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

Jiawei Duan* Byron J. Gajewski[†] Matthew S. Mayo[‡] Scott Weir[§]
Jo A. Wick[¶]

Abstract
Clinical safety data 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 the incidence rate and severity are equivalent between groups. The p value of the test of the composite null hypothesis is obtained by combining the p values of the Fisher's exact test for AE incidence and the test for AE severity. The test for AE severity is based on a biased sampling model, which is an extension of the work by Gilbert et al. (2003, Biometrics 59, 531-541) to ordinal response. 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.

Key Words: Adverse events; Causal inference; Composite null hypothesis; Posttreatment selection bias; Principal stratification; Safety monitoring; Severity.

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, FDA, 2012, FDA, 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 improves the timing of identifying risks and justify the safety of the treatment that warrant a next stage clinical trial or regulatory agency approval.

Drug safety evaluation include 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

[*] Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS
[†] Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS
[‡] Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS
[§] Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center, Kansas City, KS
[¶] Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS

"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, 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.

As we can see from the literatures, 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. For example, suppose the severity of an adverse event has three levels: mild, moderate or severe. The probabilities that the severity of an adverse event is moderate or severe are both higher for one treatment versus the other even if the incidence rates of an adverse event are the same in both groups. 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 carry 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., 2019] 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 composite 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] and [Mehrotra et al., 2006] 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, we introduce notation and define the composite null hypothesis. In Section 3, we describe the combined test for testing the composite null hypothesis. In Section 4, we introduce a proposed method for adjusting for selection bias. In Section 5, we compare the power of different tests in a comprehensive simulation study and then in Section 6 we apply the proposed method in a clinical trial safety data. We conclude the article in Section 7.

2. Notations and Composite Null

Suppose the safety evaluation is performed in a two-arm trial for a drug: a control arm and a treatment arm. Our goal is to detect those AEs with safety signals among all the AEs. An AE has a safety signal if it has greater incidence rate or greater severity in the treatment arm. Greater AE severity will be defined later.

To establish the safety of the drug, two primary endpoints will be used for each AE: AE occurrence and AE severity. AE occurrence is a binary endpoint, indicating whether an AE occurs or not. AE severity is an ordinal categorical endpoint. 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 in 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 $y_{1i} = 1(y_{2i} = 1)$. Also denote $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ as two vectors of severity scores of the subjects with adverse event 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

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

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

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

Denote $F_c(z)$ and $F_t(z)$ $(z = 1, 2, 3)$ as the cumulative density functions of the severity score of 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.

Greater AE severity is defined as follows: The AE severity is greater in the treatment group than that in the control group if $F_c(z) > F_t(z), z = 1, 2$, or equivalently, if $1 - F_c(z) < 1 - F_t(z), x = 1, 2$. 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 AE 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.

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 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.

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

The non-model based methods for testing $H_0^{(1)}$ is the one-sided Fisher's exact test. Denote p_1 as the p value for one-tailed test of

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

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 endpoints 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 proposed a test for comparing the severity score $\boldsymbol{z}_1, \boldsymbol{z}_2$ of the subjects who experience the AE. The test is based on a biased sampling model proposed by [Gilbert et al., 2003]. Details about the test will be introduced in Section 4. Denote p_2 as the p value of the test:

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

3.2 Methods for Combining Separate Tests

We consider the following two methods for testing the composite null hypothesis of an AE at one-sided level α based on a combination of the p-values (p_1, p_2) introduced above. 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 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 test to be superior than Simes test.

4. Test for Causal Treatment Effect on AE Severity

In this section we will introduce a biased sampling model and how we use it to test the causal treatment effect on the AE severity, 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.

4.1 Biased Sampling Model

Following [Gilbert et al., 2003] and [Mehrotra et al., 2006], 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 $F_{(c)}^{alw.}(z)$ and for subjects in the set $\{Z_i(t) : Y_i(c) = Y_i(t) = 1\}$, suppose $Z_i(t)$ are identically distributed as $F_{(t)}^{alw.}(z)$, also denote $f_{(c)}^{alw.}(z)$ and $f_{(t)}^{alw.}(z)$ as the probability mass function that corresponds to $F_{(c)}^{alw.}(z)$ and $F_{(t)}^{alw.}(z)$ respectively. Then any functional that measures a contrast of the distributions

$$\begin{aligned} F_{(c)}^{alw.}(z) &= Pr(Z_i(c) \leq z | Y_i(c) = Y_i(t) = 1) \quad \textit{and} \\ F_{(t)}^{alw.}(z) &= Pr(Z_i(t) \leq z | Y_i(c) = Y_i(t) = 1) \end{aligned}$$

