1, w(1|\gamma, r_{true}) = \gamma, w(2|\gamma, r_{true}) = \gamma r_{true}, w(3|\gamma, r_{true}) = \gamma r_{true}^2$. γ is determined by solving the equation of $f_{(t)}(1) + f_{(t)}(2) + f_{(t)}(3) = 1$. Besides, by assumption 2,

$(f_{(c)}(1), f_{(c)}(2), f_{(c)}(3))$ is equivalent to $(f_{(c)}^{alw.}(1), f_{(c)}^{alw.}(2), f_{(c)}^{alw.}(3))$. We consider three possible values of the true amount of selection bias r_{true} (1.25, 1, 0.8), representing moderate selection bias that is in favor of not flagging the AE, no selection bias and moderate selection bias that is in favor of flagging the AE. With data generated in this way, we can investigate the power and type I error rate of PS, PF and SEV when the prior knowledge of the degree of selection bias is correctly set ($r = r_{true}$) and when it is not ($r \neq r_{true}$).

Different parameter configurations include $\theta_1 = 0.05$, $\theta_2 = 0.05$ or 0.1, $f_{(c)}^{alw.} = (0.6, 0.3, 0.1)^T$ and $f_{(t)}^{alw.} = (0.5, 0.3, 0.2)^T$, $(0.4, 0.2, 0.4)^T$ or $(0.3, 0.2, 0.5)^T$. To measure the true difference between $f_{(c)}^{alw.}$ and $f_{(t)}^{alw.}$ (or equivalently $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$), we use an ordinal effect size measure $g = P(z_{1i} < z_{2i}) + 0.5P(z_{1i} = z_{2i})$ (Ryu & Agresti, 2008; Agresti, 2010). This measure summarizes the probability that an outcome from one distribution falls above an outcome from the other, adjusted for ties. Vargha & Delaney (1998) called *g a measure of stochastic superiority of* z_{2i} *over* z_{1i}. The measure can be written as: $g = \boldsymbol{\pi}_2^T A \boldsymbol{\pi}_1$ where

$$A = \begin{pmatrix} 0.5 & 0 & 0 \\ 1 & 0.5 & 0 \\ 1 & 1 & 0.5 \end{pmatrix}$$

g has range $[0, 1]$. If z_{1i} and z_{2i} are identically distributed, then $g = 0.5$. If z_{2i} is stochastically larger than z_{1i}, then $g > 0.5$. We finally obtain the following scenarios for simulation study.

Table 2.2: Six scenarios for simulation study

Scenario	θ_1	$\boldsymbol{\pi}_1(f_{(c)}^{alw.}(z))$	θ_2	$\boldsymbol{\pi}_2(f_{(t)}^{alw.}(z))$	RR	g
1	0.05	$(0.6, 0.3, 0.1)^T$	0.05	$(0.6, 0.3, 0.1)^T$	1	0.5
2	0.05	$(0.6, 0.3, 0.1)^T$	0.05	$(0.5, 0.3, 0.2)^T$	1	0.565
3	0.05	$(0.6, 0.3, 0.1)^T$	0.05	$(0.4, 0.2, 0.4)^T$	1	0.65
4	0.05	$(0.6, 0.3, 0.1)^T$	0.1	$(0.6, 0.3, 0.1)^T$	2	0.5
5	0.05	$(0.6, 0.3, 0.1)^T$	0.1	$(0.5, 0.3, 0.2)^T$	2	0.565
6	0.05	$(0.6, 0.3, 0.1)^T$	0.1	$(0.4, 0.2, 0.4)^T$	2	0.65

For each of 5000 datasets simulated under each parameter configuration, p values for the proposed test for severity and the proposed tests for the composite null are determined using 500

bootstrap replications.

Using a nominal 5% type I error level, Table 2.3 shows the estimated type I error rate and power of the proposed test for AE severity (SEV) and the proposed test for the composite null hypothesis based on Simes' method (PS) and Fisher's method (PF) with different presumed degree of selection bias (r) when the number of subjects who experience the AE in both groups is either 50 or 100 and when $RR = 1$ with $\theta_1 = 0.05$ and $g = 0.5, 0.565, 0.65$ with $\boldsymbol{\pi}_1 = (0.6, 0.3, 0.1)^T$ (Scenario 1-3). Since $RR = 1$, i.e., $\theta_1 = \theta_2$, according to Section 2.4, there is no selection bias, so r_{true} always has to be 1. The estimated type I error rate and power of the traditionally used Fisher's exact test for incidence rate (FET) and the Wilcoxon rank sum test for stochastic order of 4 level toxicity endpoint (WT) under corresponding parameter configurations are also included in the table. The estimated type I error rates (4th and 5th column) that correspond to PS and PF are controlled at the desired significance level of 0.05 if r is presumed to be 1. When one conservatively presume r to be less than 1, the estimated type I error rates are deflated and when r is set to be greater than 1, the estimated type I error rates are inflated. As long as the true amount of selection bias is specified (r is set to 1), PS and PF perform well in detecting the safety signal (rejecting the composite null hypothesis) when the total number of subjects who experience the AE in both groups and/or the ordinal effect size g is large enough. In addition, PS has larger power than PF. This is because the control group and the treatment group differ in one aspect (AE severity) but not in the other (AE incidence rate). This is consistent with the conclusion made by Shih & Quan (1997). In contrast, FET and WT did not effectively detect safety signal. When one conservatively presume r to be less than 1 or greater than 1, the estimated power is deflated or inflated accordingly.

Table 2.4 shows the estimated power of SEV, PS and PF with different presumed degree of selection bias (r) when the true degree of selection bias is determined by $r_{true} = 1.25, 1, 0.8$ and the number of subjects who experience the AE in both groups is either 50 or 100 and when $RR = 2$ with $\theta_1 = 0.05$ and $g = 0.5, 0.565, 0.65$ with $\boldsymbol{\pi}_1 = (0.6, 0.3, 0.1)^T$ (Scenario 4-6). The estimated power of FET for incidence rate and WT under different parameter configurations are also included in the table. As long as the true amount of selection bias is specified, PS and PF perform well in

Table 2.3: Type I error rate $\times 100\%$ and power $\times 100\%$ of the five tests under scenario 1 to 3

Method	True	Presumed	$RR = 1, g = 0.5$		$RR = 1, g = 0.565$		$RR = 1, g = 0.65$	
	r	r	$x = 50$	$x = 100$	$x = 50$	$x = 100$	$x = 50$	$x = 100$
FET	n/a	n/a	3.1	4.3	3.0	4.4	2.8	4.2
WT	1	n/a	5.9	5.3	5.6	6.3	5.5	6.2
SEV	1	1.25	8.7	11.3	31.5	49.7	75.8	93.8
	1	1	5.4	5.2	22.4	35.0	65.1	89.3
	1	0.8	3.2	3.0	15.9	23.1	50.0	72.0
PS	1	1.25	6.5	7.6	20.0	35.4	62.1	86.8
	1	1	5.0	4.7	14.0	22.5	48.9	78.5
	1	0.8	4.3	4.1	10.1	15.1	36.4	58.8
PF	1	1.25	7.9	10.8	23.5	38.7	61.6	84.8
	1	1	4.6	4.9	15.5	25.2	51.2	79.6
	1	0.8	2.8	2.6	9.2	14.6	37.6	62.7

detecting the safety signal (rejecting the composite null hypothesis) when the total number of subjects who experience the AE in both groups and/or the ordinal effect size g is large enough. In addition, PF has larger power than PS. This is because the control group and the treatment group differ consistently in both aspect of the composite null hypothesis (AE incidence rate and AE severity) (Shih & Quan, 1997). In contrast, the power of FET and WT are as good as or better than PS and PF when the total number of subjects who experience the AE in both groups is small, but as the ordinal effect size g increases, especially with large value of $g(\geq 0.65)$, the powers of PS and PF are both greater than FET and WT, meaning they can detect the safety signal more effectively. However, in scenarios when the AE severity is the same in both groups but the AE incidence rate is greater in the treatment group compare to that in the control group, FET and WT perform better than PS and PF when the sample size $(x_1 + x_2)$ is small. This is not surprising because we tend to lose some power to gain the ability to detect the safety signal with respect to AE severity.

We next illustrate the power of the proposed tests when an incorrect amount of selection bias is presumed. When there is actually no selection bias $(r_{true} = 1)$, but one conservatively presumes $r = 0.8$, the power is deflated. For larger presumed amount of selection bias, larger price will be paid (we lose more power). When there is actually no selection bias, but one presumes $r = 1.25$, the power is inflated. Thus making an incorrect assumption of selection bias can cause certain degree of power loss or power gain. If zero selection bias is presumed $(r = 1)$ but in truth there is

moderate selection bias that is in favor of flagging the AE ($r_{true} = 0.8$), the power is inflated. Since we are concerned about the composite null hypothesis, incorrectly presume the degree of selection bias when the true value of $r < 1$ does not cost us much and we might even gain some power. If zero bias is presumed ($r = 1$) but in truth there is moderate selection bias that is in favor of not flagging the AE ($r_{true} = 1.25$), we are losing power. We will lose more power to detect safety signal if in reality the selection bias is even larger ($r_{true} > 1.25$). This illustrates the importance of accounting for the possibility of selection bias to avoid missing potential safety signal.

2.6 Application

We apply the proposed method in the analysis of safety data obtained from a randomized, double-blinded phase III clinical trial conducted by National Cancer Institute (NCI). The safety data were published and analyzed by L.G. Leon-Novelo & Muller (2010). The purpose of this trial is to verify the efficacy of isotretinoin that may help control second primary tumors and mortality for stage I non-small-cell lung cancer (NSCLC) patients. One thousand, one hundred and sixty-six patients with stage I NSCLC were randomly assigned to receive either placebo or isotretinoin (30 mg/day) for 3 years. There were 589 patients who received isotretinoin while the remaining patients received placebo.

The safety data collected from the trial (shown in Table 2.5) consists of the number of patients who experienced each of the 7 AEs of interest and the corresponding number of patients within each severity level. The severity of AEs was graded using Common Toxicity Criteria for Adverse Events used by the NCI. We combine the last two severity levels into one in our analysis.

We first conduct one sided Fisher's exact test of $\theta_1 \leq \theta_2$ versus $\theta_1 > \theta_2$ to verify the assumption that placebo does not increase the risk of experiencing the AE compare to the intervention used in the treatment group. The two columns under "Fisher's exact test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$" in Table 2.6 show the p values and adjusted p values of the Fisher's exact test for incidence for each adverse event. The adjusted p values were obtained by Hochberg procedure. We can see that it is statistically significant to conclude that the incidence rate of "Headache" is greater in the control

Table 2.4: Power $\times 100\%$ of the five tests under scenario 4 to 6

Method	True	Presumed	$RR = 2, g = 0.5$		$RR = 2, g = 0.565$		$RR = 2, g = 0.65$	
	r	r	$x = 50$	$x = 100$	$x = 50$	$x = 100$	$x = 50$	$x = 100$
FET	n/a	n/a	73.2	96.3	73.1	96.3	72.2	96.4
WT	1.25	n/a	81.8	96.9	81.6	97.6	81.4	97.6
	1	n/a	81.9	97.3	80.7	97.8	82.2	97.8
	0.8	n/a	81.3	97.3	80.3	97.8	81.7	98.1
SEV	1.25	1.25	4.3	4.5	18.7	30.2	60.5	85.7
	1.25	1	2.0	1.2	7.9	9.7	31.1	50.8
	1.25	0.8	0.5	0.1	1.9	2.0	10.0	13.9
	1	1.25	11.7	16.2	38.3	64.5	85.2	98.9
	1	1	5.4	5.1	18.7	32.1	60.1	86.5
	1	0.8	1.6	0.9	6.0	8.8	28.2	49.6
	0.8	1.25	24.1	40.9	65.5	89.3	96.1	100
	0.8	1	12.6	16.8	41.5	64.4	84.8	98.4
	0.8	0.8	4.5	3.9	18.8	29.1	58.1	85.3
PS	1.25	1.25	73.4	94.2	75.1	95.4	85.8	98.8
	1.25	1	72.9	94.2	73.4	94.7	78.0	96.9
	1.25	0.8	72.8	94.1	72.6	94.3	73.1	94.8
	1	1.25	74.0	95.2	78.7	97.7	94.1	99.9
	1	1	73.1	94.5	74.4	95.8	85.6	99.0
	1	0.8	72.4	94.4	72.0	94.9	77.3	96.9
	0.8	1.25	76.5	95.7	86.2	99.4	98.4	100
	0.8	1	74.2	94.2	79.6	97.7	94.1	99.9
	0.8	0.8	72.5	93.6	73.6	95.5	85.0	99.0
PF	1.25	1.25	62.7	91.7	74.7	96.4	91.8	99.6
	1.25	1	57.7	89.1	66.9	93.2	81.6	97.9
	1.25	0.8	54.2	87.6	59.8	90.2	68.9	94.7
	1	1.25	69.3	94.5	83.7	99.0	97.5	100
	1	1	63.2	91.9	75.4	97.1	91.7	99.8
	1	0.8	57.0	89.5	66.1	93.5	80.4	98.2
	0.8	1.25	78.0	97.6	92.0	99.9	99.4	100
	0.8	1	70.1	94.3	84.4	99.0	97.4	100
	0.8	0.8	62.6	91.1	74.6	96.7	90.9	99.8

group compare to that in the treatment group. Thus the assumption that placebo does not increase the risk of experiencing the AE compare to the intervention used in the treatment group does not hold for this adverse event. So our method is inappropriate for the analysis of AE "Headache" and we exclude it from our safety analysis.

Figure 2.1 shows how p values of the following tests change as we change the degree of selection bias (either in favor of flagging the AE or of not flagging the AE) for each adverse event:

Table 2.5: Toxicity frequency and corresponding proportion for randomized eligible patients by study arms in L.G. Leon-Novelo & Muller (2010)

Toxic effect	No tox	G1	G2	G3	G4
		Placebo			
Abnormal vision	565 (0.979)	9 (0.016)	0 (0)	2 (0.003)	1 (0.002)
Arthralgia	548 (0.95)	19 (0.033)	10 (0.017)	0 (0)	-
Cheilitis	493 (0.854)	76 (0.132)	8 (0.014)	0 (0)	-
Conjunctivitis	530 (0.919)	43 (0.075)	3 (0.005)	1 (0.002)	-
Fatigue	558 (0.967)	12 (0.021)	5 (0.009)	2 (0.003)	-
Headache	554 (0.96)	16 (0.028)	3 (0.005)	4 (0.007)	-
Hyper-triglyceride	551 (0.955)	22 (0.038)	4 (0.007)	0 (0)	-
		Isotretinoin			
Abnormal vision	579 (0.983)	8 (0.014)	1 (0.002)	1 (0.002)	0 (0)
Arthralgia	544 (0.924)	30 (0.051)	10 (0.017)	5 (0.008)	-
Cheilitis	212 (0.36)	245 (0.416)	122 (0.207)	10 (0.017)	-
Conjunctivitis	449 (0.762)	98 (0.166)	31 (0.053)	11 (0.019)	-
Fatigue	572 (0.971)	14 (0.024)	3 (0.005)	0 (0)	-
Headache	580 (0.985)	9 (0.015)	0 (0)	0 (0)	-
Hyper-triglyceride	514 (0.873)	64 (0.109)	10 (0.017)	1 (0.002)	-

Table 2.6: P values of the Fisher's exact test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$ for each AE and the corresponding adjusted p values using Hochberg procedure

	Fisher's Exact Test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$	
Toxic effect	Raw	Adj.
Abnormal vision	0.396	1
Arthralgia	0.975	1
Cheilitis	1	1
Conjunctivitis	1	1
Fatigue	0.408	1
Headache	0.008	0.056
Hyper-triglyceride	1	1

Fisher's exact test, the proposed test for severity, the proposed test for incidence and severity using Fisher's method. In each plot, the red dotted line shows how the p value of the proposed test for incidence and severity using Fisher's method changes with degree of selection bias. The black solid line shows how the p value of the proposed test for severity changes with degree of selection bias. The blue dashed line represents the p value of the Fisher's exact test of $\theta_1 \geq \theta_2$ vs $\theta_1 < \theta_2$ and it does not change with the degree of selection bias. To analyze each adverse event individually,

as for AE incidence rate, "Abnormal vision" and 'Fatigue" both have same incidence rate in the control and treatment group. All other AEs have greater incidence rate in the treatment group.

As for AE severity, "Abnormal vision" and "Fatigue" both have similar overall AE severity in the control and treatment group. Note here that severity for these two AEs does not change dramatically as we change the degree of selection bias, this is because both AEs seem to have same incidence rate in the control and treatment group and there will be no selection bias according to our model, no matter what degree of selection bias we set. "Conjunctivitis" has great overall severity in the treatment group. "Arthralgia" and "Hyper-triglyceride" may have greater overall severity in the treatment group if we believe that the selection bias is in favor of not flagging the AE $(r > 1)$.

If we were to evaluate all the AEs simultaneously, p value of the proposed test for incidence and severity for each AE can be reported with multiple testing procedure such as Holmes, Hochberg and Benjamini-Hochberg procedure being used to adjust for multiplicity.

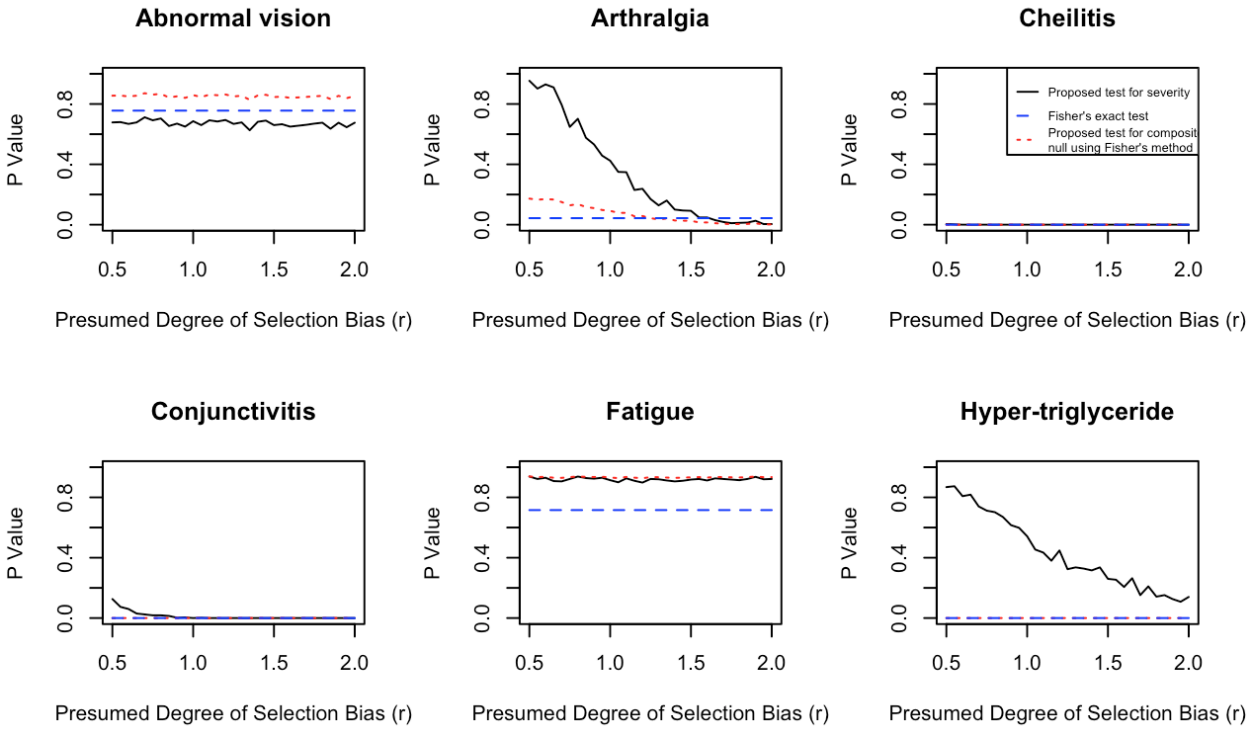

Figure 2.1: Plot of p-values versus the degree of selection bias for each adverse event in the safety data of isotretinoin trial

2.7 Discussion

Traditional analysis of safety data for AEs in clinical trials simply groups the toxicities levels into no toxicity and some toxicity and compares only the AE incidence rate between two randomized groups using a Chi-squared test or Fisher's exact test. In this article, we improve the traditional evaluation of safety data by proposing to test a composite null hypothesis for each AE that both the AE incidence rate and severity are the same in two groups. The test for the composite null involves the combination of the traditional Fisher's exact test for the AE incidence rate and a proposed conditional testing procedure for AE severity. The proposed test for AE severity is based on an extension of a bias sampling model originally developed for continues HIV viral load outcome in a vaccine trial by Gilbert et al. (2003). It is an innovative applications of causal inference method to safety signal detection area that has not been previously employed. The bias sampling model provides us a way of adjusting for severity scores of subjects in the treatment group in order to control for selection bias. The Wilcoxon rank sum test statistic is calculated to compare the adjusted severity scores of subjects in the treatment and the unadjusted severity of scores of subjects in the control group. The test does not reply on large sample theory and is applicable to rare event.

In addition to the Wilcoxon rank sum test statistic, different test statistic can also be used, for example, if we believe the distribution is skewed, Anderson-Darling type and Kolmogorov-Smirnov-type statistic may also be considered. It is worthy to note that the test statistic introduced by Lu et al. (2013) can be treated as the mean difference statistic being used on our proposed adjusted severity scores in the treatment group and the unadjusted severity scores of subjects in the control group.

The proposed method can also be applied to general randomized clinical trials, for testing causal treatment effects in the subpopulation of subjects who would experience a postrandomization event and the outcome is a ordinal categorical variable.

Some limitations remain. With only a p value of the test for the composite null reported for each AE, we may not be able to identify whether the AE has greater incidence rate or greater

severity in the treatment group. One solution to this problem is to report both the p value of the Fisher's exact test for AE incidence rate and the p value of the proposed test for AE severity.

