Efficient methods for signal detection from correlated adverse events in clinical trials

Guoqing Diao[1] | **Guanghan F. Liu**[2] | **Donglin Zeng**[3] | **William Wang**[2] | **Xianming Tan**[3] | **Joseph F. Heyse**[2] | **Joseph G. Ibrahim**[3]

[1]Department of Statistics, George Mason University, Fairfax, Virginia

[2]Merck & Co., Inc., North Wales, Pennsylvania

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Correspondence
Guoqing Diao, Department of Statistics, George Mason University, Fairfax, VA
Email: gdiao@gmu.edu

Abstract

It is an important and yet challenging task to identify true signals from many adverse events that may be reported during the course of a clinical trial. One unique feature of drug safety data from clinical trials, unlike data from post-marketing spontaneous reporting, is that many types of adverse events are reported by only very few patients leading to rare events. Due to the limited study size, the p-values of testing whether the rate is higher in the treatment group across all types of adverse events are in general not uniformly distributed under the null hypothesis that there is no difference between the treatment group and the placebo group. A consequence is that typically fewer than 100α percent of the hypotheses are rejected under the null at the nominal significance level of α. The other challenge is multiplicity control. Adverse events from the same body system may be correlated. There may also be correlations between adverse events from different body systems. To tackle these challenging issues, we develop Monte-Carlo-based methods for the signal identification from patient-reported adverse events in clinical trials. The proposed methodologies account for the rare events and arbitrary correlation structures among adverse events within and/or between body systems. Extensive simulation studies demonstrate that the proposed method can accurately control the family-wise error rate and is more powerful than existing methods under many practical situations. Application to two real examples is provided.

KEYWORDS

family-wise error rate, multiplier bootstrap, permutation test, rademacher sequence, score test

1 | INTRODUCTION

Assuring the safety of human drug products has been a key component of the mission of the Food and Drug Administration (FDA). Drug safety is typically studied and evaluated in clinical trials. Across the clinical development program, it is a continuous effort and mandatory to collect safety data including patient-reported adverse events (AEs). For example, serious cardiac AEs are closely monitored for investigative diabetes drugs (Food and Drug Administration, 2008); or serious vaccine related AEs are monitored throughout the study for vaccines (Mehrotra and Heyse, 2004a). Patient-reported AEs are often grouped into different body systems (BSs) or system organ classes (SOCs). Examples of common SOCs include cardiac disorders, gastrointestinal disorders, infections and infestations, nervous systems, renal and urinary, and respiratory systems, etc. The detailed SOC listing is maintained by Medical Dictionary for Regulatory Activities

(MedDRA, 2018). An important task in drug safety analysis is to identify true AEs from many possible AEs that may be reported (Jiang and Snapinn, 2016; Wang et al., 2018).

To detect drug safety signals, two types of data arise in practice: data from spontaneous reporting (SR) and data from clinical trials. In spontaneous reporting, health professionals or patients report suspected harm from a medicine to their local or national drug administration. Such data are often stored in some databases. For example, the Vaccine Adverse Event Reporting System (VAERS), which is co-sponsored by the FDA and the Centers for Disease Control and Prevention (CDC), collects AEs that occur after the administration of US licensed vaccines. The FDA Adverse Event Reporting System (FAERS) is a database that contains AE reports, medication error reports and product quality complaints resulting in AEs that were submitted to FDA. Data from such databases include many different drugs and many different types of AEs across different SOCs. Various methods have been developed for safety signal detection (or detection of drug-AE pairs) using data from SR; for example, proportional reporting ratios (Evans et al., 2001), reporting odds ratios (Rothman et al., 2004), the likelihood ratio tests (Huang et al., 2014, 2011, 2013, 2017; Nam et al., 2017; Zhao et al., 2018), and Bayesian methods (Bate et al., 1998; DuMouchel, 1999; DuMouchel and Pregibon, 2001; Norén et al., 2006; Hu et al., 2015).

On the other hand, in clinical trials, patients are typically randomly allocated to a treatment group and a placebo (control) group. During the course of the clinical trials, adverse events will be reported to the investigators. The interest is to know whether the risk of an AE is higher in the treatment group than the placebo group. Classical statistical tests such as Fisher's exact test, Pearson's chi-square test, and chi-square test for comparison of event rates can be used to perform hypothesis testing for each possible AE (Miettinen and Nurminen, 1985). More recently, Bayesian hierarchical mixture models and their variations have been developed for detecting safety signals in clinical trials (Berry and Berry, 2004; Xia et al., 2011, DuMouchel, 2012; Price et al., 2014; Odani et al., 2017).