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 ($H_0^{(2)} : F_c(z) = F_t(z)$), we compare $F_{(c)}^{alw.}(z)$ and $F_{(t)}^{alw.}(z)$. Or equivalently, to compare $f_{(c)}^{alw.}(z)$ and $f_{(t)}^{alw.}(z)$. The second null hypothesis can thus be rewritten as $H_0^{(2)} : F_{(c)}^{alw.}(z) = F_{(t)}^{alw.}(z)$.

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 subject 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 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.

Table 1: Principal Stratum

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

From Table 1 we know $F_{(c)}^{alw.}(z) = F_{(c)}(z)$, or equivalently, $f_{(c)}^{alw.}(z) = f_{(c)}(z)$. Thus $F_{(c)}^{alw.}(z)$ is identified from the observed data. $F_{(t)}^{alw.}(z)$ cannot be identified by the above assumptions. However, from Table 1 we know the subjects who experience the AE in the treatment group consists of the subjects who will always experience 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]:

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

It can be proved 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.

Thus

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

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 $f_{(t)}^{alw\cdot}(z)$ and $F_{(t)}^{alw\cdot}(z)$ are

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

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

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

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 $H_0^{(2)} : F_{(c)}^{alw\cdot}(z) = F_{(t)}^{alw\cdot}(z)$, the corresponding alternative hypothesis of interest is: $H_0^{(2)} : F_{(c)}^{alw\cdot}(z) > F_{(t)}^{alw\cdot}(z)$. This means that the severity endpoint of the subjects in the treatment group is stochastically larger than that in the control group. Thus the composite null hypothesis can be rewritten as

$$H_0^{(1)} : \theta_1 = \theta_2 (RR = 1) \quad \textit{and} \quad H_0^{(2)} : F_{(c)}^{alw\cdot}(z) = F_{(t)}^{alw\cdot}(z)$$

We obtain the p value (p_1) for testing $H_0^{(1)}$ using Fisher's exact and we obtain the p value ($p_{2,r}$) for testing $H_0^{(2)}$ using a proposed test to be introduced in this section. Simes's method and Fisher's method are then used to combine p_1 and $p_{2,r}$. The combined test using Simes's method and Fisher's method are referred to as Simes test and Fisher test respectively.

To test the second null hypothesis $H_0^{(2)} : F_{(c)}^{alw\cdot}(z) = F_{(t)}^{alw\cdot}(z)$, 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 $F_{(t)}^{alw\cdot}(z)$. $(\boldsymbol{n}_{2,r})$ 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.

4.3 Bootstrap Resampling

Because the data we obtained $(\boldsymbol{z}_1, \boldsymbol{z}_{2,r})$ are not exactly from the distributions $F^{alw\cdot}_{(c)}(z)$ and $F^{alw\cdot}_{(t)}(z)$ ($\boldsymbol{z}_{2,r}$ is the estimated data from $F^{alw\cdot}_{(t)}(z)$), 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 $H_0^{(2)} : F^{alw\cdot}_{(c)}(z) = F^{alw\cdot}_{(t)}(z)$. The p-value based on T_r, denoted by p_2, r 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 2.5.2. 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