The metric (r) that describes the degree of selection bias is determined after we review the subjects' characteristic information, thus the determination of r is subjective. We may further develop methods (for example Bayesian method) to more accurately and objectively estimate r or $w(z)$ from the data. In addition, we can assign values to each weight $w(z), z = 1,2,3$ to incorporate our prior knowledge about the potential selection bias instead of assuming that two consecutive weights $(w(z))$ have same ratio r, and thus making the proposed method more flexible.

Chapter 3

Assessing the Incidence and Severity of Drug Adverse Events: a Bayesian Hierarchical Cumulative Logit Model

Abstract

Detection of safety signals from many types of adverse events (AEs) that are reported in a two-arm clinical trial involves difficult multiplicity problems. A Bayesian hierarchical mixture model proposed by Berry and Berry in 2004 is a good solution to this problem as it borrows information across subgroups and moderates extremes due merely to chance. However, it compares only the incidence rates of AEs between the control and treatment group, regardless of the severity of AEs. In this article, we propose a three-level Bayesian hierarchical non-proportional version of the cumulative logit model. The primary drug safety outcome is a four level ordinal categorical variable, representing four AE severity levels from none, mild, moderate to severe. Our model allows for testing the equivalence of incidence rate and severity for all the AEs simultaneously between the control group and the treatment group while addressing multiplicities. We conduct simulation study to investigate the operating characteristics of the proposed hierarchical model. The simulation results show that, in general, the proposed method not only controls for false discovery rate but also performs well in detecting safety signals when either the incidence rate or the severity is greater in the treatment group. The proposed method is demonstrated via a simulated dataset from a vaccine trial.

Key words: Adverse events; Bayesian hierarchical model; Mixture model; Multiplicity; Safety signal detection; Severity.

3.1 Introduction

In clinical trials, the evaluation of drug efficacy and drug safety are two major goals. Drug safety is evaluated on the basis of adverse events periodically by data monitoring committee. The analysis of safety data typically involves many types of tier 2 AEs, which are routinely collected in clinical trials but about which no specific hypotheses have been formulated in advance (Berry & Berry, 2004). In a two-arm randomized clinical trial where participants are randomly allocated to a treatment group and a control group, interest often lies in comparing the incidence rate of many AEs

between the two randomized groups. If the incidence rates of some of the AEs in the treatment group are significantly greater than that in the control group, these AEs will be flagged and further investigation on the safety of the drug might be conducted.

However, simultaneously comparing the incidence rate of many AEs causes multiplicity issues. This is a very common challenge that face statisticians, frequentists and Bayesians alike. Failure to adjust for multiplicity increases the risk of false positives, thereby needlessly complicating the interpretation of the safety profile of the experimental treatment(s).

In the literature, some frequentist and Bayesian methods have been proposed for detecting potential inequivalence of AE incidence rate between two randomized groups and for addressing multiplicity issue in the context of clinical AE analyses.

From the frequentist perspective, to assess the equality of incidence rate of all the AEs encountered within each of several body systems while adjusting for multiplicity issues, p value for testing the equality of incidence rate is generated for every AE and is adjusted by the multiple testing procedure. Among several multiplicity adjustment methods, a double false discover rate (DFDR) procedure proposed by Mehrotra & Heyse (2004) is a novel method for controlling the false discovery rate (FDR) to a desired level. It is a two-step application of the false discovery rate procedure proposed by Benjamini & Hochberg (1995). Mehrotra & Adewale (2012) improved DFDR procedure that significantly lowers the FDR without materially compromising the power for detecting true signals.

In addition, some Bayesian methods have been developed to evaluate AEs simultaneously while controlling for overall type I error rate. Berry & Berry (2004) proposed a three-level Bayesian hierarchical mixture model to test the equivalence of incidence rate of many AEs between the control and treatment group. The hierarchical structure accounts for multiplicities in AE assessment. It provides an explicit method for borrowing information across types of AEs. The posterior probability that the odds ratio of the AE incidence rate between the treatment group and the control group is larger than a cutoff point is obtained for each adverse event.

Gould (2008, 2013, 2018) proposed an alternative Bayesian screening approach to detect po-

tential safety issues when event counts arise from binomial or Poisson distributions. The method assumes that the adverse event incidences are realizations from a mixture of distributions and seeks to identify the element of the mixture corresponding to each adverse event.

DuMouchel (2012) described a multivariate Bayesian logistic regression (MBLR) method for model-based analysis of safety data when there are rare events and sparse data from a pool of clinical trials. As with Berry's method, MBLR allows information from the different AEs to "borrow strength" from each other. The logistic regression model also examines the relationship between AE frequencies to multiple covariates and to treatment by covariate interactions, which enables a search for vulnerable subgroups.

One limitation with these methods is that they are based on comparing only the AE incidence between two groups, regardless of the severity of adverse event. It is probably of great interest for us to also compare the severity of the AEs, especially when two treatments have similar incidence for a given AE but the severity of the AE is consistently greater for one treatment versus the other. In this case it would be unappealing for the AE not to be flagged.

To fully capture the presence and severity of adverse events, it is important to incorporate an endpoint that describes the severity of each AE. Klingenberg et al. (2009) proposed a method for investigating the toxicity effect of a chemical compound on animals in an environmental study. They introduced a single primary endpoint to represent the presence and severity of every type of toxicity effect of the chemical compound. Each primary endpoint is an ordinal categorical variable that consists of four levels, representing the severity of none, mild, moderate and severe. They used permutation test and a bootstrap method for testing the simultaneous marginal homogeneity for all the toxicity effect of the chemical compound and adjusted the p values to control for family wise error rate (FWER). The method can be readily carry over to safety analysis in clinical trials. However, power of using one endpoint for each type of adverse effect is low for detecting certain alternatives of interest, for instance, when the AE incidence rate is the same but the severity is different. L.G. Leon-Novelo & Muller (2010) introduced an approach for flexible, model-based inference for the incidence and severity of AEs reported in a phase III clinical trial. The approach

is based on a mixture of normal distributions for latent variables associated with the ordinal AE severity data and it takes into account possible dependence among the outcomes in different AE severity categories. But little is known about the FDR and power properties associated with the method.

In this article, from the Bayesian perspective, we extend the hierarchical model of Berry & Berry (2004) for assessing both the incidence rate and severity of several AEs in a two-arm randomized clinical trial. The primary drug safety outcome is a four level ordinal categorical variable, representing four AE severity levels from none, mild, moderate to severe. The non-proportional odds cumulative logit model is used to model the relationship between the safety outcome and the treatment effect. A three-level hierarchical structure is proposed for the prior distribution of the model parameters. The method adjusts for the multiplicity issue with the help of the hierarchical nature as it borrows information across types of AEs, especially across the adverse events within the same body system. Simulation study is performed to investigate the FDR and power of the method.

The rest of the paper is organized as follows. In Section 3.2, we briefly review a Bayesian hierarchical logit model proposed by Berry & Berry (2004) for assessing the incidence rate of AEs in a two-arm randomized trial. In Section 3.3, we introduce a new Bayesian hierarchical cumulative logit model for simultaneously testing the equivalence of incidence rate and severity of all the AEs between the control group and the treatment group while addressing multiplicities. In Section 3.4, we conduct a comprehensive simulation study to investigate the power and FDR of the proposed Bayesian hierarchical method and compare it with a Bayesian non-hierarchical method and Berry & Berry (2004)'s Bayesian hierarchical logit model. In Section 3.5 we applied the proposed method to a simulated safety dataset. We conclude the article in Section 3.6.

3.2 Bayesian hierarchical logit model: a review of Berry's method

In this section, we briefly review a Bayesian hierarchical logit model proposed by Berry & Berry (2004) for simultaneously testing the equivalence of incidence rate between a control group and

a treatment group for all the AEs in a two-arm randomized trial while addressing multiplicities. AE(s) with significantly higher incidence rate in the treatment group will be flagged. We first introduce the structure of the safety data. We then review the logit model, the hierarchical prior structure of the model parameters and the decision rule for conducting the simultaneous Bayesian test.

3.2.1 Safety data structure

Suppose we were to compare the incidence rates of many AEs between a control group and a treatment group in a two-arm randomized trial. All the AEs are classified into B body systems. Body system is made in advance of and separate from the data from the trial (Berry & Berry, 2004). Within body system b there are k_b types of AEs. Denote the j^{th} AE within the b^{th} body system as A_{bj}, $b = 1,...,B; j = 1,...,k_b$. Of the N_1 subjects in the control group, $x_{1,bj}$ experience A_{bj}. Of the N_2 subjects in the treatment group, $x_{2,bj}$ experience A_{bj}. Table 3.1 shows the layout of of the safety data.

Table 3.1: Safety data structure with only AE incidence information

		Control(N_1)		Treatment(N_2)	
b	j	$x_{1,bj}$	$N_1 - x_{1,bj}$	$x_{2,bj}$	$N_2 - x_{2,bj}$
1	1	$x_{1,11}$	$N_1 - x_{1,11}$	$x_{2,11}$	$N_2 - x_{2,11}$
1	2	$x_{1,12}$	$N_1 - x_{1,12}$	$x_{2,12}$	$N_2 - x_{2,12}$
⋮	⋮	⋮	⋮	⋮	⋮
2	1	$x_{1,21}$	$N_1 - x_{1,21}$	$x_{2,21}$	$N_2 - x_{2,21}$
⋮	⋮	⋮	⋮	⋮	⋮
B	k_B	x_{1,Bk_B}	$N_1 - x_{1,Bk_B}$	x_{2,Bk_B}	$N_2 - x_{2,Bk_B}$

3.2.2 Likelihood functions and priors

Let Y_{bj} be a binary outcome, indicating whether a subject experiences A_{bj}. $Y_{bj} = 1$ if a subject experiences the AE A_{bj} and $Y_{bj} = 0$ if the subject does not experience the AE A_{bj}. Let T be the group indicator of a subject. $T = 1$ if the subject is from the treatment group and $T = 0$ if the

subject is from the control treatment. Denote $c_{bj} = P(Y_{bj} = 1|T = 0), t_{bj} = P(Y_{bj} = 1|T = 1)$. The model Berry & Berry (2004) proposed is equivalent to the following logit model:

$$\log\Big(\frac{P(Y_{bj}=1|T)}{1-P(Y_{bj}=1|T)}\Big) = \gamma_{bj} + \theta_{bj}T, b = 1,...,B; j = 1,...,k_b$$

In fact,

$$\gamma_{bj} = \log\Big(\frac{P(Y_{bj}=1|T=0)}{1-P(Y_{bj}=1|T=0)}\Big) = \log\frac{c_{bj}}{1-c_{bj}}$$

$$\theta_{bj} = \log\Big(\frac{P(Y_{bj}=1|T=1)}{1-P(Y_{bj}=1|T=1)}\Big) - \gamma_{bj} = \log\frac{t_{bj}}{1-t_{bj}} - \gamma_{bj}$$

Our interest focuses on θ_{bj}, which is the log-odds ratio of the incidence rate of A_{bj} between the treatment group and the control group. If $\theta_{bj} = 0$ then the probability that a subject experience A_{bj} is the same in the control and the treatment group. If $\theta_{bj} > 0$ then the probability that a subject experience A_{bj} is greater in the treatment group compare to that in the control group.

From the Bayesian perspective, Berry & Berry (2004) proposed a three-level hierarchical prior structure for the model parameters γ_{bj} and θ_{bj} in order to address multiplicities. Figure 3.1 shows an overview of this hierarchical prior structure.

In the first stage of the prior structure:

$$\gamma_{bj} \sim N(\mu_{\gamma b}, \sigma^2_{\gamma})$$
$$\theta_{bj} \sim \pi_b I_{[0]} + (1-\pi_b)N(\mu_{\theta b}, \sigma^2_{\theta b}), b = 1,...,B; j = 1,...,k_b$$

where the prior distribution for θ_{bj} is a mixture of a point mass at 0 and a normal distribution. Xia et al. (2011) set a different prior distribution for γ_{bj}, which is $N(\mu_{\gamma b}, \sigma^2_{\gamma b})$. We will use this prior

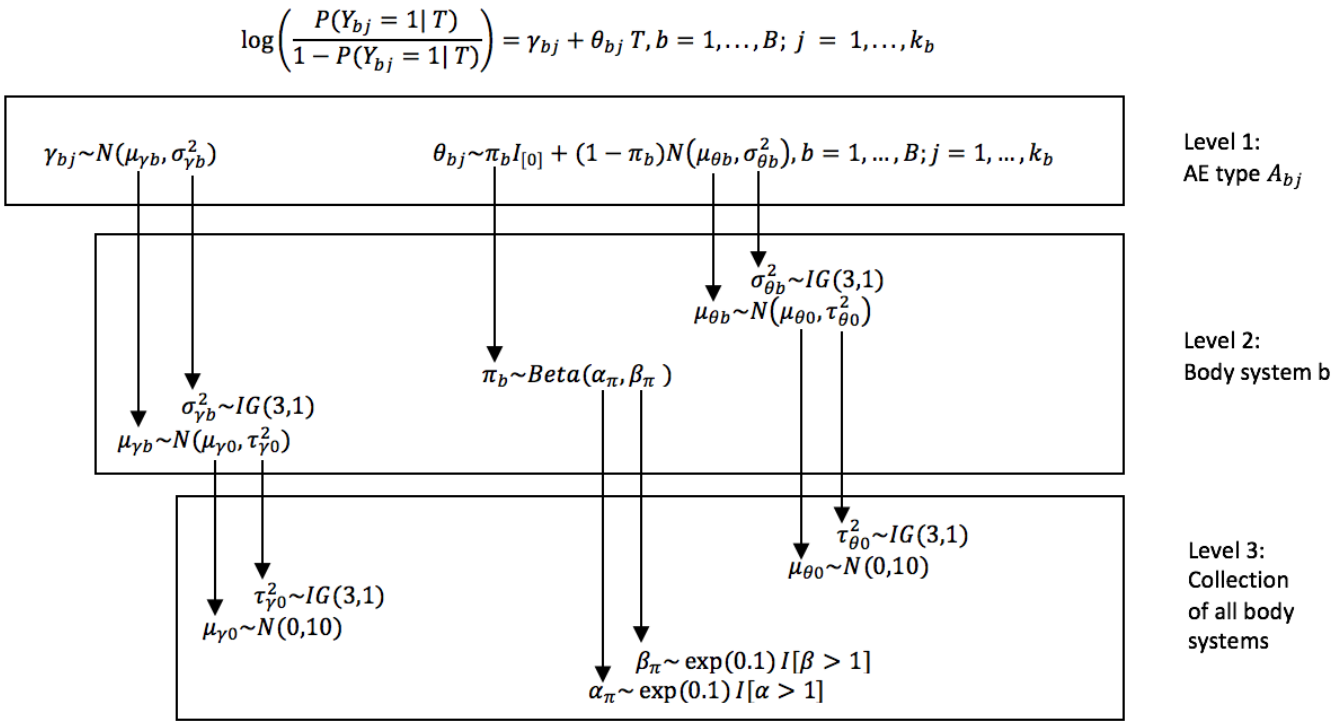

Figure 3.1: Overview of the Bayesian hierarchical prior structure of Berry & Berry (2004)'s model

distribution instead. In the second stage of the prior structure:

$$\mu_{\gamma b}\sim N(\mu_{\gamma 0},\tau^2_{\gamma 0}),\sigma^2_{\gamma b}\sim IG(\alpha_{\sigma\gamma},\beta_{\sigma\gamma})\quad(\alpha_{\sigma\gamma}=3,\beta_{\sigma\gamma}=1)$$
$$\mu_{\theta b}\sim N(\mu_{\theta 0},\tau^2_{\theta 0}),\sigma^2_{\theta b}\sim IG(\alpha_{\sigma\theta},\beta_{\sigma\theta})\quad(\alpha_{\sigma\theta}=3,\beta_{\sigma\theta}=1)$$
$$\pi_b\sim Beta(\alpha_\pi,\beta_\pi)$$

where IG represents the inverse gamma distribution. In the third stage of the prior structure:

$$\mu_{\gamma 0}\sim N(\mu_{\gamma 00},\tau^2_{\gamma 00}),\tau^2_{\gamma 0}\sim IG(\alpha_{\tau\gamma},\beta_{\tau\gamma})\quad(\mu_{\gamma 00}=0,\tau^2_{\gamma 00}=10,\alpha_{\tau\gamma}=3,\beta_{\tau\gamma}=1)$$
$$\mu_{\theta 0}\sim N(\mu_{\theta 00},\tau^2_{\theta 00}),\tau^2_{\theta 0}\sim IG(\alpha_{\theta 0},\beta_{\theta 0})\quad(\mu_{\theta 00}=0,\tau^2_{\theta 00}=10,\alpha_{\theta 0}=3,\beta_{\theta 0}=1)$$
$$\alpha_\pi\sim\frac{\lambda_\alpha\exp(-\alpha\lambda_\alpha)}{\exp(-\lambda_\alpha)}I_{[\alpha>1]},\beta_\pi\sim\frac{\lambda_\beta\exp(-\beta\lambda_\beta)}{\exp(-\lambda_\beta)}I_{[\beta>1]}\quad(\lambda_\alpha=\lambda_\beta=0.1)$$

Independent left-truncated exponential prior distribution is assigned to α_π and β_π because restrict-

ing the parameters to greater than 1 prevents the posterior density of π from becoming too heavily concentrated at one of its edges (Berry & Berry, 2004). This model provides an explicit method for borrowing information across types of AEs. The hierarchical nature of the model gives rise to a regression effect, which is appealing in the context of multiplicities because it modulates extremes (Berry & Berry, 2004). The calculation for the posterior distribution of the model parameters are carried out using Markov Chain Monte Carlo (MCMC) methods. This can be done in WinBUGS or JAGS.

3.2.3 Hypothesis test and decision rule

The goal of simultaneously testing the equivalence of incidence rates of all the AEs between the control and treatment group is equivalent to testing the null hypothesis that $\theta_{bj} = 0$ versus the alternative hypothesis that $\theta_{bj} > 0$ for $A_{bj}, b = 1, ..., B; j = 1, ..., k_b$.

After obtaining the posterior distribution random sample of θ_{bj} from MCMC, we can calculate the posterior probability that $\theta_{bj} > 0$. The decision rule is that if $P(\theta_{bj} > 0) > p_0$, A_{bj} will be flagged, where p_0 is a cutoff point that is usually set to 0.6,0.7 or 0.8. Xia et al. (2011) investigated the performance of using a cutoff point of 0.6, 0.7 and 0.8 and they suggested that 0.7 be used in the model.

3.3 Cumulative logit model for safety signal detection

The Bayesian hierarchical logit model introduced in Section 3.2 compares only the incidence rates of all the AEs between the control and treatment group, regardless of the severity of AEs. In this section, we introduce a new Bayesian hierarchical cumulative logit model for simultaneously testing the equivalence of incidence rate and severity of all the AEs between the control group and the treatment group while addressing multiplicities. AE(s) with greater incidence rate or severity in the treatment group compare to that in the control group will be flagged. Greater severity is defined in terms of stochastic order. The primary drug safety outcome is a four-level ordinal categorical variable, representing four AE severity levels of none, mild, moderate and severe. We first introduce

some basic notations, the hypotheses we will be testing and the structure of the safety data. The safety data consists of the number of subjects with outcomes classified into each of the severity levels in the control group and the treatment group. We then describe a cumulative logit model, the hierarchical prior structure of the model parameters and the decision rule for conducting the simultaneous Bayesian test. A cumulative logit model with non-hierarchical prior (solo Bayesian model) is also introduced for comparison purpose. Note that some notations introduced in the last section continue to be used in this section.

3.3.1 Data structure and notations

It is probably of great interest for us to compare both the incidence rates and severity of many AEs between a control group and a treatment group in a two-arm randomized trial. This motivates the collection of safety data that is more complicated than Table 3.1. Table 3.2 shows the safety data that consists of the number of subjects with outcomes classified into each of the severity levels in the control group and the treatment group. All the AEs are classified into B body systems. Within body system b there are k_b types of AEs. Of the N_1 subjects in the control group, $m_{11,bj}$ does not experience A_{bj}, $m_{12,bj}$ experience A_{bj} and the severity level is mild, $m_{13,bj}$ experience A_{bj} and the severity level is moderate, $m_{14,bj}$ experience A_{bj} and the severity level is severe. Of the N_2 subjects in the treatment group, $m_{21,bj}$ does not experience A_{bj}, $m_{22,bj}$ experience A_{bj} and the severity level is mild, $m_{23,bj}$ experience A_{bj} and the severity level is moderate, $m_{24,bj}$ experience A_{bj} and the severity level is severe.

Let Z_{bj} be a four level severity outcome variable that corresponds to adverse event A_{bj} for any subject. Without loss of generality, we assume that there are four AE severity levels: none, mild, moderate and severe. This can be easily extended to five severity levels, eg, Common Terminology Criteria for Adverse Events (CTCAE) by NCI. Let $Z_{bj} = 1$ if the subject does not experience A_{bj}. $Z_{bj} = 2$ if the subject experience A_{bj} and the severity of the AE is mild. $Z_{bj} = 3$ if the subject experience A_{bj} and the severity of the AE is moderate. $Z_{bj} = 4$ if a subject experience A_{bj} and the severity is severe. Different from that in Section 3.2, we set $T = 0.5$ if the subject is from the

Table 3.2: Safety data structure with AE incidence and AE severity information

		Control(N_1)				Treatment(N_2)			
b	j	$m_{11,bj}$ (No AE)	$m_{12,bj}$ (Mild)	$m_{13,bj}$ (Moderate)	$m_{14,bj}$ (Severe)	$m_{21,bj}$ (No AE)	$m_{22,bj}$ (Mild)	$m_{23,bj}$ (Moderate)	$m_{24,bj}$ (Severe)
1	1	$m_{11,11}$	$m_{12,11}$	$m_{13,11}$	$m_{14,11}$	$m_{21,11}$	$m_{22,11}$	$m_{23,11}$	$m_{24,11}$
1	2	$m_{11,12}$	$m_{12,12}$	$m_{13,12}$	$m_{14,12}$	$m_{21,12}$	$m_{22,12}$	$m_{23,12}$	$m_{24,12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	1	$m_{11,21}$	$m_{12,21}$	$m_{13,21}$	$m_{14,21}$	$m_{21,21}$	$m_{22,21}$	$m_{23,21}$	$m_{24,21}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B	k_B	m_{11,Bk_B}	m_{12,Bk_B}	m_{13,Bk_B}	m_{14,Bk_B}	m_{21,Bk_B}	m_{22,Bk_B}	m_{23,Bk_B}	m_{24,Bk_B}

treatment group and $T = -0.5$ if the subject is from the control treatment. The prior distribution is then symmetric in the sense that the logits in each treatment have the same prior variability as well as the same prior means, yet θ_{bj} still has the usual interpretation of a log odds ratio (Agresti, 2010).