An important issue in drug safety signal detection is multiplicity. Failing to adjust for multiple testing can yield an abundance of false positive results. The Bonferroni correction can control the family-wise error rate (FWER); however this procedure assumes that all AE terms are mutually independent and therefore can be overly conservative. Benjamini and Hochberg (1995) introduced the concept of false discovery rate (FDR) and the so-called BH procedure to control FDR based on sequential Bonferroni-type adjustment. Several variations of FDR have been proposed in literature, see for example, pFDR (Storey, 2002) and DFDR (Mehrotra and Heyse, 2004b; Mehrotra and Adewale, 2012). FDR-based methods for drug safety signal detection have been discussed in Ahmed et al. (2010, 2012). Huang et al. (2011) proposed likelihood ratio test methods to handle the multiplicity issue using gate keep procedures and Monte Carlo simulations. Chen et al. (2015) compared the performance of various methods for SR data and clinical trial data through simulation studies.

Essentially none of the above procedures accounts for possible correlations among the risks of AEs within the same SOC or across different SOCs. To account for such correlations, one can incorporate correlations by using random effects without specifying the correlation structure; however, inference of the unknown parameters under the random effects model involves high-dimensional integrals and the computation is intensive especially when the total number of AE terms is large. An alternative approach is the permutation resampling approach by shuffling the treatment assignments. However, the permutation approach is computationally demanding since repeated analysis over the permuted data sets is needed. The computation can be infeasible when the number of hypotheses is large and the analysis of each data set is time-consuming. Furthermore, the permutation approach requires complete exchangeability under the null hypothesis, which may not be satisfied in practice.

Furthermore, unlike data from SR, in clinical trials for many AE terms only very few patients experience the events, leading to rare events. As is shown in the Web Appendix A, due to the limited study size, the p-values of testing whether the rate is higher in the treatment group across all AE terms are in general not uniformly distributed between 0 and 1 under the null hypothesis. Particularly, with rare events, the distribution of the p-values under the null hypothesis is highly left skewed. Consequently, fewer than 100α percent of the hypotheses are rejected under the null at the nominal significance level of α and methods using large sample approximations may not perform well.

Recently, in the context of genome-wise associate studies (GWAS), Diao and Vidyashankar (2013) and Diao et al. (2014) developed Monte-Carlo-based procedures to account for multiplicity in the identification of genes that impact a certain phenotype. These procedures allow arbitrary correlations among the single nucleotide polymorphisms (SNPs) across the whole genome and are shown to be substantially more powerful than other existing methods. Unlike in GWAS where there are high-dimensional covariates (i.e., SNPs), in the drug safety analysis, we have high-dimensional outcomes from different types of AEs while there is only one covariate involved (i.e., the treatment indicator). In this paper, we develop Monte-Carlo-based procedures for the detection of drug safety signals which can account for both rare events and multiplicity.

The rest of the paper is organized as follows. In Section 2, we describe the Monte-Carlo procedure for signal detection. Extensive simulation studies examining the finite-sample performance of the proposed method are conducted in Section

3. Applications to two real examples are provided in Section 4. We conclude the paper with some discussions and future research in Section 5.

2 | METHODS

Suppose that there are B BSs, with k_b AE terms for BS b. The jth type within BS b is labeled as A_{bj}, $b = 1, \ldots, B$ and $j = 1, \ldots, k_b$. Suppose that there are n independent subjects, n_0 of which are in the control group and the remaining $n_1 = n - n_0$ subjects are in the treatment group. For $i = 1, \ldots, n$, we define X_i as the treatment indicator taking value 0 for the control group and 1 for the treatment group, and $Y_{ibj} = 1$ ($b = 1, \ldots, B$; $j = 1, \ldots, k_b$) if subject i experiences the jth type AE in the bth BS and 0 otherwise. We assume that the time at risk for all patients are the same although this assumption can be relaxed; some discussions are provided in Section 5.

Our objective is to detect the AEs that have higher rates in the treatment group than in the placebo group. Multiple comparisons arise as there are $M = \sum_{b=1}^{B} k_b$ AE terms and the occurrences of certain AE terms may be correlated. Standard methods may lead to inflated type I error rates without appropriate multiple comparison adjustment. On the other hand, Bonferroni correction is known to be conservative in the presence of correlations.