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 was generated in two steps: in the first step, the number of subjects who experience the AE in each group was generated and then the number of subjects who experience the AE with severity outcomes classified into each severity level was 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 generated a random variable from Bernoulli distribution with parameter θ_1 for the control group, and then we generated a random variable from Bernoulli distribution with parameter θ_2 for the treatment group. We continued generating Bernoulli random variable like this for each group until the summation of the Bernoulli random variables generated in control group (x_1) plus that in treatment group (x_2) is at least x. x was given in advance. The reason why we fix x is to investigate how the power changes as we increase x, which can be directly observed from the data, and with what sample size does the proposed tests perform as well as or outperform the Fisher's exact test for incidence rate. In the second step, the number of subjects who experience the AE with severity outcomes classified into each of the 3 severity levels in the control group $(\boldsymbol{n}_1)$ was generated from a multinomial distribution with parameter x_1 and $(f_{(c)}(1), f_{(c)}(2), f_{(c)}(3))$ and the number of subjects under each AE severity level in the treatment group $(\boldsymbol{n}_2)$ was 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 $f_{(c)}^{alw\cdot}(z)$ and $f_{(t)}^{alw\cdot}(z)$ and the true degree of selection bias r_{true}. Thus we set true values for $f_{(c)}^{alw\cdot}(z)$ and $f_{(t)}^{alw\cdot}(z)$ respectively and obtained the true values of $f_{(c)}(z)$ and $f_{(t)}(z)$ by transforming $f_{(c)}^{alw\cdot}$ and $f_{(t)}^{alw\cdot}$ according to the equations introduced in Section 4 as follows, $f_{(t)}(z) = \frac{f_{(t)}^{alw\cdot}(z)}{RR \times w(z|\gamma, r_{true})}, z = 1, 2, 3$. $RR = \theta_2/\theta_1, w(1|\gamma, r_{true}) = \gamma, w(2|\gamma, r_{true}) = \gamma r_{true}, w(3|\gamma, r_{true}) = \gamma r_{true}^2$. 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 $(f_{(c)}^{alw\cdot}(1), f_{(c)}^{alw\cdot}(2), f_{(c)}^{alw\cdot}(3))$. We considered 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, $f_{(c)}^{alw\cdot} = (0.6, 0.3, 0.1)^T$ and $f_{(t)}^{alw\cdot} = (0.5, 0.3, 0.2)^T$, $(0.4, 0.2, 0.4)^T$ or $(0.3, 0.2, 0.5)^T$. To measure the true difference between θ_1 and θ_2, we introduce odds ratio $OR = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}$ as an effect size measure and to measure the true difference between $f_{(c)}^{alw\cdot}$ and $f_{(t)}^{alw\cdot}$ (or equivalently $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$), we use a 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

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

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: Scenarios for simulation study

Scenario	θ_1	$\boldsymbol{\pi}_1(f_{(c)}^{alw\cdot}(z))$	θ_2	$\boldsymbol{\pi}_2(f_{(t)}^{alw\cdot}(z))$	OR	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	$(0.6, 0.3, 0.1)^T$	2.11	0.5
5	0.05	$(0.6, 0.3, 0.1)^T$	0.1	$(0.5, 0.3, 0.2)^T$	2.11	0.565
6	0.05	$(0.6, 0.3, 0.1)^T$	0.1	$(0.4, 0.2, 0.4)^T$	2.11	0.65

For each of 500 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 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 $OR = 1$ 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 $OR = 1$, i.e., $\theta_1 = \theta_2$, according to Section 4, there is no selection bias, so r_{true} always has to be 1. The 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 were also included in the table. The estimated type I error rates (4th and 5th column) that correspond to different tests were 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 decrease accordingly 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, 6th and 9th column) 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, the estimated power (6th and 9th column) decrease accordingly.

Table 3: Type I error rate $\times 100\%$ and power $\times 100\%$ of 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), given that $x = 50, 100$, $OR = 1, g = 0.5, 0.565, 0.65$ with $\theta_1 = 0.05, 0.1$ and $\boldsymbol{\pi}_1 = (0.6, 0.3, 0.1)^T$

Method	True	Presumed	$OR = 1, g = 0.5$		$OR = 1, g = 0.565$		$OR = 1, g = 0.65$	
	r	r	50	100	50	100	50	100
$\boldsymbol{\theta_1 = 0.05}$								
FET	n/a	n/a	3	4.8	4.2	3.6	2.8	3.4
WT	1	n/a	6.6	6.6	7.2	5	5.8	5
	1	1.25	6.2	10.2	35.8	46.6	74.8	94.6
SEV	1	1	3.8	4	23.6	33.2	63.2	90.8
	1	0.8	3	2.6	16	23	48.8	72.4
	1	1.25	5.4 (7)	7.4 (9.4)	27.8 (27.2)	36.8 (34.4)	66.4 (60)	90.2 (85.2)
PS(PF)	1	1	4.2 (4.6)	3 (3.4)	17.6 (16.2)	24.6 (24)	53.6 (47.2)	84.6 (82.2)
	1	0.8	3.6 (2.4)	2.2 (2)	11 (9.6)	17 (11)	40.2 (37)	63 (63.6)