Denote $P(Z_{bj} = 1|T = -0.5) = \phi_{11,bj}, P(Z_{bj} = 2|T = -0.5) = \phi_{12,bj}, P(Z_{bj} = 3|T = -0.5) = \phi_{13,bj}, P(Z_{bj} = 4|T = -0.5) = \phi_{14,bj}$ $P(Z_{bj} = 1|T = 0.5) = \phi_{21,bj}, P(Z_{bj} = 2|T = 0.5) = \phi_{22,bj}, P(Z_{bj} = 3|T = 0.5) = \phi_{23,bj}, P(Z_{bj} = 4|T = 0.5) = \phi_{24,bj}$. We make the following assumptions about the distributions of $\boldsymbol{m}_{1,bj}$ and $\boldsymbol{m}_{2,bj}$.

$$\boldsymbol{m}_{1,bj} = (m_{11,bj}, m_{12,bj}, m_{13,bj}, m_{14,bj}) \sim Multinomial(N_1, (\phi_{11,bj}, \phi_{12,bj}, \phi_{13,bj}, \phi_{14,bj}))$$

$$\boldsymbol{m}_{2,bj} = (m_{21,bj}, m_{22,bj}, m_{23,bj}, m_{24,bj}) \sim Multinomial(N_2, (\phi_{21,bj}, \phi_{22,bj}, \phi_{23,bj}, \phi_{24,bj}))$$

A_{bj} is said to have a safety signal if its incidence rate or the severity is greater in the treatment group compare to that in the control group. This definition of safety signal is an enhancement of that in Berry & Berry (2004) as it takes into consideration not only the increase of incidence rate but also the increase of severity. The severity of A_{bj} is greater in the treatment group compare to that in the control group if $F_c(z) > F_t(z), z = 1,2,3$ (i.e., stochastically greater), or equivalently, if $1 - F_c(z) < 1 - F_t(z), z = 1,2,3$, where $F_c(z)$ and $F_t(z)$ are cumulative density functions of the severity outcome in the control group and the treatment group respectively. Hence, A_{bj} has a safety

signal if its corresponding probabilities ($\phi_{1l,bj}, \phi_{2l,bj}, l = 1,2,3,4$) in each severity level under two groups belong to the set:

$$\{\phi_{14,bj} \leq \phi_{24,bj},$$
$$\phi_{13,bj} + \phi_{14,bj} \leq \phi_{23,bj} + \phi_{24,bj},$$
$$\phi_{12,bj} + \phi_{13,bj} + \phi_{14,bj} \leq \phi_{22,bj} + \phi_{23,bj} + \phi_{24,bj}, \text{ at least one inequality is strict}\}$$
$$\cup \{\phi_{12,bj} + \phi_{13,bj} + \phi_{14,bj} < \phi_{22,bj} + \phi_{23,bj} + \phi_{24,bj}\}$$

Greater severity of an AE is defined in terms of stochastic order because among several formally defined notions, it is the least stringent. See Cohen & Sackrowitz (2000) and Cohen et al. (2000).

3.3.2 Bayesian hierarchical non-proportional odds version of cumulative logit model

We propose a cumulative logit model to quantify the relationship between the ordinal severity outcome Z_{bj} and the treatment. The simplest cumulative logit model is the proportional odds version of the cumulative logit model (Agresti, 2010):

$$\log\Big(\frac{P(Z_{bj} \leq k|T)}{1 - P(Y_{bj} \leq k|T)}\Big) = \eta_{bj,k} - \omega_{bj}T, b = 1,...,B; j = 1,...,k_b, k = 1,2,3$$

Note that

$$\eta_{bj,k} = \log\Big(\frac{P(Z_{bj} \leq k|T = -0.5)}{1 - P(Z_{bj} \leq k|T = -0.5)}\Big), k = 1,2,3$$

$$\omega_{bj} = \log \frac{\frac{P(Z_{bj}\leq 1|T=0.5)}{1-P(Z_{bj}\leq 1|T=0.5)}}{\frac{P(Z_{bj}\leq 1|T=-0.5)}{1-P(Z_{bj}\leq 1|T=-0.5)}} = \log \frac{\frac{P(Z_{bj}\leq 2|T=0.5)}{1-P(Z_{bj}\leq 2|T=0.5)}}{\frac{P(Z_{bj}\leq 2|T=-0.5)}{1-P(Z_{bj}\leq 2|T=-0.5)}} = \log \frac{\frac{P(Z_{bj}\leq 3|T=0.5)}{1-P(Z_{bj}\leq 3|T=0.5)}}{\frac{P(Z_{bj}\leq 3|T=-0.5)}{1-P(Z_{bj}\leq 3|T=-0.5)}}$$

Parameters ω_{bj}, the log odds ratios of cumulative probabilities between the treatment group and the control group, are of main interest. If $\omega_{bj} = 0$ then the incidence rate and the severity of

A_{bj} are the same for the control and treatment group. If $\omega_{bj} > 0$ then the incidence rate and the severity of A_{bj} are both greater in the treatment group compare to those in the control group. It is simple with the same parameter ω_{bj} for each severity level. However, with only one parameter, we cannot identify whether a safety signal is due to the increase in incidence rate or severity. In addition, the proportional odds assumption is unrealistic to hold for all AEs in reality. Thus, we propose to use the following non-proportional odds version of the cumulative logit model (Agresti, 2010) instead:

$$\log\Big(\frac{P(Z_{bj} \leq k|T)}{1-P(Y_{bj} \leq k|T)}\Big) = \eta_{bj,k} - \omega_{bj,k}T, b = 1,...,B; j = 1,...,k_b, k = 1,2,3$$

Compare to the proportional odds version of the cumulative logit model, the parameter that is associated with the treatment covariate is allowed to vary with severity level. Note that

$$\eta_{bj,k} = \log\Big(\frac{P(Z_{bj} \leq k|T=-0.5)}{1-P(Z_{bj} \leq k|T=-0.5)}\Big)$$

$$\begin{aligned}\omega_{bj,k} &= \log\frac{\frac{P(Z_{bj}\leq k|T=0.5)}{1-P(Z_{bj}\leq k|T=0.5)}}{\frac{P(Z_{bj}\leq k|T=-0.5)}{1-P(Z_{bj}\leq k|T=-0.5)}} \\ &= \log\frac{\frac{1-P(Z_{bj}\leq k|T=0.5)}{P(Z_{bj}\leq k|T=0.5)}}{\frac{1-P(Z_{bj}\leq k|T=-0.5)}{P(Z_{bj}\leq k|T=-0.5)}}, k = 1,2,3\end{aligned}$$

Parameters $\omega_{bj,k}, k = 1,2,3$ are the log odds ratios of the probability that the severity of A_{bj} is at least mild, moderate or severe between the treatment group and the control group. If $\omega_{bj,1} = 0$ then the incidence rate of A_{bj} is the same for the control and treatment group. If $\omega_{bj,1} > 0$ then the incidence rate of A_{bj} is greater in the treatment group compare to that in the control group. If $\omega_{bj,2} = 0$ then the probability that a subject experience A_{bj} and the severity level is at least moderate is the same for the control and treatment group. If $\omega_{bj,2} > 0$ then the probability that a subject experience A_{bj} and the severity level is at least moderate is greater in the treatment group

compare to that in the control group. If $\omega_{bj,3} = 0$ then the probability that a subject experience A_{bj} and the severity level is at least severe is the same for the control and treatment group. If $\omega_{bj,3} > 0$ then the probability that a subject experience A_{bj} and the severity level is at least severe is greater in the treatment group compare to that in the control group.

The key challenge for the non-proportional odds version of the cumulative logit model is that in order for stochastic ordering to hold, it is necessary that

$$-\infty < \eta_{bj,1} - \omega_{bj,1}T < \eta_{bj,2} - \omega_{bj,2}T < \eta_{bj,3} - \omega_{bj,3}T < \infty$$

for $T = -0.5$ or 0.5. This can be rewritten as (McKinley et al., 2015):

$$\begin{aligned}
&\eta_{bj,l} - \eta_{bj,l+1} < \omega_{bj,l}T - \omega_{bj,l+1}T, l = 1,2 \\
&\implies \eta_{bj,l} - \eta_{bj,l+1} < \min\{-0.5(\omega_{bj,l} - \omega_{bj,l+1}), 0.5(\omega_{bj,l} - \omega_{bj,l+1})\}, l = 1,2
\end{aligned}$$

Denote $C_{bj,l} = \min\{-0.5(\omega_{bj,l} - \omega_{bj,l+1}), 0.5(\omega_{bj,l} - \omega_{bj,l+1})\}, l = 1,2$, then we have

$$\eta_{bj,l} - \eta_{bj,l+1} < C_{bj,l}, l = 1,2$$

In light of this fact, we can therefore specify prior distributions with three-level hierarchies. Figure 3.2 shows the three-level prior structure of the model parameters.

In the first stage of the prior structure, the joint prior distribution of $\eta_{bj,1}, \eta_{bj,2}, \eta_{bj,3}, \omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}$ is:

$$\begin{aligned}
&f(\eta_{bj,1}, \eta_{bj,2}, \eta_{bj,3}, \omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}) \\
&= f(\eta_{bj,1}, \eta_{bj,2}, \eta_{bj,3} | \omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}) f(\omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}) \\
&= f(\eta_{bj,1}, \eta_{bj,2}, \eta_{bj,3} | \omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}) f(\omega_{bj,1}) f(\omega_{bj,2}) f(\omega_{bj,3}) \\
&= f(\eta_{bj,1}) f(\eta_{bj,2} | \eta_{bj,1}, \omega_{bj,1}, \omega_{bj,2}) f(\eta_{bj,3} | \eta_{bj,2}, \omega_{bj,2}, \omega_{bj,3}) f(\omega_{bj,1}) f(\omega_{bj,2}) f(\omega_{bj,3})
\end{aligned}$$

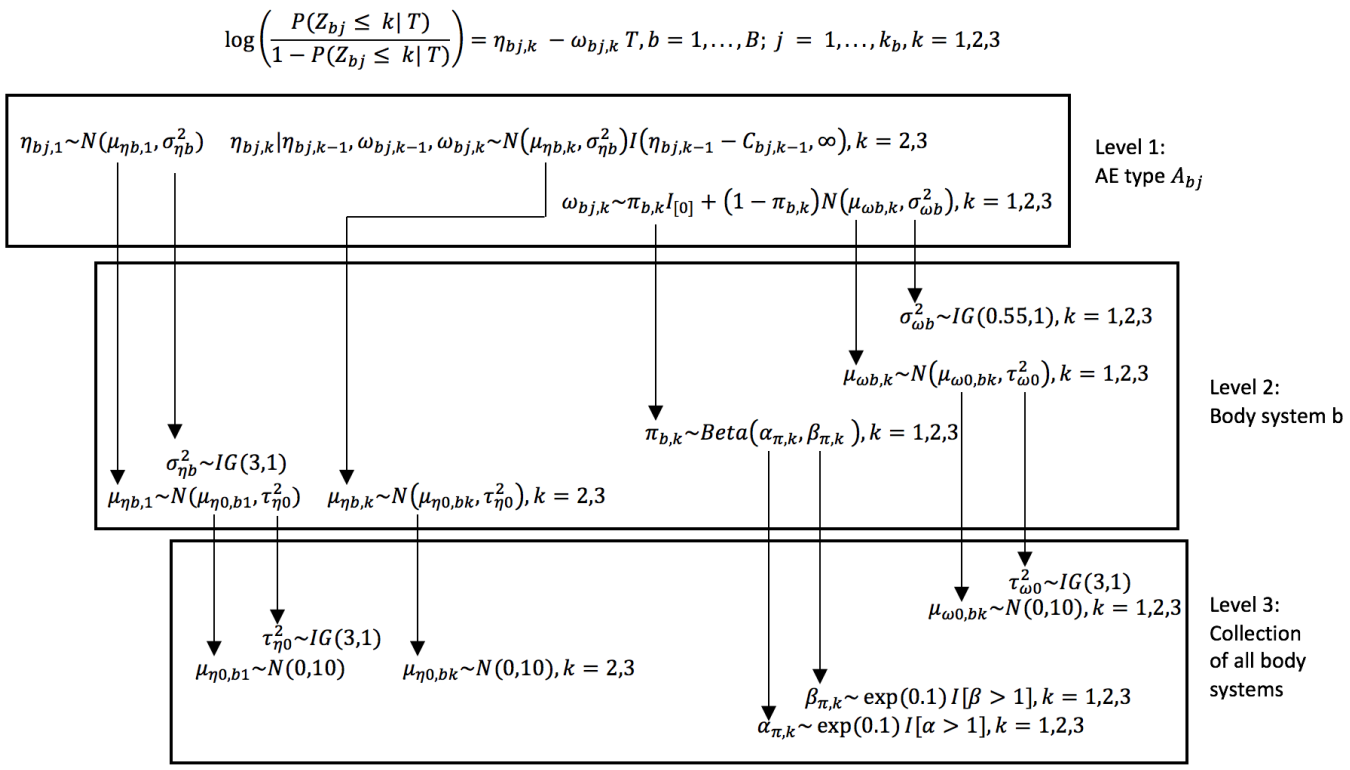

Figure 3.2: Overview of the Bayesian hierarchical prior structure of the non-proportional odds version of the cumulative logit model

where $f(\eta_{bj,1}), f(\eta_{bj,2}|\eta_{bj,1}, \omega_{bj,1}, \omega_{bj,2})$ and $f(\eta_{bj,3}|\eta_{bj,2}, \omega_{bj,2}, \omega_{bj,3})$ are the probability density functions of random variables defined as follows:

$$\eta_{bj,1} \sim N(\mu_{\eta b,1}, \sigma^2_{\eta b})$$

$$\eta_{bj,2}|\eta_{bj,1}, \omega_{bj,1}, \omega_{bj,2} \sim N(\mu_{\eta b,2}, \sigma^2_{\eta b}) I[\eta_{bj,1} - C_{bj,1}, \infty]$$

$$C_{bj,1} = \min\{-0.5(\omega_{bj,1} - \omega_{bj,2}), 0.5(\omega_{bj,1} - \omega_{bj,2})\}$$

$$\eta_{bj,3}|\eta_{bj,2}, \omega_{bj,2}, \omega_{bj,3} \sim N(\mu_{\eta b,3}, \sigma^2_{\eta b}) I[\eta_{bj,2} - C_{bj,2}, \infty]$$

$$C_{bj,2} = \min\{-0.5(\omega_{bj,2} - \omega_{bj,3}), 0.5(\omega_{bj,2} - \omega_{bj,3})\}$$

$f(\eta_{bj,1})$ is a normal distribution defined in $(-\infty, \infty)$. $f(\eta_{bj,2}|\eta_{bj,1}, \omega_{bj,1}, \omega_{bj,2})$ is a truncated normal distribution defined in the range $(\eta_{bj,1} - C_{bj,1}, \infty)$ and $f(\eta_{bj,3}|\eta_{bj,2}, \omega_{bj,2}, \omega_{bj,3})$ is a truncated normal distribution defined in the range $(\eta_{bj,2} - C_{bj,2}, \infty)$. $f(\omega_{bj,1}), f(\omega_{bj,2}), f(\omega_{bj,3})$ are

the probability density functions of random variables $\omega_{bj,1}, \omega_{bj,2}, \omega_{bj,3}$. Each of them is defined as a mixture of a distribution having unit point mass at 0 and a normal distribution:

$$\omega_{bj,1} \sim \pi_{b,1} I_{[0]} + (1-\pi_{b,1}) N(\mu_{\omega b,1}, \sigma^2_{\omega b})$$
$$\omega_{bj,2} \sim \pi_{b,2} I_{[0]} + (1-\pi_{b,2}) N(\mu_{\omega b,2}, \sigma^2_{\omega b})$$
$$\omega_{bj,3} \sim \pi_{b,3} I_{[0]} + (1-\pi_{b,3}) N(\mu_{\omega b,3}, \sigma^2_{\omega b})$$

In the second stage of the prior structure:

$$\mu_{\eta b,1} \sim N(\mu_{\eta 0,b1}, \tau^2_{\eta 0}), \mu_{\eta b,2} \sim N(\mu_{\eta 0,b2}, \tau^2_{\eta 0}), \mu_{\eta b,3} \sim N(\mu_{\eta 0,b3}, \tau^2_{\eta 0})$$
$$\mu_{\omega b,1} \sim N(\mu_{\omega 0,b1}, \tau^2_{\omega 0}), \mu_{\omega b,2} \sim N(\mu_{\omega 0,b2}, \tau^2_{\omega 0}), \mu_{\omega b,3} \sim N(\mu_{\omega 0,b3}, \tau^2_{\omega 0})$$
$$\sigma^2_{\eta b} \sim IG(\alpha_{\sigma\eta}, \beta_{\sigma\eta}), \sigma^2_{\omega b} \sim IG(\alpha_{\sigma\omega}, \beta_{\sigma\omega}) (\alpha_{\sigma\eta} = 3, \beta_{\sigma\eta} = 1, \alpha_{\sigma\omega} = 0.55, \beta_{\sigma\omega} = 1)$$
$$\pi_{b,1} \sim Beta(\alpha_{\pi,1}, \beta_{\pi,1}), \pi_{b,2} \sim Beta(\alpha_{\pi,2}, \beta_{\pi,2}), \pi_{b,3} \sim Beta(\alpha_{\pi,3}, \beta_{\pi,3})$$

Specifically, the hyper parameters of the prior distribution for $\sigma^2_{\omega b}$ are set to 0.55 and 1 after doing experiments on some values. Other hyper parameter values may be (0.1,1). Parameter values like (0.001,1) may be too conservative that will lead to low power. In the third stage of the prior

structure:

$$\mu_{\eta 0,b1} \sim N(\mu_{\eta 00,1}, \tau^2_{\eta 00}) \quad (\mu_{\eta 00,1} = 0, \tau^2_{\eta 00} = 10)$$

$$\mu_{\eta 0,b2} \sim N(\mu_{\eta 00,2}, \tau^2_{\eta 00}) \quad (\mu_{\eta 00,2} = 0, \tau^2_{\eta 00} = 10)$$

$$\mu_{\eta 0,b3} \sim N(\mu_{\eta 00,3}, \tau^2_{\eta 00}) \quad (\mu_{\eta 00,3} = 0, \tau^2_{\eta 00} = 10)$$

$$\mu_{\omega 0,b1} \sim N(\mu_{\omega 00,1}, \tau^2_{\omega 00}) \quad (\mu_{\omega 00,1} = 0, \tau^2_{\omega 00} = 10)$$

$$\mu_{\omega 0,b2} \sim N(\mu_{\omega 00,2}, \tau^2_{\omega 00}) \quad (\mu_{\omega 00,2} = 0, \tau^2_{\omega 00} = 10)$$

$$\mu_{\omega 0,b3} \sim N(\mu_{\omega 00,3}, \tau^2_{\omega 00}) \quad (\mu_{\omega 00,2} = 0, \tau^2_{\omega 00} = 10)$$

$$\tau^2_{\eta 0} \sim IG(\alpha_{\tau\eta}, \beta_{\tau\eta}), \tau^2_{\omega 0} \sim IG(\alpha_{\omega 0}, \beta_{\omega 0}) \quad (\alpha_{\tau\eta} = 3, \beta_{\tau\eta} = 1, \alpha_{\omega 0} = 3, \beta_{\omega 0} = 1)$$

$$\alpha_{\pi,1} \sim \frac{\lambda_\alpha \exp(-\alpha\lambda_\alpha)}{\exp(-\lambda_\alpha)} I_{[\alpha>1]}, \beta_{\pi,1} \sim \frac{\lambda_\beta \exp(-\beta\lambda_\beta)}{\exp(-\lambda_\beta)} I_{[\beta>1]} \quad (\lambda_\alpha = \lambda_\beta = 0.1)$$

$$\alpha_{\pi,2} \sim \frac{\lambda_\alpha \exp(-\alpha\lambda_\alpha)}{\exp(-\lambda_\alpha)} I_{[\alpha>1]}, \beta_{\pi,2} \sim \frac{\lambda_\beta \exp(-\beta\lambda_\beta)}{\exp(-\lambda_\beta)} I_{[\beta>1]}$$

$$\alpha_{\pi,3} \sim \frac{\lambda_\alpha \exp(-\alpha\lambda_\alpha)}{\exp(-\lambda_\alpha)} I_{[\alpha>1]}, \beta_{\pi,2} \sim \frac{\lambda_\beta \exp(-\beta\lambda_\beta)}{\exp(-\lambda_\beta)} I_{[\beta>1]}$$

3.3.3 Solo Bayesian non-proportional odds cumulative logit model

To show how the three-stage hierarchical prior structure affects the conclusions, we further consider a Solo Bayesian model. This is simply a Bayesian model with priors being set as the first stage of the three-stage hierarchical prior structure introduced above (Figure 3.3). The parameters of the distribution in the first stage of the prior structure are considered fixed. No information is borrowed across different AEs within the same body system and the individual types of AEs are considered in isolation. With vague prior information, hypothesis test based on this model should have results similar to the frequentist approach without multiplicity adjustment.The parameters of the distribution in the first stage of prior structure are set as: $\pi_{b,1} = \pi_{b,2} = \pi_{b,3} = 0.5$, $\mu_{\omega b,1} = \mu_{\omega b,2} = \mu_{\omega b,3} = 0, \sigma^2_{\omega b} = 1$ $\mu_{\eta b,1} = \mu_{\eta b,2} = \mu_{\eta b,3} = -1, \sigma^2_{\eta b} = 10$.

$$\log\left(\frac{P(Z_{bj} \leq k|\, T)}{1-P(Z_{bj} \leq k|\, T)}\right) = \eta_{bj,k} - \omega_{bj,k}\, T, b = 1,\ldots,B;\ j = 1,\ldots,k_b, k = 1,2,3$$

$$\eta_{bj,1} \sim N(-1,10) \quad \eta_{bj,k}|\eta_{bj,k-1}, \omega_{bj,k-1}, \omega_{bj,k} \sim N(-1,10)I(\eta_{bj,k-1} - C_{bj,k-1}, \infty), k = 2,3$$

$$\omega_{bj,k} \sim 0.5I_{[0]} + 0.5N(0,1), k = 1,2,3$$

Figure 3.3: Priors of the solo Bayesian non-proportional odds version of the cumulative logit model

3.3.4 Hypothesis test and decision rule

Our goal of simultaneously comparing the equivalence of incidence rate and severity of all the AEs between the control group and the treatment group is hence equivalent to testing the following hypotheses:

$$H_0^{(bj)} : \omega_{bj,1} = \omega_{bj,2} = \omega_{bj,3} = 0$$

$$H_1^{(bj)} : \omega_{bj,1} \geq 0, \omega_{bj,2} \geq 0, \omega_{bj,3} \geq 0, \text{at least one inequality is strict} \qquad \text{OR} \quad \omega_{bj,1} > 0$$

From Bayesian perspective, the decision rule is based on the posterior probability of the alternative hypothesis, i.e., $P(H_1^{(bj)})$. The calculations for the posterior distribution of the parameters are carried out using Markov Chain Monte Carlo (MCMC) methods in JAGS. We simulate 10000 observations from the posterior distribution after a burn-in of 2000 observations. From the MCMC posterior sample, we can calculate the posterior probability of $H_1^{(bj)}$. The decision rule is that if $P(H_1^{(bj)}) > p_0$, A_{bj} will be flagged, where p_0 is a cutoff point that is set to 0.6,0.7 or 0.8. We will investigate how the choice of p_0 affect the decision in the next section.