We consider the following logistic regression model for each AE:

$$\log \frac{P(Y_{ibj} = 1|X_i)}{P(Y_{ibj} = 0|X_i)} = \alpha_{bj} + \beta_{bj} X_i, \quad i = 1, \ldots, n; \quad b = 1, \ldots, B; j = 1, \ldots, k_b.$$

To test the null hypothesis of $\beta_{bj} = 0$, we can use the standard likelihood ratio test, the Wald test, or the score test. All three test statistics have the same asymptotic distribution. Additionally, we can show that, under the null hypothesis, the likelihood ratio test statistic can be represented in the form $U_{bj}^2/V_{bj} + o_p(1)$, where $U_{bj} = \sum_{i=1}^{n} U_{ibj}$, $V_{bj} = \sum_{i=1}^{n} U_{ibj}^2$, $U_{ibj} = (Y_{ibj} - \pi_{bj})(X_i - \bar{X})$, and $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Here π_{bj} is the AE rate under the null hypothesis and can be estimated by $\widetilde{\pi}_{bj} = \sum_{i=1}^{n} Y_{ibj}/n$. Write $W_{bj} = U_{bj}/\sqrt{V_{bj}}$. For fixed b and j, W_{bj} converges to a standard normal distribution under the null hypothesis. Furthermore, the correlation between W_{bj} and $W_{b'j'}$ can be consistently estimated by $\sum_{i=1}^{n} \widetilde{U}_{ibj}\widetilde{U}_{ib'j'}/\sqrt{\widetilde{V}_{bj}\widetilde{V}_{b'j'}}$, where $\widetilde{U}_{ibj} = (Y_{ibj} - \widetilde{\pi}_{ibj})(X_i - \bar{X})$ and $\widetilde{V}_{bj} = \sum_{i=1}^{n} \widetilde{U}_{ibj}^2$. It is worth to note that the above formulation allows us to account for correlations among AEs not only from the same BS but also from different BSs.

When the number of AE terms is small, we can estimate the correlations of the test statistics and then draw samples from the joint null distribution of the test statistics. However, in most applications, M is large and possibly greater than sample size n. In this case, we can adopt the multiplier bootstrap procedure using the Rademacher sequence as described in Diao et al. (2014).

We now consider a sequence of i.i.d. random variables $\mathbf{G} \equiv (G_1, \ldots, G_n)$ independent of the data with $E(G_1) = 0$ and $\mathrm{Var}(G_1) = 1$. It can be shown that under the null hypothesis, given the observed data, $\widetilde{W}_{bj}(\mathbf{G}) \equiv \sum_{i=1}^{n} \widetilde{U}_{ibj} G_i/\sqrt{\widetilde{V}_{bj}} \to N(0, 1)$. Furthermore, conditional on the observed data, $\mathrm{Cov}\{\widetilde{W}_{bj}(\mathbf{G}), \widetilde{W}_{b'j'}(\mathbf{G})\} = \sum_{i=1}^{n} \widetilde{U}_{ibj}\widetilde{U}_{ib'j'}/\sqrt{\widetilde{V}_{bj}\widetilde{V}_{b'j'}}$. Therefore, given the observed data, the conditional joint distribution of $\{\widetilde{W}_{bj}(\mathbf{G}) : b = 1, \ldots, B; j = 1, \ldots, k_b\}$ can be used to approximate the joint null distribution of $\{\widetilde{W}_{bj} : b = 1, \ldots, B; j = 1, \ldots, k_b\}$, where $\widetilde{W}_{bj} = \sum_{i=1}^{n} \widetilde{U}_{ibj}/\sqrt{\widetilde{V}_{bj}}$.

While it is challenging to derive the analytic form of the joint distribution of the test statistics, based on the above results, one can generate a large number of realizations from the joint (conditional) distribution of $\{\widetilde{W}_{bj}(\mathbf{G}) : b = 1, \ldots, B; j = 1, \ldots, k_b\}$ by generating a large number of random sequences of $(G_1, \ldots, G_n)$. The subsequent inference is based on these realizations.

Several choices of the distribution of G_i are available; for example, Diao et al. (2004) and Lin (2005) considered the standard normal distribution. Empirical studies by Diao et al. (2014) suggested that the choice of Rademacher sequence, in which $P(G_i = 1) = P(G_i = -1) = 1/2$, performs well especially when $M \gg n$. Furthermore, for a multiplier process $n^{-1/2} \sum_{i=1}^{n} Z_i G_i$, where Z_i's are i.i.d. random variables and Z_i and G_i are independent, a Rademacher sequence preserves variance and other even moments and hence can lead to more accurate higher order approximations. More discussions on the properties of the Rademacher sequence can be found in Koltchinskii (2006) and Tao (2009).