Table 4 shows the estimated power of the proposed test for AE severity (SEV) and the proposed test for the composite null based on Simes' method (PS) or Fisher's method (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 $OR = 2$ 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 the traditionally used Fisher's exact test (FET) for incidence rate and the Wilcoxon rank sum test for stochastic order of 4 level AE toxicity endpoint (WT) under different parameter configurations were also included in the table. As long as the true amount of selection bias is specified, 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, 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 of 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 decreases. 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 conservatively presumes $r = 1.25$, the power increases. Thus making a conservative assumption of selection bias can cause certain degree of power loss and 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 increases. 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.

6. Application

We applied 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. 1166 patients with stage I NSCLC were randomly assigned to receive either placebo or isotretinoin (30 mg/day) for 3 years. 589 patients received isotretinoin while the remaining patients received placebo.

The safety data collected from the trial (shown in Table 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 combined the last two severity levels into one in our analysis.

We first conducted 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 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 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 excluded it from our safety analysis.

Figure 1 shows how p values of the following tests change as we change the degree of

Table 4: Power $\times 100\%$ of 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), given that $x = 50, 100$, $OR = 2, g = 0.5, 0.565, 0.65$ with $\theta_1 = 0.05, 0.1$ and $\boldsymbol{\pi}_1 = (0.6, 0.3, 0.1)^T$

Method	True	Presumed	$OR = 2.11, g = 0.5$		$OR = 2.11, g = 0.565$		$OR = 2.11, g = 0.65$	
	r	r	50	100	50	100	50	100
				$\boldsymbol{\theta_1 = 0.05}$				
FET	n/a	n/a	73.8	96.6	73.8	96.6	73.8	96.6
	1.25	n/a	80.2	97	81.8	97.6	82	97.2
WT	1	n/a	82.2	98.4	81.2	96.2	83.4	96.6
	0.8	n/a	75.4	98.4	80.2	98.2	80.8	98.2
	1.25	1.25	4.4	5.4	17.8	26.2	58.8	85.6
	1.25	1	2	1	7.8	7.2	29.8	57.2
	1.25	0.8	0	0.4	1.8	1.4	10	17.4
	1	1.25	13.4	18.4	40.2	59	88.2	99.2
SEV	1	1	4.8	8.4	19.2	32.8	64.6	87.8
	1	0.8	1.6	1.6	6.4	10	31.2	51.2
	0.8	1.25	24.4	40.4	66	89	94.6	100
	0.8	1	13.2	15.8	42.4	66.4	84.6	98
	0.8	0.8	4.2	2.6	20.8	28.8	61.4	85.2
	1.25	1.25	61.2 (61.6)	95 (92.8)	67.4 (73.6)	96 (96.8)	80 (91.2)	98.8 (98.6)
	1.25	1	60.6 (57)	94.8 (89.4)	64.4 (65.8)	95.2 (93.4)	68.6 (80)	96.8 (98.2)
	1.25	0.8	60.2 (52.6)	94.8 (87.8)	63 (60)	95.2 (90.6)	63.2 (98.2)	93.2 (93.8)
	1	1.25	64.2 (69.8)	96.4 (95.8)	73.8 (83.2)	97.6 (98.8)	95.2 (97.6)	100 (100)
PS(PF)	1	1	62.4 (64.2)	96 (93.4)	66.6 (75.6)	96 (96.2)	85.8 (93.4)	99 (100)
	1	0.8	61.8 (58.8)	96 (91.6)	62 (65.4)	94.4 (93.2)	70.2 (85.4)	95.6 (97.8)
	0.8	1.25	64.8 (72)	95.8 (96.2)	82.2 (92)	99.4 (99.8)	98.4 (99.6)	100 (100)
	0.8	1	60.8 (64.2)	94.4 (94.4)	75.8 (84.2)	98.4 (99)	93.4 (96.8)	100 (100)
	0.8	0.8	57.8 (58.4)	94 (91.2)	66.8 (76.2)	96.8 (97.8)	81.6 (91.4)	99.4 (100)

selection bias (either in favor of flagging the AE or of not flagging the AE) for each adverse event: 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

Table 5: Toxicity frequency for randomized eligible patients by study arms. In the placebo (isotretinoin) group, 171 (427) of 577 (589) patients exhibited some type of toxicity. The proportion of patients in the study arm belonging to the cell is given in the parenthesis

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 6: P values and adjusted p values using Hochberg procedure of the Fisher's exact test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$

	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

Holmes, Hochberg and Benjamini-Hochberg procedure being used to adjust for multiplicity.