3.4 Simulation study

In this section we describe a simulation study that evaluates the operating characteristics (FDR and power) of the Bayesian hierarchical cumulative logit model and the solo Bayesian cumulative logit model under different scenarios. Comparisons are also made between the two cumulative logit

models and Berry & Berry (2004)'s Bayesian hierarchical logit model.

3.4.1 Simulation setup

The goal of this simulation study is to investigate the FDR and power of the above methods in detecting safety signals in a two-arm phase III randomized clinical trial under 7 data generation scenarios. Suppose there are 800 subjects in the control group and 800 subjects in the treatment group. We consider 26 types of AEs distributed across 3 body systems, with 10,6,10 AE types in body system 1 through 3.

Power and FDR are estimated by simulation. The estimation accuracy is 500 simulated trials. In each simulated trial, for the control group, we generate the number of subjects who experience the AE $A_{bj}, b = 1,...,B; j = 1,...,k_b$ with severity being classified into different severity levels (i.e., $m_{11,bj}, m_{12,bj}, m_{13,bj}, m_{14,bj}$) from the multinomial distribution with probability vector $(\phi_{11,bj}, \phi_{12,bj}, \phi_{13,bj}, \phi_{14,bj})$. Similarly, for the treatment group, we generate $m_{21,bj}, m_{22,bj}, m_{23,bj}, m_{24,bj}$ from the multinomial distribution with probability vector $(\phi_{21,bj}, \phi_{22,bj}, \phi_{23,bj}, \phi_{24,bj})$.

We consider 7 scenarios of data generation. Probability vector $(\phi_{11,bj}, \phi_{12,bj}, \phi_{13,bj}, \phi_{14,bj})$ is the same in all scenarios and is given in Table 3.3. Probability vector $(\phi_{21,bj}, \phi_{22,bj}, \phi_{23,bj}, \phi_{24,bj})$ varies among different scenarios. Details are given as follows.

Scenario 1 Global null-no true signal in any body system and probability vector for treatment group $(\phi_{21,bj}, \phi_{22,bj}, \phi_{23,bj}, \phi_{24,bj})$ is the same as that for control group.

Scenario 2 We signal within specific AE and within the first two severity levels. The first two AEs in body system 1 and 3 and the first AE in body system 2 in the treatment group have true signal. $\phi_{21,11} = \phi_{21,12} = 0.85, \phi_{22,11} = \phi_{22,12} = 0.13$, $\phi_{21,21} = 0.75, \phi_{22,21} = 0.21$, $\phi_{21,31} = \phi_{21,32} = 0.65, \phi_{22,31} = \phi_{22,32} = 0.29$. All other probabilities remain to be the same as the control group.

Scenario 3 We signal within specific body system. All AEs in body system 1 have true signal. $\phi_{21,11} = ... = \phi_{21,1,10} = 0.85, \phi_{22,11} = ... = \phi_{22,1,10} = 0.13$. All other probabilities

Table 3.3: Probability vectors of AEs in the control group for simulation study

		Control($N_1 = 800$)			
b	j	$\phi_{11,bj}$ (No AE)	$\phi_{12,bj}$ (Mild)	$\phi_{13,bj}$ (Moderate)	$\phi_{14,bj}$ (Severe)
1	1	0.95	0.03	0.015	0.005
1	2	0.95	0.03	0.015	0.005
1	3	0.95	0.03	0.015	0.005
1	4	0.95	0.03	0.015	0.005
1	5	0.95	0.03	0.015	0.005
1	6	0.95	0.03	0.015	0.005
1	7	0.95	0.03	0.015	0.005
1	8	0.95	0.03	0.015	0.005
1	9	0.95	0.03	0.015	0.005
1	10	0.95	0.03	0.015	0.005
2	1	0.9	0.06	0.03	0.01
2	2	0.9	0.06	0.03	0.01
2	3	0.9	0.06	0.03	0.01
2	4	0.9	0.06	0.03	0.01
2	5	0.9	0.06	0.03	0.01
2	6	0.9	0.06	0.03	0.01
3	1	0.85	0.09	0.045	0.015
3	2	0.85	0.09	0.045	0.015
3	3	0.85	0.09	0.045	0.015
3	4	0.85	0.09	0.045	0.015
3	5	0.85	0.09	0.045	0.015
3	6	0.85	0.09	0.045	0.015
3	7	0.85	0.09	0.045	0.015
3	8	0.85	0.09	0.045	0.015
3	9	0.85	0.09	0.045	0.015
3	10	0.85	0.09	0.045	0.015

remain to be the same as the control group.

Scenario 4 We signal within specific AE and within the first two severity levels. The first two AEs in body system 1 and 3 and the first AE in body system 2 in the treatment group have true signal. $\phi_{22,11} = \phi_{22,12} = 0.01, \phi_{23,11} = \phi_{23,12} = 0.035$, $\phi_{22,21} = 0.02, \phi_{23,21} = 0.07$, $\phi_{22,31} = \phi_{22,32} = 0.04, \phi_{23,31} = \phi_{23,32} = 0.095$. All other probabilities remain to be the same as the control group.

Scenario 5 We signal within specific body system. All AEs in body system 1 have true signal. $\phi_{22,11} = ... = \phi_{22,1,10} = 0.01, \phi_{23,11} = ... = \phi_{23,1,10} = 0.035$. All other probabilities

remain to be the same as the control group.

Scenario 6 We signal within specific AE and within the first two severity levels. The first two AEs in body system 1 and 3 and the first AE in body system 2 in the treatment group have true signal. $\phi_{23,11} = \phi_{23,12} = 0.005, \phi_{24,11} = \phi_{24,12} = 0.015$, $\phi_{23,21} = 0.01, \phi_{24,21} = 0.03$, $\phi_{23,31} = \phi_{23,32} = 0.015, \phi_{24,31} = \phi_{24,32} = 0.045$. All other probabilities remain to be the same as the control group.

Scenario 7 We signal within specific body system. All AEs in body system 1 have true signal. $\phi_{23,11} = ... = \phi_{23,1,10} = 0.005, \phi_{24,11} = ... = \phi_{24,1,10} = 0.015$. All other probabilities remain to be the same as the control group.

As introduced in the last section, the decision rule for comparing the incidence rate and severity between two groups for A_{bj} is based on the posterior probabilities. If cumulative logit model is used, A_{bj} will be flagged if $P(H_1^{(bj)}) > p_0$. If logit model is used, A_{bj} will be flagged if $P(\theta_{bj} > 0) > p_0$. p_0 is a cutoff point that is set to 0.6,0.7 or 0.8 respectively.

For each simulated trial, we calculate the following posterior probabilities for each AE under Solo Bayesian cumulative logit model and Bayesian hierarchical cumulative logit model respectively: the posterior probability of $H_1^{(bj)}$ and the posterior probability that $\omega_{bj,k} \geq 0, k = 1,2,3$. Hence for each simulated trial we will obtain a Table 3.4 of posterior probabilities. In addition, we calculate the posterior probabilities $P(\theta_{bj} > 0)$ for each AE under Bayesian hierarchical logit model proposed by Berry & Berry (2004).

We estimate the FDR using the average value of F/S, where S is the number of AE types that are flagged in a simulated trial and F is the corresponding number of AE types that are incorrectly flagged. Power is estimated with the average value of C/T, where C is the number of AE types that are correctly flagged in a simulated trial and T is the corresponding number of AE types with underlying true signals. If no AE types are flagged in a simulated trial, the FDR is defined as 0. Under scenario 1, the global null, where there is no true signal, power is not defined. False discovery rate (FDR) and power are then estimated. Specifically for scenario 1, the global null,

Table 3.4: The posterior probability that $H_1^{(bj)}$ is true and the posterior probability that $\omega_{bj,k} > 0, k = 1,2,3$ for each adverse event A_{bj}

b	j	$P(H_1^{(bj)}\|\text{Data})$	$P(\omega_{bj,1} > 0\|\text{Data})$	$P(\omega_{bj,2} > 0\|\text{Data})$	$P(\omega_{bj,3} > 0\|\text{Data})$
1	1	$P(H_1^{(11)}\|\text{Data})$	$P(\omega_{11,1} > 0\|\text{Data})$	$P(\omega_{11,2} > 0\|\text{Data})$	$P(\omega_{11,3} > 0\|\text{Data})$
⋮	⋮	⋮	⋮	⋮	⋮
1	10	$P(H_1^{(1,10)}\|\text{Data})$	$P(\omega_{1,10,1} > 0\|\text{Data})$	$P(\omega_{1,10,2} > 0\|\text{Data})$	$P(\omega_{1,10,3} > 0\|\text{Data})$
2	1	$P(H_1^{(21)}\|\text{Data})$	$P(\omega_{21,1} > 0\|\text{Data})$	$P(\omega_{21,2} > 0\|\text{Data})$	$P(\omega_{21,3} > 0\|\text{Data})$
⋮	⋮	⋮	⋮	⋮	⋮
2	6	$P(H_1^{(26)}\|\text{Data})$	$P(\omega_{26,1} > 0\|\text{Data})$	$P(\omega_{26,2} > 0\|\text{Data})$	$P(\omega_{26,3} > 0\|\text{Data})$
3	1	$P(H_1^{(31)}\|\text{Data})$	$P(\omega_{31,1} > 0\|\text{Data})$	$P(\omega_{31,2} > 0\|\text{Data})$	$P(\omega_{31,3} > 0\|\text{Data})$
⋮	⋮	⋮	⋮	⋮	⋮
3	10	$P(H_1^{(3,10)}\|\text{Data})$	$P(\omega_{3,10,1} > 0\|\text{Data})$	$P(\omega_{3,10,2} > 0\|\text{Data})$	$P(\omega_{3,10,3} > 0\|\text{Data})$

family wise error rate (FWER) is equal to FDR.

3.4.2 Simulation results

In Table 3.5, we summarize the results of the simulation study by presenting the FDR of the methods based on Solo Bayesian cumulative logit model, Bayesian hierarchical cumulative logit model and Bayesian hierarchical logit model under the global null scenario where there is no true safety signal. Both the Bayesian hierarchical cumulative logit model and Bayesian hierarchical logit model outperform the Solo Bayesian model in the sense that they have lower FDR.

Table 3.5: FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenario 1, the global null

		Cumulative logit model (Solo Bayesian)	Cumulative logit model (Bayesian hierarchical)	Logit model (Bayesian hierarchical)
Scenario	p_0	FDR(FWER)	FDR(FWER)	FDR(FWER)
1	0.6	0.713	0.020	0.018
1	0.7	0.327	0.007	0.01
1	0.8	0.113	0.000	0.001

Table 3.6 shows the power and FDRs of the three methods under scenario 2 and 3 where the incidence rates of some of the AEs are greater in the treatment group but the severities are the

same for all AEs in both groups. As we expect, the FDRs of the methods based on Bayesian hierarchical cumulative logit model and Bayesian hierarchical logit model are controlled at a low level. However, the FDR of the method based on Solo Bayesian cumulative logit model is greater, especially when there are more AEs with true safety signal (scenario 2) and/or the cut-off value p_0 is set to 0.6. The power of the methods based on two hierarchical models is greater, showing their ability to detect true safety signal. On the other hand, the power of the method based on Solo Bayesian cumulative logit model for detecting the true safety signal is comparatively low. This is because the decision rule is based on the posterior probability that $H_0^{(bj)}$ is true, which is affected not only by $P(\omega_{bj,k} > 0)$ $(k = 1,2,3)$ but also by $P(\omega_{bj,k} < 0)$ $(k = 1,2,3)$. If any of the $P(\omega_{bj,k} < 0)$ is large due to the randomness of sampling, $P(H_0^{(bj)})$ will decrease to some extent, leading to low power of rejecting the null hypothesis.

Table 3.6: Powers and FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenarios 2 to 3

		Cumulative logit model (Solo Bayesian)		Logit model (Bayesian hierarchical)		Cumulative logit model (Bayesian hierarchical)	
Scenario	p_0	FDR	Power	FDR	Power	FDR	Power
2	0.6	0.179	0.749	0.016	1	0.015	0.996
2	0.7	0.083	0.535	0.008	1	0.009	0.992
2	0.8	0.027	0.263	0.002	1	0.006	0.989
3	0.6	0.087	0.641	0.006	1	0.002	0.993
3	0.7	0.053	0.408	0.005	1	0.002	0.987
3	0.8	0.033	0.136	0.003	1	0.001	0.977

Table 3.7 shows the power and FDRs of the three methods under scenario 4 to 7 where the incidence rates of all the AEs are the same in both groups but the severities of some of the AEs are greater in the treatment group. As we expect, the FDRs of the methods based on Bayesian hierarchical cumulative logit model and Bayesian hierarchical logit model are controlled at a low level while the FDR of the method based on Solo Bayesian cumulative logit model is greater, especially when there are more AEs with true safety signal (scenario 4, 6) and/or the cut-off value p_0 is set to 0.6. The power of the proposed method based on the Bayesian hierarchical cumulative logit model is greater, showing its ability to detect true safety signal with respect to AE severity.

However, the power of the method based on the Bayesian hierarchical logit model is very low, which is not surprising as the method is developed for detecting the inequivalence with respect to AE incidence rate only. On the other hand, the power of the method based on Solo Bayesian cumulative logit model for detecting the true safety signal is comparatively low.

Table 3.7: Powers and FDRs of the two Bayesian cumulative logit models and the Bayesian logit model under scenarios 4 to 7

		Cumulative logit model (Solo Bayesian)		Logit model (Bayesian hierarchical)		Cumulative logit model (Bayesian hierarchical)	
Scenario	p_0	FDR	Power	FDR	Power	FDR	Power
4	0.6	0.192	0.663	0.013	0	0.031	0.917
4	0.7	0.118	0.501	0	0	0.020	0.891
4	0.8	0.045	0.270	0	0	0.015	0.861
5	0.6	0.109	0.540	0.007	0	0.004	0.986
5	0.7	0.084	0.327	0.003	0	0.002	0.977
5	0.8	0.051	0.120	0	0	0.001	0.953
6	0.6	0.164	0.727	0.003	0	0.055	0.872
6	0.7	0.071	0.598	0	0	0.032	0.834
6	0.8	0.020	0.394	0	0	0.016	0.794
7	0.6	0.101	0.538	0.003	0	0.010	0.969
7	0.7	0.051	0.354	0.003	0	0.008	0.939
7	0.8	0.032	0.143	0.003	0	0.006	0.887

3.5 Application

We use a simulated dataset to validate the proposed method. The generation of the simulated dataset is based on a safety dataset collected from a vaccine trial involved a quadrivalent vaccine containing measles, mumps, rubella and varicella (MMRV)(Mehrotra & Heyse, 2004). Participants were 296 healthy toddlers ages 12-18 months who were randomly assigned to two groups. Group 1 received MMRV on day 0 and group 2 received MMR on day 0 followed by V on day 42. Safety follow-up used standard AE reporting. The primary purpose is to compare the AE incidence rates between Group 1, days 0 to 42, and Group 2, days 42 to 84 in order to compare the safety profile of MMRV to that of V alone for the varicella component.

To generate the simulated dataset, we assume that the true control incidence rates of AEs are

the same as the observed proportion of control events (except that for those with 0 events, we assume a proportion of 1/132). The ratios of the probability that the AE severity is mild, moderate and severe is 6:3:1.

We signal within specific AEs: A_{53} has true safety signal with respect to the first and second severity levels, the probability vector of the AE in the treatment group is $(0.49, 0.3796, 0.0978, 0.0326)$. A_{68} has true safety signal with respect to the second and third severity levels, the probability vector of the AE in the treatment group is $(0.9, 0.03, 0.06, 0.01)$. A_{15} have true safety signal with respect to the third and fourth severity levels, the probability vector of the AE in the treatment group is $(0.85, 0.09, 0.015, 0.045)$. All other AEs have same probability vectors in the control group.

A total of 1000 subjects were assigned into two groups of equal size. The simulated dataset shows in Table 3.8, which consists of the number of subjects who experience A_{bj} with AE severity being classified into different severity level in the control and treatment groups respectively.

We apply the proposed Bayesian hierarchical non-proportional odds version of the cumulative logit model to analyze the simulated dataset. For comparison purpose, we also apply the Solo Bayesian model. Table 3.9 gives the posterior probability that $H_1^{(bj)}$ is true and the posterior probabilities that $\omega_{bj,1} > 0, \omega_{bj,2} > 0, \omega_{bj,3} > 0$, respectively under Bayesian hierarchical cumulative logit model and Solo Bayesian cumulative logit model.

We are specifically interested in the three AEs with true safety signal: A_{53}, A_{68}, A_{15}. Based on the Bayesian hierarchical model, we obtain the posterior probability that $H_1^{(bj)}$ is true for A_{53}, A_{68}, A_{15}. $P(H_1^{(53)}) = 0.893, P(H_1^{(68)}) = 0.757, P(H_1^{(15)}) = 0.988$. Based on the Solo Bayesian model, $P(H_1^{(53)}) = 0.551, P(H_1^{(68)}) = 0.905, P(H_1^{(15)}) = 0.816$. The reason why $P(H_1^{(68)})$ is smaller under Bayesian hierarchical model compare to Solo Bayesian model is because A_{68} is part of the largest body system (6), one in which there is not consistent evidence of a safety signal with respect to the second and third severity levels, thus the estimated treatment effect for A_{68} is "shrunk" toward 0. However, despite this shrinkage, the estimated $P(H_1^{(68)})$ is moderately large. Therefore the difference is likely due to a treatment effect.