We describe the multiplier bootstrap procedure using the Rademacher sequence to control for the FWER as follows:

1. For $l = 1, \ldots, L$, generate i.i.d. Rademacher random variables $\{G_1^{(l)}, \ldots, G_n^{(l)}\}$, which are independent of the data.
2. Calculate $\widetilde{U}_{bj}^{(l)} = \sum_{i=1}^{n} \widetilde{U}_{ibj} G_i^{(l)}$ and $\widetilde{W}_{bj}^{(l)} = \widetilde{U}_{bj}^{(l)}/\sqrt{\widetilde{V}_{bj}}$. Define $\widetilde{\mathbf{W}}^{(l)} = (\widetilde{W}_{11}^{(l)}, \ldots, \widetilde{W}_{Bk_B}^{(l)})$.
3. For a given FWER α, compute the $1-\alpha$ sample quantile of $\{||\widetilde{\mathbf{W}}^{(l)}||_\infty, l = 1, \ldots, L\}$, where $||\widetilde{\mathbf{W}}^{(l)}||_\infty = \sup\{\widetilde{W}_{bj}^{(l)} : b = 1, \ldots, B; j = 1, \ldots, k_b\}$. Reject the null hypothesis if $||\widetilde{\mathbf{W}}||_\infty$ exceeds this threshold, where $\widetilde{\mathbf{W}} = \{\widetilde{W}_{bj} : b = 1, \ldots, B; j = 1, \ldots, k_b\}$.

In step 3, we can also calculate (an estimate of) the multiple testing adjusted p-value given as the proportion of $\{||\widetilde{\mathbf{W}}^{(l)}||_\infty : l = 1, \ldots, L\}$ that are greater than or equal to $||\widetilde{\mathbf{W}}||_\infty$. The standard error of the p-value estimate is bounded

from above by $\sqrt{p(1-p)/L} \leq 1/(2\sqrt{L})$, where p is the true (adjusted) p-value. Therefore, we can choose the value of L to control the standard error of the p-value estimate at a specified level.

Remark 1. The proposed multiplier bootstrap procedure accounts for arbitrary correlations among the test statistics at AEs within the same BS or across different BSs.

Remark 2. Since the proposed multiplier bootstrap procedure is simulation-based, its performance does not rely on the large sample approximations which require relatively large sample sizes and event rates away from 0 and 1. Therefore this procedure is applicable in the presence of rare events as shown in the numerical studies.

3 | SIMULATION STUDIES

We conduct extensive simulation studies to examine the performance of the proposed procedure to control FWER. We consider two simulation settings mimicking the two real examples in Section 4: one from a vaccine trial (*Example One*) and the other from an anti-depression trial (*Example Two*). In the first simulation setting, there are eight BSs. The corresponding numbers of AEs are 5, 7, 1, 1, 3, 11, 9, and 3. In the second simulation setting, there are twenty-one BSs, with 1, 9, 1, 7, 1, 10, 43, 27, 2, 28, 13, 15, 4, 24, 47, 41, 10, 18, 19, 25, and 6 AE terms in BSs 1-21, respectively. We use the observed proportion of event data for each AE in the combined data to generate data for the control group. In *Example One*, the observed proportion of event ranges from 0.008 to 0.490 with a median of 0.019 and a third quartile of 0.087. In *Example Two*, the observed proportion of event ranges from 0.002 to 0.214 with a median of 0.002 and a third quartile of 0.006. In particular, for 206 AE terms, the event rate is less than or equal to 0.002.

To generate correlated binary data, we first generate a multivariate normal vector $\mathbf{Z} \equiv (Z_{11}, \ldots, Z_{1k_1}, \ldots, Z_{B1}, \ldots, Z_{Bk_B}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a variance-covariance matrix. All the diagonal elements of $\boldsymbol{\Sigma}$ are equal to one. We then determine $\mathbf{Y}$ by

$$Y_{bj} = \begin{cases} 1, & \text{if } Z_{bj} \leq \Phi^{-1}(p_{bj}) \\ 0, & \text{otherwise} \end{cases},$$

where p_{bj} is the event rate for the jth AE in the bth BS, $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$ and $\Phi^{-1}(\cdot)$ is its inverse function.