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 the AE incidence rate and severity are equivalent. 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

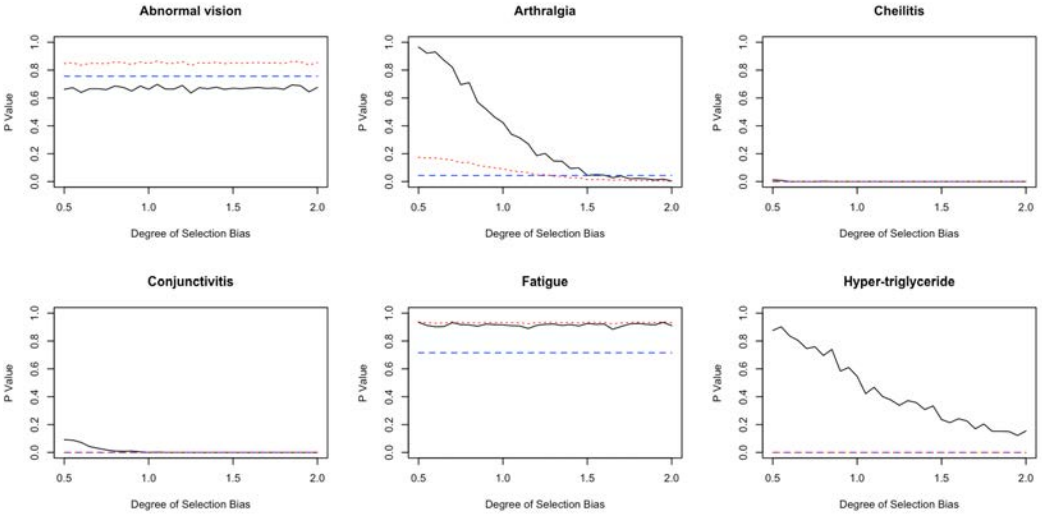

Figure 1: Plot of p values of the following tests versus the degree of selection bias for each adverse event: Fisher's exact test, the proposed test for severity, the proposed test for incidence and severity using Fisher's method. The red dotted line represents the p value of the proposed test for incidence and severity using Fisher's method. The black solid line represents the p value of the proposed test for severity. The blue dashed line represents the p value of the Fisher's exact test of $\theta_1 \leq \theta_2$ vs $\theta_1 > \theta_2$

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 the adjusted and unadjusted severity scores of subjects in the treatment and control group respectively.

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 when 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.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute Cancer Center Support Grant P30 CA168524.

References

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

[Benjamini & Hochberg, 1995] 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 & Berry, 2004] Berry, S. M. & Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60, 418–426.

[Cohen & Sackrowitz, 2000] 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 et al., 2000] 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.

[Duan et al., 2019] Duan, J., Wick, J., Gajewski, B., Mahnken, J., Mayo, M., & Weir, S. (2019). Statistical monitoring of causal treatment eect on the incidence andseverity of adverse events in clinical trials. *Submitted to Biometrics*.

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

[FDA, 2010] 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] FDA (2012). *Guidance for Industry and Investigators Safety Reporting Requirements for INDs and BA/BE Studies- Small Entity Compliance Guide*.

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

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

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

[Gilbert et al., 2003] 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.

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

[Gould, 2013] 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, 2018] Gould, A. L. (2018). Unified screening for potential elevated adverse event risk and other associations. *Statistics in Medicine*, 37, 2667–2689.

[Hudgens et al., 2003] 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 et al., 2009] Klingenberg, B., Solari, A., Salmaso, L., & Pesarin, F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, 65, 452–462.

[L.G. Leon-Novelo & Muller, 2010] 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.

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

[Mehrotra & Adewale, 2012] 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 & Heyse, 2004] 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 et al., 2006] 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.

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

[Shih & Quan, 1997] 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, 1986] Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3), 751–754.

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

[Xia et al., 2011] 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.