Also note that the method based on solo Bayesian model cannot detect the safety signal that is

Table 3.8: Simulated safety dataset for data analysis

		Control ($N_1 = 500$)				Treatment ($N_2 = 500$)			
b	j	$m_{11,bj}$	$m_{12,bj}$	$m_{13,bj}$	$m_{14,bj}$	$m_{21,bj}$	$m_{22,bj}$	$m_{23,bj}$	$m_{24,bj}$
1	1	355	93	39	13	348	91	49	12
1	2	391	64	33	12	392	66	33	9
1	3	496	3	1	0	498	2	0	0
1	4	497	2	1	0	498	2	0	0
1	5	420	46	30	4	424	45	12	19
2	1	494	3	2	1	490	5	3	2
2	2	493	5	2	0	493	5	1	1
2	3	498	0	2	0	496	0	3	1
2	4	451	28	16	5	457	28	13	2
2	5	498	2	0	0	498	2	0	0
2	6	477	15	7	1	470	18	9	3
2	7	421	53	21	5	445	30	19	6
3	1	492	4	1	3	492	5	2	1
4	1	490	8	1	1	496	1	2	1
5	1	494	3	2	1	496	1	3	0
5	2	494	4	2	0	496	2	2	0
5	3	335	100	53	12	249	199	34	18
6	1	491	7	1	1	497	2	0	1
6	2	496	3	1	0	497	2	1	0
6	3	496	1	2	1	494	4	2	0
6	4	468	19	8	5	468	19	11	2
6	5	432	46	17	5	420	45	25	10
6	6	498	1	1	0	496	3	1	0
6	7	476	13	8	3	464	23	10	3
6	8	456	26	15	3	440	14	37	9
6	9	494	4	1	1	494	4	1	1
6	10	499	1	0	0	497	1	1	1
6	11	494	5	0	1	498	2	0	0
7	1	495	1	2	2	493	5	2	0
7	2	497	0	1	2	498	1	1	0
7	3	497	2	1	0	499	0	1	0
7	4	490	6	3	1	490	4	4	2
7	5	494	3	3	0	490	5	4	1
7	6	496	2	2	0	494	6	0	0
7	7	494	4	2	0	493	5	2	0
7	8	496	1	2	1	491	7	1	1
7	9	488	7	3	2	492	5	3	0
8	1	494	3	2	1	495	5	0	0
8	2	444	33	15	8	453	22	17	8
8	3	494	3	0	3	496	4	0	0

Table 3.9: The posterior probability that $H_1^{(bj)}$ is true and the posterior probability that $\omega_{bj,k} > 0, k = 1,2,3$ for A_{bj} based on two Bayesian cumulative logit models

		Solo Bayesian cumulative logit model				Bayesian hierarchical cumulative logit model			
b	j	$P(H_1^{(bj)})$	$P(\omega_{bj,1} > 0)$	$P(\omega_{bj,2} > 0)$	$P(\omega_{bj,3} > 0)$	$P(H_1^{(bj)})$	$P(\omega_{bj,1} > 0)$	$P(\omega_{bj,2} > 0)$	$P(\omega_{bj,3} > 0)$
1	1	0.216	0.058	0.174	0.075	0.036	0.003	0.014	0.02
1	2	0.158	0.06	0.068	0.086	0.029	0.003	0.005	0.022
1	3	0.228	0.138	0.157	0.304	0.288	0.007	0.025	0.345
1	4	0.236	0.186	0.144	0.299	0.272	0.012	0.02	0.331
1	5	0.816	0.053	0.056	0.997	0.988	0.002	0.004	0.996
2	1	0.383	0.255	0.187	0.249	0.115	0.016	0.02	0.09
2	2	0.356	0.154	0.144	0.4	0.313	0.012	0.023	0.306
2	3	0.383	0.128	0.167	0.337	0.23	0.006	0.009	0.221
2	4	0.151	0.072	0.069	0.108	0.04	0.004	0.008	0.029
2	5	0.298	0.217	0.269	0.25	0.254	0.023	0.131	0.175
2	6	0.429	0.168	0.173	0.344	0.152	0.01	0.016	0.133
2	7	0.078	0.005	0.13	0.236	0.073	0	0.027	0.059
3	1	0.232	0.189	0.183	0.089	0.069	0.014	0.047	0.018
4	1	0.141	0.014	0.41	0.204	0.544	0.002	0.612	0.056
5	1	0.256	0.089	0.329	0.157	0.106	0.013	0.081	0.044
5	2	0.28	0.095	0.274	0.266	0.175	0.016	0.068	0.13
5	3	0.551	1	0.02	0.514	0.893	1	0.002	0.194
6	1	0.134	0.037	0.152	0.285	0.13	0.002	0.071	0.095
6	2	0.29	0.127	0.272	0.255	0.215	0.01	0.137	0.121
6	3	0.292	0.407	0.13	0.165	0.108	0.046	0.04	0.055
6	4	0.173	0.087	0.167	0.069	0.061	0.005	0.049	0.019
6	5	0.437	0.087	0.291	0.279	0.123	0.003	0.067	0.066
6	6	0.344	0.368	0.185	0.272	0.163	0.038	0.059	0.112
6	7	0.418	0.446	0.126	0.179	0.094	0.041	0.023	0.047
6	8	0.905	0.26	0.994	0.44	0.757	0.015	0.765	0.122
6	9	0.319	0.162	0.191	0.236	0.129	0.011	0.07	0.069
6	10	0.412	0.205	0.359	0.27	0.241	0.016	0.174	0.104
6	11	0.177	0.076	0.251	0.153	0.144	0.006	0.127	0.048
7	1	0.272	0.498	0.116	0.1	0.063	0.057	0.01	0.018
7	2	0.215	0.259	0.123	0.108	0.038	0.017	0.015	0.018
7	3	0.242	0.076	0.327	0.254	0.15	0.007	0.104	0.068
7	4	0.356	0.084	0.265	0.271	0.093	0.006	0.032	0.063
7	5	0.415	0.252	0.176	0.323	0.1	0.014	0.02	0.073
7	6	0.191	0.405	0.05	0.316	0.076	0.04	0.006	0.097
7	7	0.318	0.194	0.201	0.272	0.09	0.011	0.026	0.064
7	8	0.42	0.672	0.094	0.257	0.174	0.154	0.012	0.052
7	9	0.188	0.088	0.194	0.112	0.044	0.007	0.026	0.021
8	1	0.172	0.237	0.061	0.21	0.039	0.02	0.004	0.043
8	2	0.244	0.025	0.222	0.148	0.046	0	0.034	0.014
8	3	0.202	0.208	0.12	0.072	0.02	0.014	0.01	0.005

associated with A_{53} as $P(H_1^{(53)}) = 0.551$. The reason for this has been explained in the simulation study.

Our hierarchical model addresses multiplicities. For all other AEs without true safety signal, the estimated $P(H_1^{(bj)})$ are generally smaller under Bayesian hierarchical model than under solo

Bayesian model because $P(H_1^{(bj)})$ tend to be "shrunk" toward 0.

If we set the cut-off point p_0 as 0.7, it is feasible to draw the conclusion that A_{53}, A_{68}, A_{15} have safety signal.

3.6 Discussion

Traditional safety analysis focuses on comparing only the AE incidence between two groups, regardless of the AE severity. It is probably of great interest for us to also compare the severities of the AEs, especially when the severities of some AEs are consistently greater for one treatment versus the other. In this article, we present a three-stage Bayesian hierarchical non-proportional odds version of the cumulative logit model for comparing both the incidence and severity of AEs reported in the two-arm randomized clinical trial. The proposed Bayesian cumulative logit model is a novel extension of a Bayesian hierarchical logit model proposed by Berry & Berry (2004). There are a few advantages of our Bayesian approach. First, the proposed model can handle the four level ordinal categorical safety outcome that contains information about AE occurrence and severity. Non-proportional odds assumption allows the treatment effect to vary with severity level. Second, the use of a point-mass mixture prior is important because many of the AEs may have same incidence and severity in both the control and treatment group. In addition, it is straightforward to assess the incidence and severity of all the AEs based on the posterior probability that the incidence rate or the overall severity is greater in the treatment group. Finally, in contrast to considering each type of AE independently in the solo Bayesian model, our Bayesian hierarchical model addresses multiplicity because it borrows information across types of AEs, especially across the AEs that are classified into the same group (for example, body system). Our simulation study shows that the proposed method has low FDR and high power for detecting safety signal.

Body system allows us to exploit information across types of AEs that are related. Assignment of AEs to different body systems may lead to different conclusions. This is the negative aspect of the method. So it is important that we only assign types of AEs that are biologically related into a body system. This requires the help from biological or clinical experts.

Another limitation with the proposed method is that with small sample size, the number of subjects under different severity level can be small, especially the number of subjects under the severity level of severe. Some types of AEs without safety signals with respect to the third and fourth severity level may be affected by the AE with safety signal in the same body system and thus lead to the potential increase in the false positive rate.

Our model is based on summary (marginal) data regarding the number of subjects under each of the severity level. In the future research, with data that contains patient level information, we can fit a cumulative logit model with a random effect to account for dependencies among the various AEs at the subject level. Conclusions based on the random effect model could be more precise.

The current application of our method is in a single trial setting. However, multiple clinical trials are usually conducted on an experimental drug. The second future direction of our research is to incorporate individual subject data from multiple studies into the model. In addition, the proposed method is applied once to the safety data when the trial is completed. It is of great interest to extend the proposed method to a continuous monitoring framework. Thus another future direction is that the proposed Bayesian hierarchical cumulative logit model can be applied in a group sequential manner for multiple interim analyses of safety events.

Chapter 4

Interactive Tools for Blinded Safety Monitoring in Clinical Trials using Bayesian Methods

Abstract

Safety monitoring is critical for ongoing clinical trials and it is often conducted by an Independent Data Monitoring Committee (IDMC) periodically based on unblinded data. As sponsors develop novel drugs, there is an increasing need for sponsors to detect potential safety signals as soon as possible ahead of IDMC schedule and prompt decisions regarding an unblinded analysis. A few quantitative methods for blinded safety monitoring have been developed. The complex nature of these methods makes the safety monitoring and reporting challenging. In this article, we develop two interactive tools to accelerate and facilitate the process of blinded safety monitoring and reporting using RShiny Dashboard package in R Studio. These two interactive tools are user-friendly and convenient for users to conduct blinded safety monitoring during the ongoing trial.

Key words: Adverse events; Bayesian model; Beta-Binomial model; Blinded safety monitoring; Poisson-Gamma model; R-Shiny.

4.1 Introduction

Safety evaluation is one of the most important goals in clinical trials. Appropriate evaluation of safety data during the conduct of a clinical trial ensures timely alteration or termination of the trial to protect the trial participants from being exposed to harmful treatment and save development costs and time.

The recent FDA draft guidance on safety assessment for Investigational New Drug (IND) Safety Reporting (FDA, 2012, 2015) states that sponsors should develop a safety assessment committee and a safety surveillance plan as key elements of a systematic approach to safety surveillance. In addition, the sponsors should oversee the evolving safety profile of the investigational drug by evaluating, at appropriate intervals, the cumulative serious adverse events from all of the trials in the development program and other available important safety information (e.g., findings from epidemiological studies and from animal or in vitro testing). They should also perform unblinded comparisons of event rates in investigational drug and control groups, as needed.

As summarized in Zhu et al. (2016), safety evaluation in clinical trials includes two major areas generally referred as "safety signal detection" and "safety monitoring". In the area of safety signal detection, all adverse events (AEs) reported from the clinical trial are to be included in the analysis. In a two arm clinical trial, the goal of safety signal detection is to compare the incidence rates of many AEs between two groups and identify a potential subset of AEs with significantly greater incidence rate in the treatment group. Simultaneously comparing the incidence rate of many AEs causes multiplicity issues. So the major challenge is the multiplicity adjustment. Some frequentist and Bayesian methods have been proposed for signal detection in clinical trial AE data. See Mehrotra & Heyse (2004), Mehrotra & Adewale (2012), Berry & Berry (2004), Xia et al. (2011), Diao et al. (2019). Some authors also have considered detecting safety signals with respect to both the incidence and severity, for example, Klingenberg et al. (2009), L.G. Leon-Novelo & Muller (2010), Duan et al. (2019a), Duan et al. (2019b).

The second area of drug safety evaluation in clinical trials is the examination of accumulating data, which we refer as safety monitoring. In this area, we aims at continuously monitoring an adverse event of special interest (AESI) that is specified upfront in the ongoing study. Zhu et al. (2016) reviewed frequentist and Bayesian methods for safety monitoring, including Sequential Probability Ratio Test (SPRT), a continuous sequential stopping rule by Goldman & Hannan (2001) and a decision criterion by Thall & Simon (1994) for single arm trial monitoring. Zhu et al. (2016) also proposed two Bayesian methods for two-arm trial safety monitoring based on the work by Yao et al. (2013).

Safety monitoring in clinical trials is often conducted by an Independent Data Monitoring Committee (IDMC) periodically based on unblinded data or solely by the sponsors. As sponsors develop novel drugs and biological agents in new diseases or therapeutic areas, there is an increasing need for them to monitor patient safety to detect potential safety signals as soon as possible ahead of IDMC schedule and prompt decisions regarding an unblinded analysis, while maintaining study blinding to minimize bias and protect the integrity of the blinded studies. For example, in some pharmaceutical companies, a Safety Monitoring Teams (SMT) is established for reviewing

blinded reports and listings of safety data on a regular basis and making determinations on whether there is a change in the risk profile of the drug based on clinical judgment without preset decision rules or criteria.

Bayesian approaches provide a natural framework to identify early safety signal for further action without unblinding ongoing studies by incorporating prior knowledge about safety profile of control group or background rate of events and by updating knowledge using cumulative data from the ongoing trial. Assumptions on the safety profile of control group or background rate must be made utilizing historical information or epidemiology data.

A few Bayesian approaches have been proposed for blinded monitoring of adverse events in two-arm clinical trials. Ball (2011) described an approach based on pooled adverse event rate from a two-arm randomized trial. The decision rule is that if there is a high posterior probability given cumulative data that the pooled event rate is greater than a background rate in the literature, then there is a potential safety signal with respect to the AE. Gould & Wang (2017) proposed a Bayesian approach for blinded safety monitoring of two-arm randomized trials where the potential adverse event risk levels can be estimated with different treatment effect metrics such as relative risk, absolute risk difference, or odds ratio. Their Bayesian approach allows explicit specification of prior beliefs (i.e., prior distributions) about the control group event rate and the treatment effect metrics. Schnell & Ball (2016) introduced a Bayesian hierarchical exposure-time model for two-arm blinded safety monitoring. Mukhopadhyay et al. (2018) proposed a two-step Bayesian method for monitoring and detecting safety signals from blinded safety data for AESI.

The complex nature of these methods makes the application challenging. In order to accelerate, facilitate and improve the process of blinded safety monitoring and reporting, it is necessary to develop ready-to-use tools based on the existing safety monitoring and reporting methods. In this article, among all the blinded safety monitoring methods, we focus on the methods proposed by Ball (2011) and Gould & Wang (2017). We first describe the details of these two methods. In addition, during interim monitoring of safety data, subjects with more exposure to study drug and longer observation time may have higher risk of having an adverse event. This is not considered

by these authors. Therefore, we further describe the extensions of these two methods by considering the exposure adjusted incidence rate. Based on these two methods and their extensions, we develop two interactive tools for blinded safety monitoring by using RShiny Dashboard package in R Studio. These two interactive tools are user-friendly and convenient for users to conduct blinded safety monitoring during the ongoing trial.

The rest of the article is organized as follows. We review two Bayesian approaches for blinded safety monitoring in Section 4.2 and Section 4.3. In Section 4.4, we describe the two interactive tools and we provide a summary of the two interactive tools and some thoughts on future enhancement of the interactive tools in Section 4.5.

4.2 Bayesian approach based on pooled rate

We briefly describe the method proposed by Ball (2011) as well as its extension with Poisson model.

4.2.1 Binary event data - Beta-binomial model

In a two-arm randomized trial, consider an AESI, denote x_n as the number of subjects who experience the AE in the pooled data at the time of interim monitoring, denote N as the total number of subjects in the trial and θ as the pooled event rate. Ball (2011) assumed that x_n follows binomial distribution, i.e., $x_n \sim Bin(N, \theta)$. The likelihood function of the pooled event rate θ given x_n is

$$L(\theta; x_n) = \binom{N}{x_n} \theta^{x_n} (1-\theta)^{N-x_n}$$

From the Bayesian perspective, Ball (2011) assumed beta conjugate prior for the pooled event rate θ. $\theta \sim Beta(\alpha_0, \beta_0)$, where α_0 can be interpreted as prior number of successes and β_0 as prior number of failures. The larger the $\alpha_0 + \beta_0$, the more informative the prior distribution.

The posterior distribution of θ given x_n is $\theta | x_n \sim Beta(\alpha_0 + x_n, \beta_0 + n - x_n)$. From the historical information, let an event rate for control group be c_1. The decision rule is that if $P(\theta_n > c_1 | x_n) \geq P$,

where P is a high probability defined a priori, e.g., 80% or 90%, then the data may suggest a potential safety signal that needs further investigation to determine if there is a safety signal that warrants further action. Otherwise, we may continue the study as is. The prior is kept weakly informative to avoid too much influence of prior information on the decision, i.e., $\alpha_0 + \beta_0$ is small.

4.2.2 Exposure adjusted incidence rate - Poisson gamma model

During interim monitoring of safety data, the exposure and observation time for each subject are often different. Subjects with more exposure to study drug and longer observation time may have higher risk of having an AE than subjects with shorter exposure and observation time. In such situation, it is more appropriate to consider exposure adjusted incidence rate of event instead of a simple binary event rate.

Denote λ as the pooled incidence rate per 100 subject years and E as the total follow up time of the subjects in the trial in subject-years, or total exposure time at the time of analysis. Let γ be the incidence rate of the pooled data during the exposure time E, which is equal to $\frac{\lambda E}{100}$.

We assume the number of subjects who experience the AE in the pooled data during the total observation period of E follows Poisson distribution, i.e., $x_n \sim Poi(\gamma)$. The likelihood function of the parameter γ, the pooled incidence rate during the exposure time E is

$$L(\gamma; x_n) = \frac{\gamma^{x_n}}{x_n!} e^{-\gamma}$$

The prior distribution of the pooled incidence rate per 100 subject-years is a gamma distribution, $\lambda \sim Gamma(\alpha_1, \beta_1)$, where the parameters can be interpreted as α_1 events in β_1 interval of time. Thus, the pooled incidence rate during the exposure time E is $\gamma = \lambda E/100 \sim Gamma(\alpha_1, \frac{100\beta_1}{E})$.

Since Gamma prior distribution is a conjugate prior, the posterior probability distribution function has a closed form. The posterior distribution for the pooled incidence rate given the observed number of subjects who experience the AESI is thus a Gamma distribution, $\gamma | x_n \sim$

$Gamma(\alpha_1 + x_n, \frac{100\beta_1}{E} + 1)$.

The decision rule is that if $P(\gamma > c_2 | x_n) \geq P$, there may be a potential signal alert where $c_2 = \frac{\alpha_1 E}{100\beta_1}$ and P are critical values. Further investigation is needed if the approach suggests a potential signal before taking action with the study.

4.3 Bayesian approach based on treatment effect metrics

We briefly describe the method proposed by Gould & Wang (2017) and its extension in this section.

4.3.1 Binary event data - Beta-binomial model

In a two-arm randomized clinical trial, denote p_C and p_T as the incidence rate of the AESI in the control group and treatment group respectively. Denote x_C and x_T as the number of subjects who experience the AESI in the control group and treatment group respectively. Let N be the total number of subjects in the trial and τ be the fraction of the total sample size allocated to the treatment group. Gould & Wang (2017) assumed that $x_T \sim Bin(N\tau, \theta_T)$ and $x_C \sim Bin(N(1-\tau), \theta_C)$. So the likelihood function of p_C and p_T given x_C, x_T, N is the product of two probability mass functions of binomial distribution.

After reparameterization, the likelihood function can be re-expressed as a function of a treatment metric M and p_C given the pooled number of subjects who experience the AE (X), the number of subjects who experience the AE in the treatment group (x_T) and the number of subjects enrolled in the trial (N) (Gould & Wang, 2017). The treatment metric M can be risk ratio $R = \frac{p_T}{p_C}$ or risk difference $D = p_T - p_C$. If the treatment metric M is risk ratio $R = \frac{p_T}{p_C}$, the likelihood function of R and p_C is

$$L(R, p_C) = f_{binom}(X; N, p_C)\Big(\frac{1 - Rp_C}{1 - p_C}\Big)^{\tau N} C(x_T, X, N, \tau)\Big(\frac{R(1-p_C)}{1 - Rp_C}\Big)^{x_T}$$

where f_{binom} is the binomial probability mass function, and $C(x_T, X, N, \tau) = \binom{\tau N}{x_T}\binom{(1-\tau)N}{X - x_T} / \binom{N}{X}$ is a hypergeometric distribution term. If the treatment metric M is risk difference $D = p_T - p_C$, the

likelihood functions of the D and p_C is

$$L(D, p_C) = f_{binom}(X; N, p_C)\Big(\frac{1 - p_C - D}{1 - p_C}\Big)^{\tau N} C(x_T, X, N, \tau)\Big(\frac{(p_C + D)(1 - p_C)}{p_C(1 - p_C - D)}\Big)^{x_T}$$

See the Appendix A.1 for the derivations of the above likelihood functions. x_T is an unobserved value. Its value ranges from 0 to X. Summing the likelihood function L over the range of x_T gives the likelihood function of M and p_C given X, $L(M, p_C|X)$.

A beta distribution is used to characterize the prior belief on the control group event rate p_C (Gould & Wang, 2017). Parameters for the prior beta distributions can be selected based on historical knowledge about the control group event rate from meta-analyses of previous trials or large electronic health record databases.

If treatment effect metric is risk difference D, a uniform prior distribution defined on the interval $[-1, 1]$ is a convenient prior. If treatment effect metric is relative risk R, lognormal distribution may be considered for the prior, with its parameters being selected to satisfy that the mean of R equals to 1 (i.e., no elevated AE risk associated with the test drug) and most of the probability density lies within a wide interval, e.g., $(0.1, 10)$ if control group event rate is 5%. It is advisable to use weakly informative prior for treatment effect so that the posterior distribution is primarily driven by the observed event rate in the current trial under monitoring (Gould & Wang, 2017).

The joint posterior density of M given X is proportional to the product of $L(M, p_C|X)$ and the prior density functions for M and p_C. Integrating the joint posterior density of M and p_C with respect to p_C yields the marginal posterior density of M, which then can be used to calculate the posterior probability. Since the posterior probability distribution function does not have a closed form, Gould & Wang (2017) used the grid approximation to calculate the posterior probability. They set up a grid with a range that covers most of the density of M and p_C and then sum it up by row or column to get the marginal probability of M or p_C.

The decision rule is that if $P(M > M_{crit}|X) \geq \gamma_{crit}$, there is a potential signal alert with critical values M_{crit} and γ_{crit}. This is equivalent to that if $X > X_{crit}$, there may be a potential signal alert

where X_{crit} is the minimum X s.t. $P(M > M_{crit}|X) \geq \gamma_{crit}$.

Implementing Bayesian approach requires complicated statistical computation and calls for tangible presentation of the decision rules to guide regulatory and clinical in monitoring safety events in blinded trials. Gould & Wang (2017) recommended using control charts as a monitoring tool. In the control chart, $\frac{X_{crit}}{N}$ is plotted against N (the number of subjects enrolled in the trial). If the observed X is greater than X_{crit} at a given N, a potential signal for elevated risk of AE is identified and further investigation including unblinding treatment assignment information on the AE cases (by an internal independent team or an external DMC) may be warranted. The translation from the decision rules to the control charts is a result of the fact that the posterior probability that $M > M_{crit}$ is a monotone increasing function of X_{crit}.

After we have obtained the critical value X_{crit}, further investigation of the likelihood of observing such critical value is needed, as the model assumption may not be consistent with the truth. (Gould & Wang, 2017) recommended that we investigate the plot of probability of observing the critical value X_{crit} given the assumptions about the control group event rate and the metric R or D. The y axis of the plot is $P(X > X_{crit}|M, p_C)$. The x axis is the critical value X_{crit}. The probability curve in the plot is a non-increasing smoothing curve.

4.3.2 Exposure adjusted incidence rate - Poisson gamma model

If the total sample size is large and the event incidence is low, Gould & Wang (2017) considered using Poisson model. However, they assumed the exposure and observation time for each subject are the same, which is unrealistic and may lead to biased monitoring results. Subjects with more exposure to study drug and longer observation time may have higher risk of having an adverse event than subjects with shorter exposure and observation time. To more accurately account for difference in exposure time, we consider exposure adjusted incidence rate.

Denote the AE incidence rate with exposure index being 100 subject year in the control group and treatment group as λ_C and λ_T respectively. Denote E as the total follow up time of the subjects in the trial, or total exposure time. Thus the exposure adjusted incidence rate of control group and

the treatment group are $\theta_C = \frac{\lambda_C E(1-\tau)}{100}$ and $\theta_T = \frac{\lambda_T E \tau}{100}$ respectively. Assumed that $x_T \sim Poi(\theta_T)$ and $x_C \sim Poi(\theta_C)$ (Gould & Wang, 2017). So the likelihood function of θ_C and θ_T given x_C, x_T is:

$$L(\theta_C, \theta_T; x_C, x_T) = f_{poiss}(x_C; \theta_C) f_{poiss}(x_T; \theta_T)$$

where f_{poiss} is the binomial probability mass function. After reparameterization, the above likelihood function can be re-expressed as a function of a treatment metric M and the control group exposure adjusted incidence rate θ_C given the pooled number of subjects who experience the AE (X), the number of subjects who experience the AE in the treatment group (x_T) and the number of subjects enrolled in the trial (N) (Gould & Wang, 2017).

If the treatment metric M is risk ratio $R = \frac{\theta_T}{\theta_C}$, the likelihood function of R and θ_C is

$$L(R, \theta_C; X, x_T) = f_{poiss}(X; (1+\xi R)\theta_C) \binom{X}{x_T} \left(\frac{\xi R}{1+\xi R}\right)^{x_T} \left(\frac{1}{1+\xi R}\right)^{X-x_T}$$

where $\xi = \frac{\tau}{1-\tau}$ and $R = \frac{\lambda_T}{\lambda_C}$.

Gould & Wang (2017) also mentioned using risk difference, but they did not give details about the derivation. If the treatment effect metric M is risk difference $D = \lambda_T - \lambda_C$, the likelihood function of D and θ_C given X, x_T is

$$\begin{aligned} L(D, \theta_C; X, x_T) &= \frac{\theta_T^{x_T}}{x_T!} e^{-\theta_T} \frac{\theta_C^{x_C}}{x_C!} e^{-\theta_C} \\ &= \frac{\left(\theta_C + D\frac{0.5E}{100}\right)^{x_T}}{x_T!} e^{-(\theta_C + D\frac{0.5E}{100})} \frac{\theta_C^{X-x_T}}{(X-x_T)!} e^{-\theta_C} \end{aligned}$$

We found that τ has to be 0.5, otherwise the likelihood function cannot be reparameterized in terms of D and θ_C. x_T is an unobserved value. Its value ranges from 0 to X. Summing the above likelihood function over the range of x_T gives the likelihood function of M and θ_C given X, $L(M, \theta_C | X)$.

Assume that our prior knowledge of the control group incidence rate per 100 subject-years has

a Gamma distribution with parameters α_1, β_1. $\lambda_C \sim Gamma(\alpha_1, \beta_1)$. Parameters for the gamma distribution can be selected based on historical knowledge about the control group event rate from meta-analyses of previous trials or large electronic health record databases. Hence, the prior distribution of the incidence rate of the control group θ_C during the total exposure in control group is gamma distribution with parameters $\alpha_1, \frac{100\beta_1}{E(1-\tau)}$.

The prior distribution of R is assumed to be log-normal with its parameters selected to satisfy that the mean of R equals to 1 (i.e., no elevated AE risk associated with the test drug) and most of the probability density lies within a wide interval, e.g., $(0.1, 10)$. The prior distribution of D is assumed to be normally distributed with large variance to represent non-informative prior knowledge.

The posterior density of R can be obtained following the same steps as outlined for the preceding binomial model. If we use ratio as the metric, the posterior probability distribution function has a closed form, i.e., gamma distribution. However, if we use the difference as the metric, the posterior pdf does not have a closed form. Same as how get the posterior probability under beta-binomial model, we use the grid approximation to calculate the posterior probability.

The decision rule is that if $P(M > M_{crit}|X) \geq \gamma_{crit}$, there may be a potential signal alert for further investigation where M_{crit} and γ_{crit} are critical values. This is equivalent to that if $X > X_{crit}$, there may be a potential signal alert where X_{crit} is the minimum X s.t. $P(M > M_{crit}|X) \geq \gamma_{crit}$.

Same as that in beta-binomial model, we can investigate the plot of probability that the observed X is greater than the critical value X_{crit} given the assumptions about the control group incidence rate and the metric R or D.

4.4 Interactive tools

In this section we introduce two tools developed by using R-Shiny Dashboard package in R Studio.

4.4.1 An interactive tool based on pooled rate

The first interactive tool is developed based on the method proposed by Ball (2011). It can be used to perform computations for posterior probability associated with the pooled event rate or pooled exposure adjusted incidence rate of an AESI and dynamically output the conclusion about whether the data suggest a safety signal. The tool has a side bar menu that consists of four tabs: About, Prior visualization, Binary event rate beta binomial model and Exposure adjusted incidence rate Poisson gamma model. See Figure 4.1. The "About" tab contains short instructions about how to

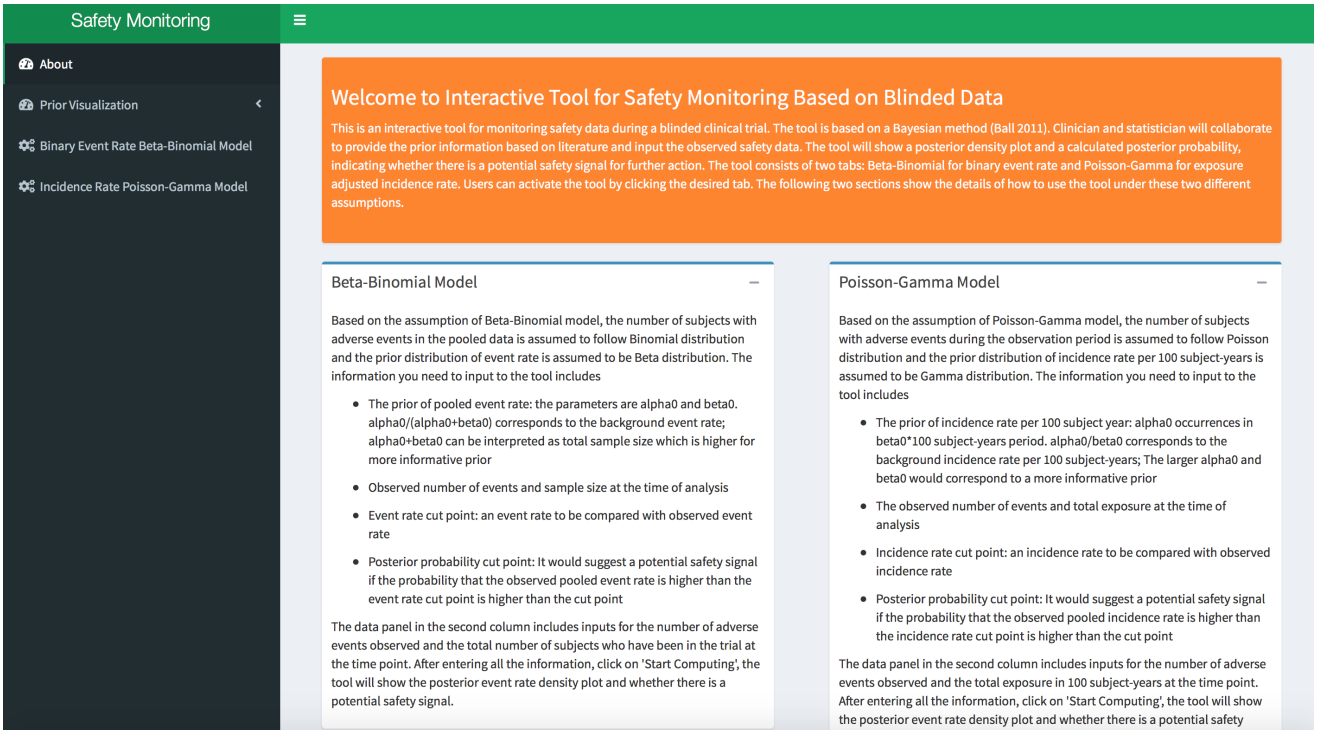

Figure 4.1: "About" tab of the first interactive tool

use the tool and descriptions about two model assumptions.

The "Prior visualization" tab is used for the overview of the prior distributions of Beta and Gamma distribution. (Figure 4.2 and Figure 4.3). It can be used to visualize how different parameter values of the prior distribution affect the shape of the density and how informative the prior is. There are two sub-tabs: "Beta distribution" and "Gamma distribution". Each of them contains the parameter value input boxes and the probability density function plot. For Gamma distribution, there are two additional input boxes for determining the minimum and maximum of the x-axis. The

probability density function plot changes dynamically as the user changes the parameter values in the input box.

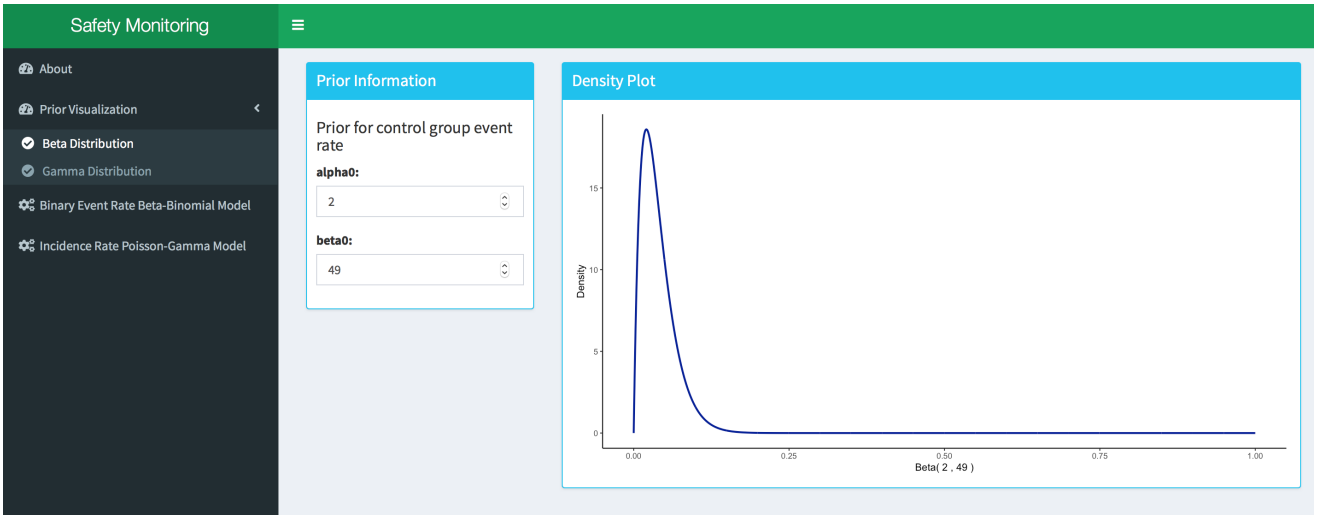

Figure 4.2: Beta distribution density plot in the "Prior Visualization" tab

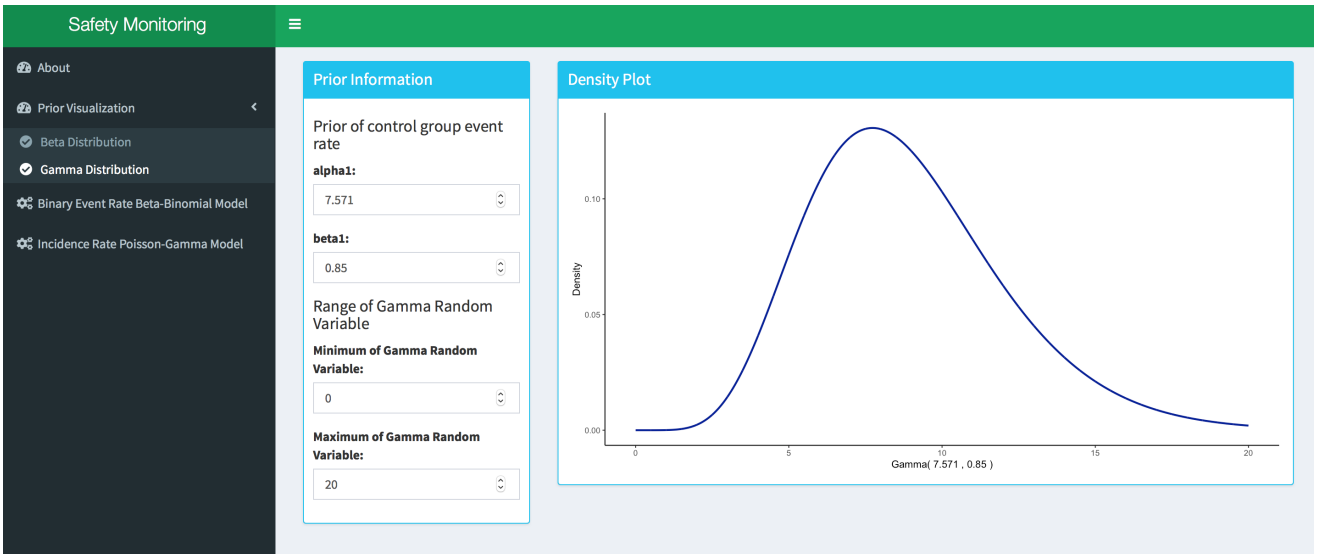

Figure 4.3: Gamma distribution density plot in the "Prior Visualization" tab

The "Binary event rate - beta-binomial model" tab and "Incidence rate - Poisson-Gamma model" tab are the main functions of the tool for safety monitoring under two different model assumptions.

The "Binary event rate - beta-binomial model" tab is developed for safety monitoring when the pooled number of subjects who experience the AESI is assumed to follow a binomial distribution.

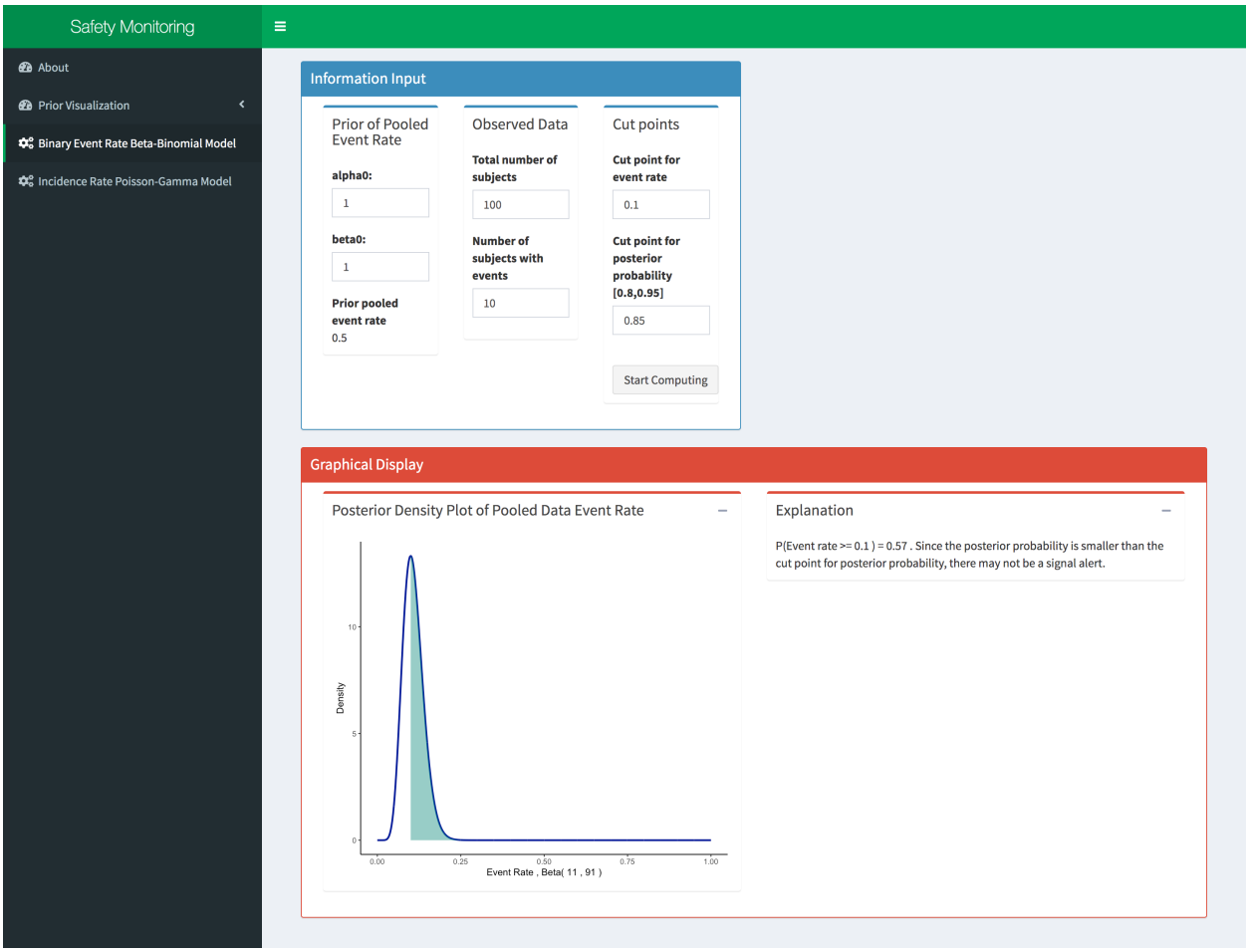

Figure 4.4: "Binary event rate - beta-binomial model" tab in the first tool

There are two panels under this tab. See Figure 4.4. The panel with blue header is the information input panel. There are three columns of input boxes for users to input values, including the parameter values of the prior distribution of the pooled event rate θ, the total number of subjects in the trial, the pooled number of subjects who experience the AESI and the critical values. After entering the parameter values of the prior distribution, the prior pooled event rate will be calculated and shown. To start conducting safety monitoring, click on "start computing" button. The analysis results then show in the second panel, the graphical display panel with red header. The results consist of a posterior density plot box and an explanation box. In the posterior density plot box, there is a plot of the posterior probability density function of pooled event rate. The green shaded area is the posterior probability that the pooled event rate is larger or equal to its critical value.

The box on the right side of the panel shows the explanations of the analysis result. The tool dynamically compares the calculated posterior probability with the posterior probability cut point. If the calculated posterior probability is larger than the posterior cut off probability, the tool outputs the conclusion that there may be a potential safety signal and recommend further action. Otherwise, it outputs the conclusion that there may not be a signal alert.

The "Poisson gamma model" tab is developed to conduct safety monitoring when the pooled number of subjects who experience the AESI is assumed to follow a Poisson distribution. There are two panels under this tab. See Figure 4.5.

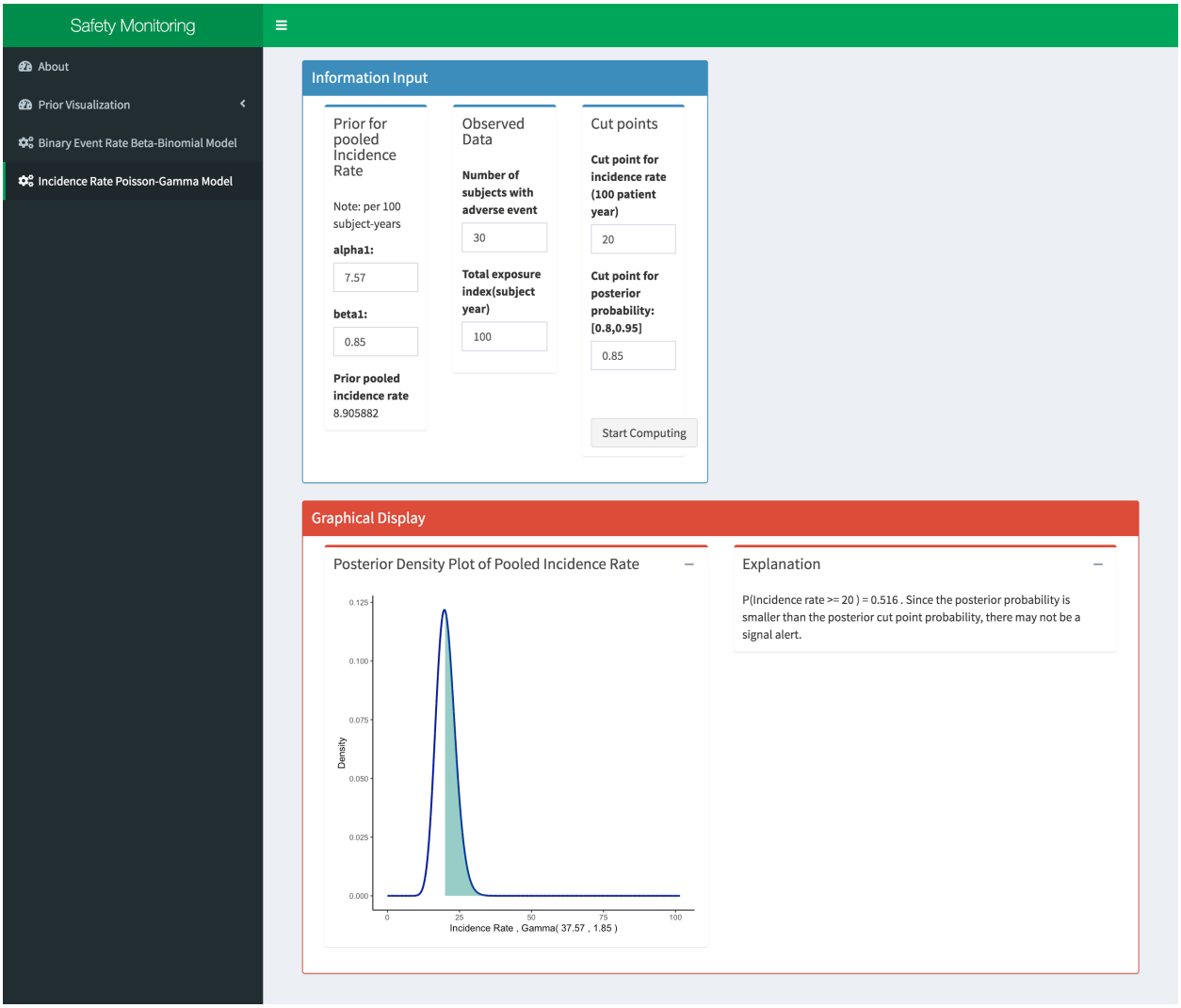

Figure 4.5: "Incidence rate - Poisson-Gamma model" tab in the first tool

The first panel with blue header is the information input panel. There are three columns of

input boxes for the users to input values, including the parameter values of the prior distribution of the pooled data incidence rate λ, the observed data, and the cut points. After entering the parameter values of the prior distribution, the prior pooled incidence rate will be calculated. The prior incidence rate and the cut point for incidence rate are both expressed in rate per 100 subject years exposure. The tool automatically adjusts the incidence rate to account for the total exposure E (in subject-years). The user can enter the calculated prior pooled incidence rate or twice the prior pooled incidence rate as the cut point after discussing with clinicians. To start conducting safety monitoring, click on "start computing" button.

The analysis results show in the second panel, the graphical display panel with red header. The results consist of a graphical display box and an explanation box. In the graphical display box, there is a plot of the posterior probability density function of pooled incidence rate. The green shaded area is the posterior probability that the pooled incidence rate is larger or equal to its critical value. The box on the right side of the panel shows the explanations of the analysis result. The tool dynamically compares the calculated posterior probability with the posterior cut off probability. If the posterior probability is larger than the posterior cut off probability, the tool outputs the conclusion that there is a potential safety signal and recommend further action. Otherwise, it outputs the conclusion that there is not a signal alert.

4.4.2 An interactive tool based on treatment effect metrics

The second interactive tool is developed based on the method proposed by Gould & Wang (2017). It can be used to perform computations for posterior probability of the treatment effect metric of an AESI and dynamically output the result about whether the data suggest a safety signal. The tool has a side bar menu that consists of four tabs: About, Prior visualization, Single trial, Multiple trials. See Figure 4.6.

The "About" tab contains short instructions about how to use the tool and descriptions about two model assumptions. The "Prior visualization" tab is used for the overview of the prior distributions of Beta and Gamma distribution, same as that in the first interactive tool.

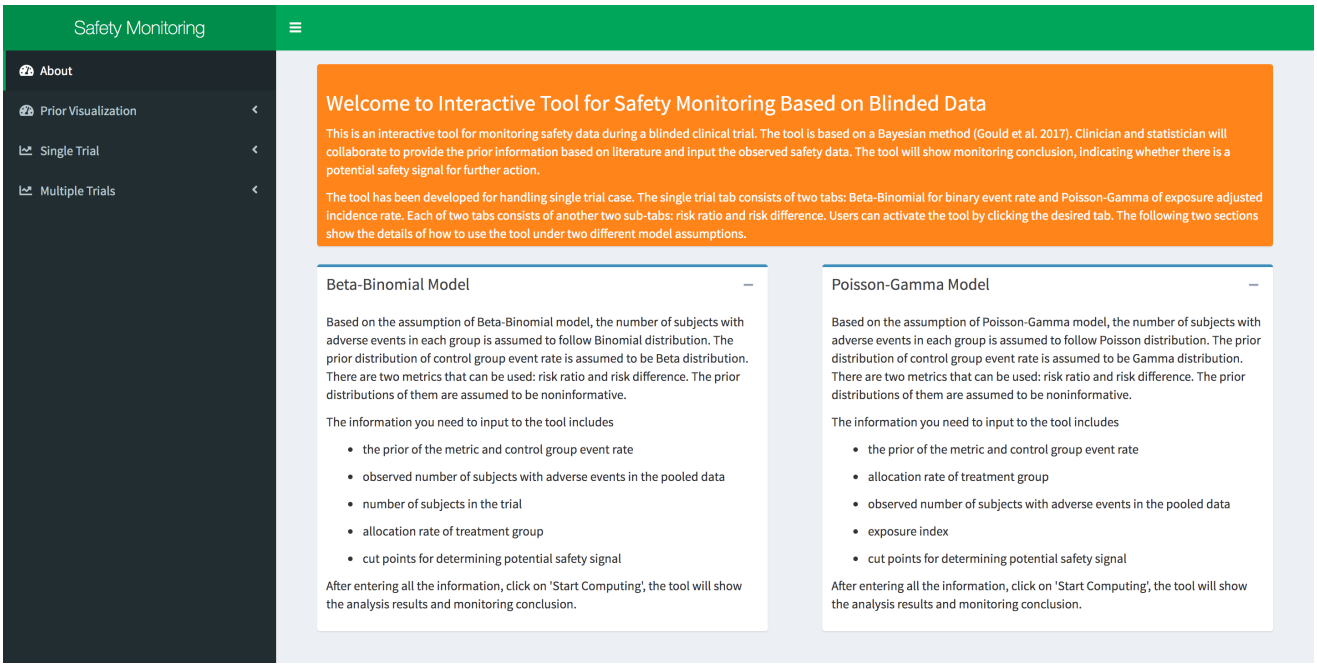

Figure 4.6: "About" tab of the second interactive tool

Under "single trial" tab, there are two sub-tabs: "Binary event rate - beta-binomial model" and "Incidence rate - Poisson-Gamma model", each of which consists of another two sub-tabs: "risk ratio" and "risk difference".

The tab "Binary event rate - beta-binomial model" tab is developed for safety monitoring when the number of subjects who experience the AESI in each group (treatment, control group) is assumed to follow Binomial distribution. This tab consists of two sub-tabs "risk ratio" (Figure 4.7), for safety monitoring when risk ratio is used as the treatment effect metric and "risk difference" (Figure 4.8), for safety monitoring when risk difference is used as the treatment effect metric. Under each of these two sub-tabs, there are three panels.

The first panel with blue header is the information input panel. There are five columns of input boxes for users to input values, including the parameter values of the prior distribution of the risk ratio R or risk difference D, the control group event rate p_C, allocation rate of treatment group, number of subjects in the trial, pooled number of subjects who experience the AESI, some assumptions about the value of the parameters and the cut points. To start conducting safety monitoring, click on "start computing" button.

The second panel with red header is the monitoring records panel. In this panel, there is table that shows the records of the historical monitoring conclusions that correspond to each input of the prior information, number of subjects in the trial, critical values of the metric R or D and γ.

The analysis results of the current safety monitoring is shown in the third panel, the graphical display panel with orange header. There are two analysis plots in this panel: control chart boundaries plot and the plot of the probability of observing certain number of AESI. The control chart boundaries plot on the left side of the panel shows the plot of critical pooled event rate (critical value of X divided by the number of subjects in the trial) versus number of the subjects in the trial. The black dots are the corresponding critical pooled event rate. The dots are connected to illustrate the overall trend. The red dot represents the observed pooled event rate of the current trial. If the observed events rate is larger than the critical event rate, a label "potential signal alert" will show next to the red dot. Otherwise, a "may not be signal alert" label will show next to the red dot.

On the right side of the panel, there is a plot that shows the probability that the observed pooled number of subjects with the AE is greater than a certain number x under three assumptions about the value of control group event rate and the metric R or D. There are three smoothing curves. Each curve corresponds to one of the three assumptions about the values of R or D the user input. The red dashed line represents the critical value of X for the current monitoring and the intersections between the red dashed line and three smoothing curves are the corresponding probabilities of observing a pooled number of subjects with the AE that is greater than the critical value of X under three assumptions. These probabilities can be found in the second panel, the monitoring records panel. Under each of the above two plots, there is a corresponding short paragraph of explanation.

The tab "Incidence rate - Poisson-Gamma model" is developed when the number of subjects who experience the AESI in each group (treatment, control group) is assumed to follow Poisson distribution. Similarly, this tab consists of two sub-tabs: "risk ratio", for safety monitoring when risk ratio is used as the treatment effect metric (Figure 4.9) and "risk difference", for safety monitoring when risk difference is used as the treatment effect metric (Figure 4.10). There are three panels under each of the sub-tabs.

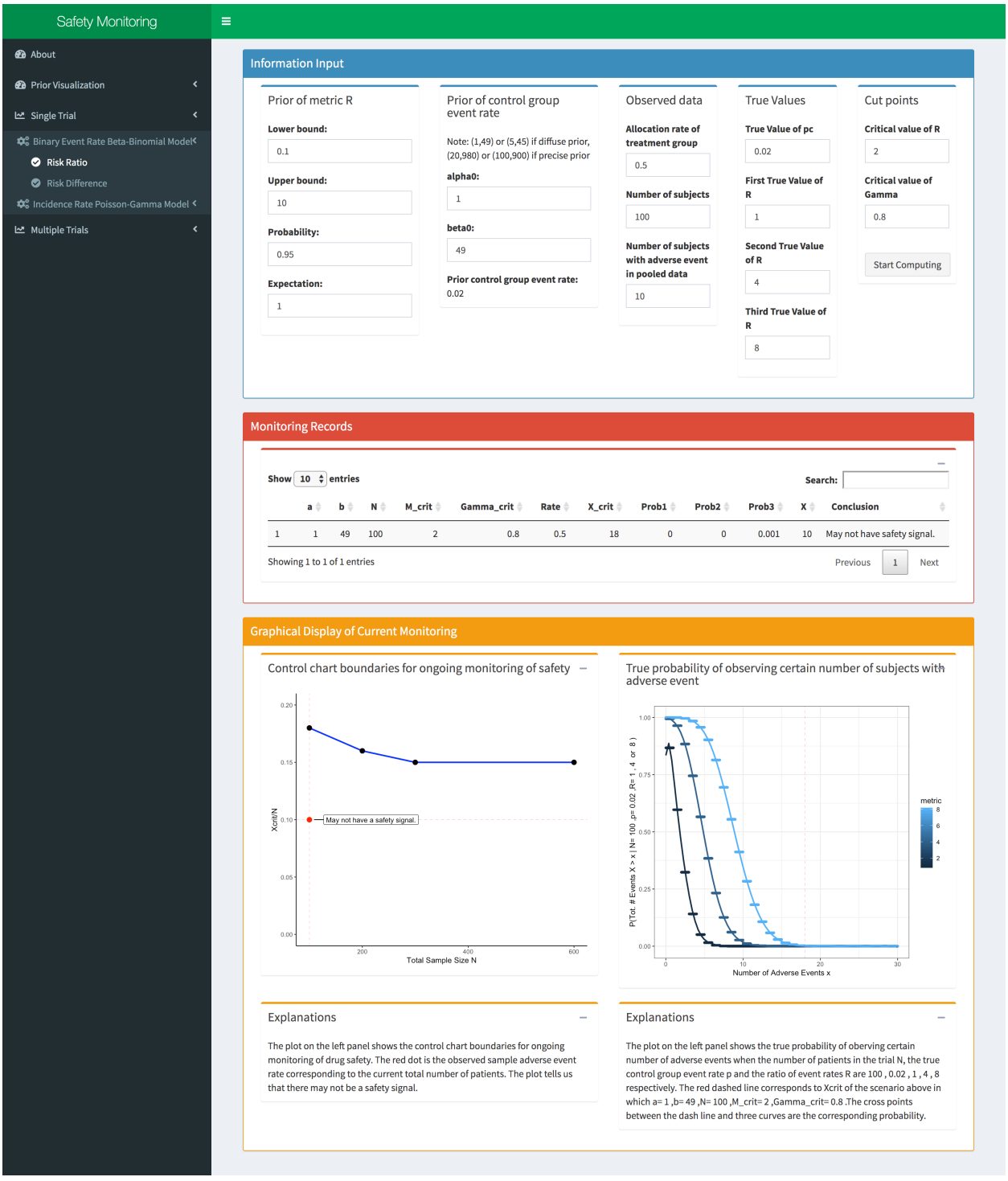

Figure 4.7: "Risk Ratio" sub-tab under "Binary event rate - beta-binomial model" tab in the second tool

The first panel with blue header is the information input panel. There are five columns of input boxes for users to input values, including the parameter values of the prior distribution of the risk

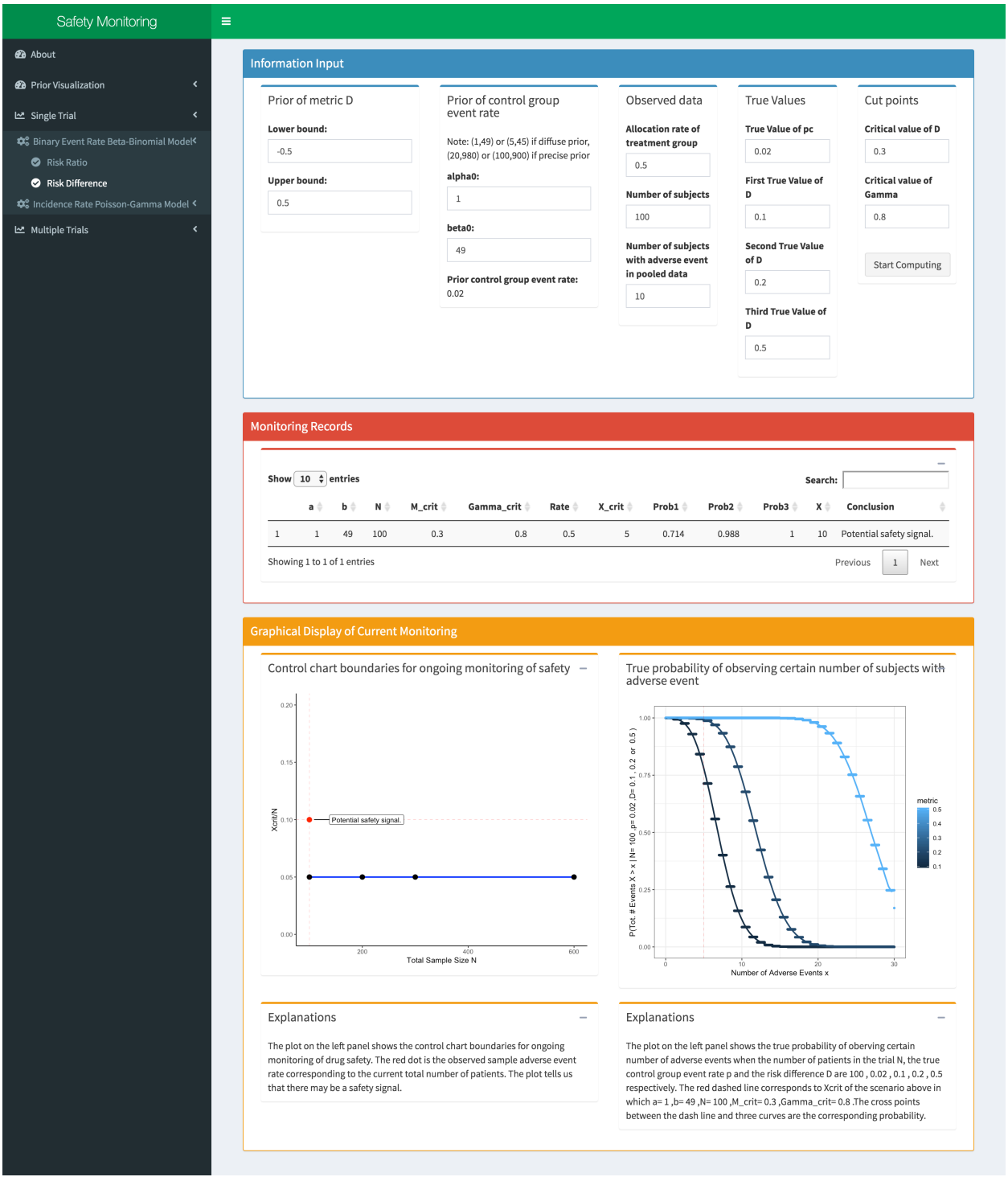

Figure 4.8: "Risk Difference" sub-tab under "Binary event rate - beta-binomial model" tab in the second tool

ratio R or risk difference D, the control group incidence rate, observed data, true values of the parameters and the cut points. Once we click on "start computing" button, the second panel shows

the monitoring conclusion that correspond to the input of the prior information, number of subjects in the trial, critical values of the metric R or D and γ.

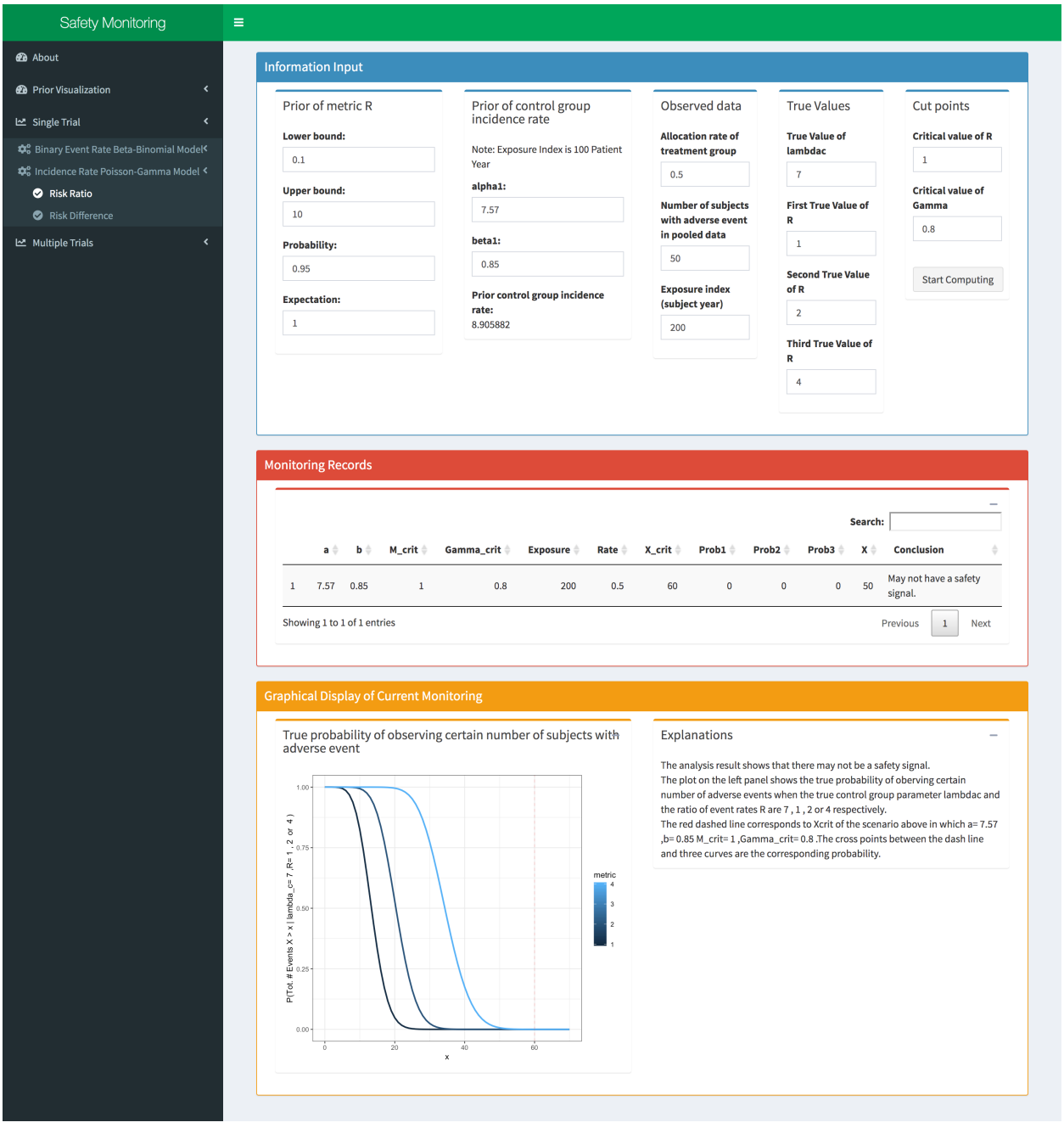

Figure 4.9: “Risk Ratio” sub-tab under “Incidence rate - Poisson-Gamma model” tab in the second tool

The analysis results of the current safety monitoring is shown in the third panel, the graphical display panel with orange header. There is a plot that shows the probability that the observed

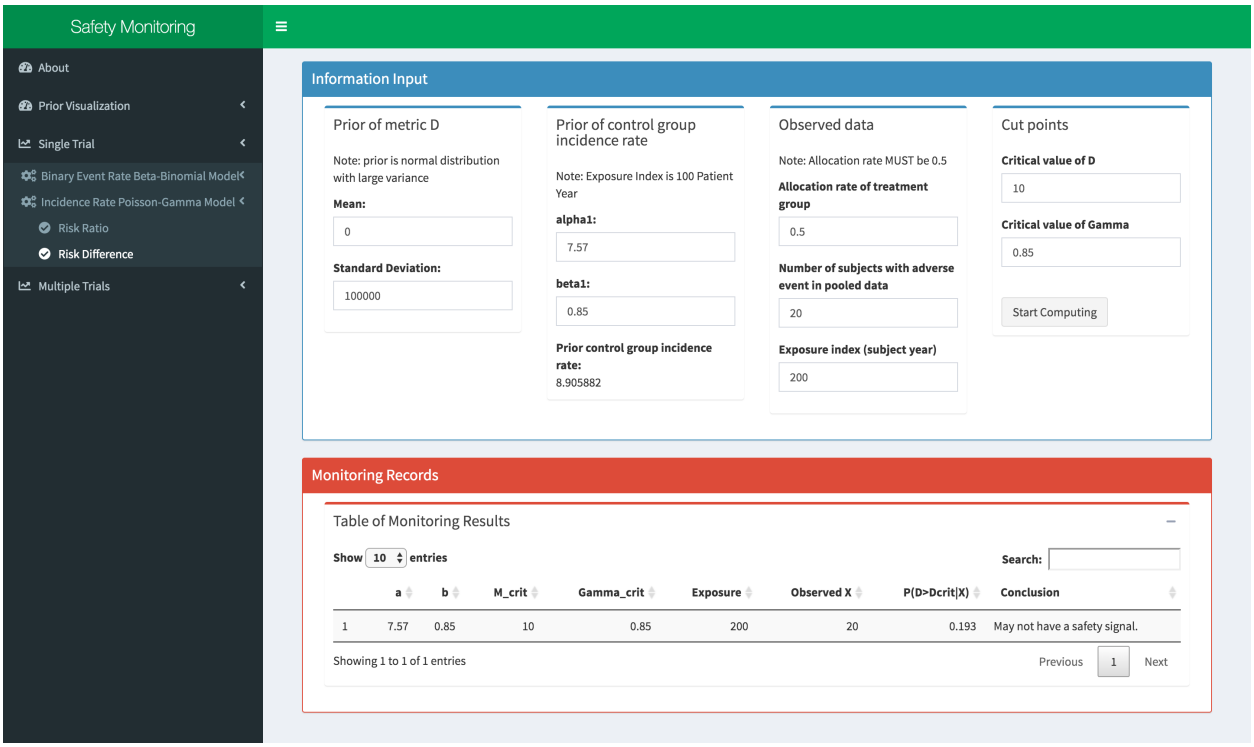

Figure 4.10: "Risk Difference" sub-tab under "Incidence rate - Poisson-Gamma model" tab in the second tool

pooled number of subjects with the AE is greater than a certain number x under three assumptions about the value of control group incidence rate and the metric R or D.

4.5 Discussion

As sponsors develop novel drugs, there is an increasing need for sponsors to detect potential safety signals as soon as possible ahead of IDMC schedule and prompt decisions regarding an unblinded analysis. A few quantitative methods for blinded safety monitoring have been developed. The complex nature of these methods makes the safety monitoring and reporting challenging. This paper describes two interactive tools to accelerate and facilitate the process of blinded safety monitoring and reporting. These two tools were developed using RShiny Dashboard package in R Studio and they are convenient for users to conduct safety monitoring during the ongoing blinded trial.

The interactive tools are only to alert trial sponsors and mangers to a potential safety issue in a principled way without compromising the blinding or integrity of the trial. Unblinding a trial has

many consequences and it would be inadvisable to do so without substantial evidence suggesting that it was necessary. In fact, it may not be necessary to unblind an ongoing trial to satisfy the spirit of the recently issued FDA guidance regarding safety reporting requirements for products still under development FDA (2015).

The performances of the interactive tools rely on two blinded safety monitoring methods. Both methods rely heavily on the assumptions of the prior distribution of event rate or incidence rate of the control group. They may provide biased information for interim safety decisions when the prior distribution is mis-specified.

For the method proposed by Ball (2011), it relies heavily on the assumption that the control rate in the study is similar to the assumed rate based on historical information. It performed reasonably well when the rate in standard therapy is similar to or lower than the assumed rate. But if the actual rate in the standard therapy is higher than the assumed rate, there is high chance to signal an alert even when the experimental group rate is similar or lower than standard therapy. A more stringent decision criterion could reduce the false positive signal, but will also reduce the sensitivity to detect a signal. As a future research, to improve the robustness of the method to departures from control rate assumption, one could consider imposing a prior on the control rate instead of using a fixed number.

Similarly, for the method proposed by Gould & Wang (2017), a finding that would justify unblinding the trial is unlikely unless the prior information provides a very precise estimate of the control group events rate, the trial to which the method would be applied is fairly large or the increase in risk from the test agent is substantial.

It is recommended that the study statisticians engage clinical and safety physicians early to obtain historical information on the safety profile of the control group and discuss how to determine other parameters for decision rules, e.g., parameters for the prior distribution, the cut point for potential safety signals. The team will need to discuss further actions together if the tools suggest potential safety signals.

Chapter 5

Summary and Future Directions

In this dissertation, we have developed two statistical methods to enhance the safety signal detection and an R-Shiny application to accelerate and facilitate the process of blinded drug safety monitoring in a two-arm clinical trial.

Traditional safety analysis focuses on comparing only the AE incidence between two groups, regardless of the severity of adverse event. It is probably of great interest for us to also compare the severity of the AEs, especially when the severities of some AEs are consistently greater for one treatment versus the other. In chapter 2 of this dissertation, from the frequentist perspective, for each AE we propose to use AE occurrence and severity as co-primary endpoints and to perform a statistical test of the composite null hypothesis that the incidence rate and severity are equivalent in the control group and treatment group. The p value of the test of the composite null hypothesis is obtained by combining the p value of the Fisher's exact test for AE incidence and the p value of our proposed test for AE severity. Our test for AE severity is a novel extension of a biased sampling model originally developed for continuous outcome by Gilbert et al. (2003). We conduct simulation studies to investigate the power and type I error rate of the proposed tests of the composite null hypothesis and compare them with the test of equality of AE incidence rate. The simulation results show that, in general, the proposed method performs as well or outperforms the test of equality of AE incidence rate in detecting a safety signal. After obtaining p values for each of the AEs, safety signal detection can be performed by applying a multiple testing procedure to all the p values of to adjust for multiplicity.

Moreover, in chapter 3 of this dissertation, from the Bayesian perspective, we propose a three-level Bayesian hierarchical non-proportional version of the cumulative logit model for detecting safety signal with respect to AE incidence rate and AE severity. Our model allows for testing the equivalence of incidence rate and severity for all the AEs simultaneously between the control group and the treatment group while addressing multiplicities. The simulation results show that, in general, the proposed method not only controls for false discovery rate but also performs well in detecting safety signals when either the incidence rate or the severity is greater in the treatment group.

Finally, due to the complex nature of the current methods for blinded safety monitoring and reporting, we have developed two R Shiny interactive tools to accelerate, facilitate and improve the process of blinded safety monitoring and reporting in chapter 4. The interactive tool can be used to perform computations for posterior probability of the pooled rate or treatment effect metrics and dynamically output the monitoring result about whether the data suggest a safety signal.

There are several topics that attract our attention for future studies.

1. For the proposed frequentist method, The metric r that describes the degree of selection bias is determined after we review the subjects' characteristic information, thus the determination of r is subjective. We may further develop methods (for example Bayesian method) to more accurately and objectively estimate r or $w(z)$ from the data.

2. The proposed Bayesian hierarchical cumulative logit model is based on summary (marginal) data regarding the number of subjects under each of the severity level. In the future research, with data that contains patient level information, we can fit a cumulative logit model with a random effect to account for dependencies among the various AEs at the subject level. Conclusions based on the random effect model could be more precise.

3. In addition, the proposed Bayesian hierarchical approach is applied once to the safety data when the trial is completed. It is of great interest to extend the proposed method to a continuous monitoring framework. Thus another future direction is that the proposed Bayesian hierarchical cumulative logit model can be applied in a group sequential manner for multiple interim analyses of safety events.

4. To accelerate and facilitate the process of safety signal detection, it would also be interest to develop R-Shiny interactive tools based on the methods proposed in this dissertation.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Inc, 2 edition.

Ball, G. (2011). Continuous safety monitoring for randomized controlled clinical trials with blinded treatment information. *Contemporary Clinical Trials*, 32, S11–S17.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & Freitas, R. M. D. (1998). A bayesian neural network method for adverse drug reactionsignal generation. *European Journal of Clinical Pharmacology*, 54, 315–321.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multipletesting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Berry, S. M. & Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60, 418–426.

Bolland, K. & Whitehead, J. (2000). Formal approaches to safety monitoring of clinical trialsin life-threatening conditions. *Statistics in Medicine*, 19, 2899–2917.

Chen, C. & Chaloner, K. (2006). A bayesian stopping rule for a single arm study:with a case study of stem cell transplantation. *Statistics in Medicine*, 25, 2956–2966.

Chen, W., Zhao, N., Qin, G., & Chen, J. (2013). A bayesian group sequential approach to safety signal detection. *Journal of Biopharmaceutical Statistics*, 23, 213–230.

Cohen, A. & Sackrowitz, H. B. (2000). Testing whether treatment is "better" than control with ordered categorical data: definitions and complete class theorem. *Statistics and Decisions*, 18, 1–25.

Cohen, A., Sackrowitz, H. B., & Sackrowitz, M. (2000). Testing whether treatment is 'better' than control with ordered categorical data: an evaluation of new methodology. *Statistics in Medicine*, 19, 2699–2712.

Crowe, B. J., Xia, H. A., Berlin, J. A., Watson, D. J., Shi, H., Lin, S. L., Kuebler, J., Schriver, R. C., Santanello, N. C., Rochester, G., Porter, J. B., Oster, M., Mehrotra, D. V., Li, Z., King, E. C., Harpur, E. S., & Hall, D. B. (2009). Recommendations for safety planning, datacollection, evaluation and reporting during drug,biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. *Clinical Trials*, 6, 430–440.

Diao, G., Liu, G. F., Zeng, D., Wang, W., Tan, X., Heyse, J. F., & Ibrahim, J. G. (2019). Efficient methods for signal detection from correlated adverseevents in clinical trials. *Biometrics*, (pp. 1–9).

Duan, J., Wick, J., Gajewski, B., Mahnken, J., Mayo, M., & Weir, S. (2019a). Assessing the incidence and severity of drug adverse events: A bayeisan hierarchical cumulative logit model. *Submitted to JASA*.

Duan, J., Wick, J., Gajewski, B., Mahnken, J., Mayo, M., & Weir, S. (2019b). Statistical evaluation of causal treatment effect on the incidence and severity of adverse events in clinical trials. *Submitted to Biometrics*.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the fda spontaneousreporting system. *The American Statistician*, 53(3), 177–190.

DuMouchel, W. (2012). Multivariate bayesian logistic regression for analysis of clinical study safety issues. *Statistical Science*, 27(3), 319–339.

Evans, S. J., Waller, P. C., & Davis, S. (2001). Use of proportional reportingratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6), 483–486.

FDA (2010). *Investigational New Drug Safety Reporting Requirements for Human Drug and Biological Products and Safety Reporting Requirements for Bioavailability and Bioequivalence Studies in Humans*.

FDA (2012). *Guidance for Industry and Investigators Safety Reporting Requirements for INDs and BA/BE Studies- Small Entity Compliance Guide*.

FDA (2015). *Safety Assessment for IND Safety Reporting Guidance for Industry*.

Fisher, R. (1932). *Statistical Methods for Research Workers*. Edinburgh and London: Oliver and Boyd, 5 edition.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–29.

Gilbert, P. B., Boschand, R. J., & Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59, 531–541.

Goldman, A. I. & Hannan, P. J. (2001). Optimalcontinuous sequentialboundaries for monitoring-toxicity in clinical trials: a restricted search algorithm. *Statistics in Medicine*, 20, 1575–1589.

Gould, A. L. (2002). Drug safety evaluation in and after clinical trials. *Deming Conference*.

Gould, A. L. (2008). Detecting potential safety issues in clinical trials by bayesian screening. *Biometrical Journal*, 5, 837–851.

Gould, A. L. (2013). Detecting potential safety issues in large clinical or observational trials by bayesian screening when event counts arise from poisson distributions. *Journal of Biopharmaceutical Statistics*, 23, 829–847.

Gould, A. L. (2018). Unified screening for potential elevated adverse event risk and other associations. *Statistics in Medicine*, 37, 2667–2689.

Gould, A. L. & Wang, W. B. (2017). Monitoring potential adverse event rate differences using data from blinded trials: the canary in the coal mine. *Statistics in Medicine*, 36, 92–104.

Hu, N., Huang, L., & Tiwarib, R. C. (2015). Signal detection in fda aers databaseusing dirichlet process. *Statistics in Medicine*, 34, 2725–2742.

Huang, L., Zalkikar, J., & Tiwari, R. (2014). Likelihood ratio based tests forlongitudinal drug safety data. *Statistics in Medicine*, 33, 2408–2424.

Huang, L., Zalkikar, J., & Tiwari, R. C. (2011). A likelihood ratio test based method for signaldetection with application to fda's drug safetydata. *Journal of the American Statistical Association*, 106(496), 1230–1241.

Huang, L., Zalkikar, J., & Tiwari, R. C. (2013). Likelihood ratio test-based method for signal detection in drug classes using fda's aers database. *Journal of Biopharmaceutical Statistics*, 23, 178–200.

Hudgens, M. G., Hoering, A., & Self, S. G. (2003). On the analysis of viral load endpoints in hiv vaccine trials. *Statistics in Medicine*, 22, 2281–2298.

Klingenberg, B., Solari, A., Salmaso, L., & Pesarin, F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, 65, 452–462.

Kulldorff, M., Davis, R. L., Kolczak, M., Lewis, E., TracyLieu, & Platt, R. (2011). A maximized sequential probability ratio test fordrug and vaccine safety surveillance. *Sequential Analysis*, 30(1), 58–78.

L.G. Leon-Novelo, X. Zhou, B. N. B. & Muller, P. (2010). Assessing toxicities in a clinical trial: Bayesian inferencefor ordinal data nested within categories. *Biometrics*, 66, 966–974.

Li, L. & Kulldorff, M. (2010). A conditional maximized sequentialprobability ratio test for pharmacovigilance. *Statistics in Medicine*, 29, 284–295.

Lin, L.-A., Zhan, Y., Li, H., Yuan, S. S., Ball, G., & Wang, W. (2019). Bridging blinded and unblinded analysis for ongoing safety monitoring and evaluation. *Contemporary Clinical Trials*, 83, 81–87.

Lu, X., Mehrotra, D. V., & Shepherd, B. E. (2013). Rank-based principal stratum sensitivity analyses. *Statistics in Medicine*, 32, 4526–4539.

McEvoy, B. W., Nandy, R. R., & Tiwari, R. C. (2013). Bayesian approach for clinical trial safety data using an ising prior. *Biometrics*, 69, 661–672.

McKinley, T. J., Morters, M., & Wood, J. L. N. (2015). Bayesian model choice in cumulative link ordinal regression models. *Bayesian Analysis*, 10(1), 1–30.

Mehrotra, D. V. & Adewale, A. J. (2012). Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Statistics in Medicine*, 31, 1918–1930.

Mehrotra, D. V. & Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13, 227–238.

Mehrotra, D. V., Li, X., & Gilbert, P. B. (2006). A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept hiv vaccine trial. *Biometrics*, 62, 893–900.

Mukhopadhyay, S., Waterhouse, B., & Hartford, A. (2018). Bayesian detection of potential risk using inference onblinded safety data. *Pharmaceutical Statistics*, 17, 823–834.

Ryu, E. & Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, 27, 1703–1717.

Schnell, P. M. & Ball, G. (2016). A bayesian exposure-time method forclinical trial safety monitoring with blinded data. *Therapeutic Innovation & Regulatory Science*, 50(6), 833–838.

Shih, M.-C., Lai, T. L., Heysed, J. F., & Chene, J. (2010). Sequential generalized likelihood ratio tests forvaccine safety evaluation. *Statistics in Medicine*, 29, 2698–2780.

Shih, W. J. & Quan, H. (1997). Testing for treatment differences with dropouts present in clinical trials, a composite approach. *Statistics in Medicine*, 16, 1225–1239.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3), 751–754.

Thall, P. F. & Simon, R. (1994). Practical bayesian guidelines for phase iib clinical trials. *Biometrics*, 50(2), 337–349.

Vargha, A. & Delaney, H. D. (1998). The kruskal-wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, 59, 137–142.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16, 117–186.

Wang, W., Whalen, E., Munsaka, M., Li, J., Fries, M., Kracht, K., Sanchez-Kam, M., Singh, K., & Zhou, K. (2017). On quantitative methods for clinical safety monitoring in drug development. *Statistics in Biopharmaceutical Research*, 10(2), 85–97.

Xia, H. A., Ma, H., & Carlin, B. P. (2011). Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics*, 21, 1006–1029.

Yao, B., Zhu, L., Jiang, Q., & Xia, H. A. (2013). Safety monitoring in clinical trials. *Pharmaceutics*, 5, 94–106.

Zhu, L., Yao, B., Xia, H. A., & Jiang, Q. (2016). Statistical monitoring of safety in clinical trials. *Statistics in Biopharmaceutical Research*, 8(1).

Appendix A

Some Derivations for Chapter 4

A.1 Single binomial trial

The likelihood of the single binomial trial:

$$
\begin{aligned}
L &= f_{binom}(x_T; N\theta, p_T) f_{binom}(x_C; N(1-\theta), p_C) \\
&= \binom{N\theta}{x_T} p_T^{x_T} (1-p_T)^{N\theta - x_T} \binom{N(1-\theta)}{x_C} p_C^{x_C} (1-p_C)^{N(1-\theta)-x_C} \\
&= \binom{N\theta}{x_T} p_T^{x_T} (1-p_T)^{N\theta} (1-p_T)^{-x_T} \binom{N(1-\theta)}{x_C} p_C^{x_C} (1-p_C)^{N(1-\theta)} (1-p_C)^{-x_C} \\
&= \binom{N\theta}{x_T} \binom{N(1-\theta)}{X-x_T} \Big(\frac{1-p_T}{1-p_C}\Big)^{N\theta} \Big(\frac{p_T}{1-p_T}\Big)^{x_T} p_C^{X-x_T} (1-p_C)^{N} (1-p_C)^{x_T - X} \\
&= \binom{N\theta}{x_T} \binom{N(1-\theta)}{X-x_T} \Big(\frac{1-p_T}{1-p_C}\Big)^{N\theta} \Big(\frac{p_T}{1-p_T}\Big)^{x_T} p_C^{X-x_T} (1-p_C)^{x_T} (1-p_C)^{N-X} \\
&= \binom{N\theta}{x_T} \binom{N(1-\theta)}{X-x_T} \Big(\frac{1-p_T}{1-p_C}\Big)^{N\theta} \Big(\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\Big)^{x_T} p_C^{X} (1-p_C)^{N-X} \\
&= \binom{N}{X} p_C^X (1-p_C)^{N-X} \frac{1}{\binom{N}{X}} \binom{N\theta}{x_T} \binom{N(1-\theta)}{X-x_T} \Big(\frac{1-p_T}{1-p_C}\Big)^{N\theta} \Big(\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\Big)^{x_T} \\
&= f_{binom}(X; N, p_C) C() \Big(\frac{1-p_T}{1-p_C}\Big)^{N\theta} \Big(\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\Big)^{x_T}
\end{aligned}
$$

Three reparameterizations:

$$p_T = R p_C$$

$$p_T = D + p_C$$

$$Q = \frac{p_T(1-p_C)}{(1-p_T)p_C}$$

By plugging into the above likelihood with p_T represented by p_C, we can get three likelihoods.

A.2 Single Poisson trial

If the total sample size is large and the event incidence is low, then the number of adverse events approximately follows Poisson distribution with exposure rates $N\theta p_T$ and $(N-N\theta)p_C$ respectively in treatment group and control group. Thus:

$$\frac{\lambda_T}{\lambda_C} = \frac{N\theta p_T}{(N-N\theta)p_C} = \frac{\theta}{1-\theta}\frac{p_T}{p_C} = \xi R$$

The likelihood of single Poisson trial:

$$\begin{aligned}
L &= f_{Poiss}(x_T;\lambda_T) f_{Poiss}(x_C;\lambda_C) \\
&= \frac{\lambda_T^{x_T}}{x_T!}e^{-\lambda_T}\frac{\lambda_C^{x_C}}{x_C!}e^{-\lambda_C} \\
&= \frac{(\lambda_C\xi R)^{X-x_C}}{(X-x_C)!}e^{-\lambda_C\xi R}\frac{\lambda_C^{x_C}}{x_C!}e^{-\lambda_C} \\
&= \frac{\lambda_C^X\lambda_C^{-x_C}(\xi R)^X(\xi R)^{-x_C}}{(X-x_C)!}e^{-\lambda_C\xi R}\frac{\lambda_C^{x_C}}{x_C!}e^{-\lambda_C} \\
&= \frac{\lambda_C^X}{X!}X!\frac{\lambda_C^{-x_C}(\xi R)^X(\xi R)^{-x_C}}{(X-x_C)!}e^{-\lambda_C\xi R}\frac{\lambda_C^{x_C}}{x_C!}e^{-\lambda_C} \\
&= \frac{\lambda_C^X(1+\xi R)^X}{X!}\frac{1}{(1+\xi R)^X}\frac{X!}{(X-x_C)!x_C!}\lambda_C^{-x_C}(\xi R)^X(\xi R)^{-x_C}e^{-\lambda_C(1+\xi R)}\lambda_C^{x_C} \\
&= \frac{[\lambda_C(1+\xi R)]^X}{X!}e^{-\lambda_C(1+\xi R)}\binom{X}{x_T}\frac{1}{(1+\xi R)^X}\lambda_C^{-x_C}(\xi R)^X(\xi R)^{-x_C}\lambda_C^{x_C} \\
&= \frac{[\lambda_C(1+\xi R)]^X}{X!}e^{-\lambda_C(1+\xi R)}\binom{X}{x_T}\frac{1}{(1+\xi R)^X}(\xi R)^{X-x_C} \\
&= f_{Poiss}(X;\lambda_C(1+\xi R))\binom{X}{x_T}\frac{1}{(1+\xi R)^X}(\xi R)^{X-x_C} \\
&= f_{Poiss}(X;\lambda_C(1+\xi R))\binom{X}{x_T}\frac{1}{(1+\xi R)^{X-x_C}}\frac{1}{(1+\xi R)^{x_C}}(\xi R)^{X-x_C} \\
&= f_{Poiss}(X;\lambda_C(1+\xi R))\binom{X}{x_T}\left(\frac{\xi R}{1+\xi R}\right)^{X-x_C}\left(\frac{1}{1+\xi R}\right)^{x_C} \\
&= f_{Poiss}(X;\lambda_C(1+\xi R))\binom{X}{x_T}q^{X-x_C}(1-q)^{x_C} \\
&= f_{Poiss}(X;\lambda_C(1+\xi R))\binom{X}{x_T}q^{x_T}(1-q)^{X-x_T}
\end{aligned}$$