We consider different values for the variance-covariance matrix $\boldsymbol{\Sigma}$. Specifically, we assume that $\boldsymbol{\Sigma}$ depends on two parameters: the correlation coefficient among AEs from the same BS and the correlation coefficient among AEs across different BSs. Specifically, we let

$$\mathrm{Cov}(Z_{bj}, Z_{b'k}) = \begin{cases} \rho, & b = b', j \neq k \\ r, & b \neq b' \end{cases}.$$

We consider seven different pairs of (ρ, r): $(0,0), (0.3,0), (0.6,0), (0.9,0), (0.3,0.15), (0.6,0.3)$, and $(0.9,0.45)$. For the first scenario, all AEs are mutually independent. For scenarios 2–4, there are within-BS correlations whereas AEs from different BSs are independent. For scenarios 5–7, all AEs are correlated; however, AEs from the same BS have higher correlations than AEs from different BSs. We consider sample sizes of 50 and 100 in each treatment group. All results are based on 1,000 replicates with 10,000 Rademacher sequences for each replicate.

We examine the performance of the proposed multiplier bootstrap method using Rademacher sequences for controlling the FWER. For comparison, we also consider the Bonferroni correction procedure and the permutation procedure using the Fisher's exact test and score-type test described in Section 2. For Bonferroni correction procedures, we first calculate the un-adjusted p-values and then compare them with α/M, where α is the nominal significance level and M is the previously-defined total number of AE terms across all BSs. The permutation tests are based on 10,000 permuted samples. All tests are one-sided testing whether the event rate is higher in the treatment group than in the control group. We reject the null hypothesis if at least one AE is detected.

Table 1 presents the type I error rates at the significance level of 0.05 for simulation settings based on *Example One* and *Example Two*. In all cases, both the proposed method and the permutation methods have accurate control of the FWER, whereas the Bonferroni correction yields conservative results, especially with the Fisher's exact test. As expected, the Bonferroni correction is more conservative when the number of AE terms is large. We also compare the computation complexity of the proposed method with the permutation methods. On average, the proposed test is about 5 times faster than the permutation method using the score-type test and 6 times faster than the permutation method using the Fisher's exact test. One would see even more computational advantage of the proposed method over the permutation methods when analyzing each permuted data set involves iterative procedures.

For the alternative hypothesis, we increase the event rate by 0.1 for the first and twenty first AEs in the treatment group. The event rates at these two AE terms for settings based on *Example One* are 0.402 and 0.087 in the control group. The corresponding relative risks at these two AE terms are 1.248 and 2.148. For settings based on *Example Two*, both event rates at these two AE terms are 0.002 in the control group leading to relative risks of 47.8. Table 2 presents the powers at the significance level of 0.05. Table 3 presents the corresponding

TABLE 1 Type I error rates for detecting at least one AE at the nominal significance level of 0.05 based on 1,000 replicates.

n	ρ	r	Bonf. (Fisher)	Bonf. (Score)	Perm. (Fisher)	Perm. (Score)	Proposed
Simulation settings mimicking *Example One*							
100	0	0	0.006	0.017	0.058	0.055	0.060
	0.3	0	0.002	0.017	0.048	0.047	0.050
	0.6	0	0.004	0.014	0.045	0.047	0.045
	0.9	0	0.004	0.012	0.042	0.039	0.043
	0.3	0.15	0.004	0.015	0.05	0.048	0.050
	0.6	0.3	< 0.001	0.008	0.044	0.048	0.046
	0.9	0.45	0.001	0.008	0.026	0.028	0.030
200	0	0	0.003	0.019	0.051	0.056	0.051
	0.3	0	0.007	0.019	0.040	0.047	0.042
	0.6	0	0.004	0.021	0.047	0.047	0.045
	0.9	0	0.006	0.013	0.037	0.037	0.036
	0.3	0.15	0.011	0.020	0.049	0.052	0.052
	0.6	0.3	0.009	0.021	0.053	0.056	0.054
	0.9	0.45	0.005	0.016	0.052	0.058	0.053
Simulation settings mimicking *Example Two*							
100	0	0	< 0.001	0.003	0.049	0.044	0.056
	0.3	0	< 0.001	0.001	0.042	0.042	0.052
	0.6	0	< 0.001	0.001	0.054	0.055	0.066
	0.9	0	< 0.001	0.001	0.049	0.049	0.049
	0.3	0.15	0.001	0.001	0.041	0.046	0.047
	0.6	0.3	< 0.001	< 0.001	0.042	0.048	0.056
	0.9	0.45	< 0.001	< 0.001	0.048	0.047	0.049
200	0	0	0.001	0.002	0.054	0.062	0.062
	0.3	0	0.001	0.003	0.044	0.052	0.046
	0.6	0	0.002	0.004	0.053	0.061	0.059
	0.9	0	< 0.001	0.002	0.044	0.050	0.044
	0.3	0.15	< 0.001	0.002	0.043	0.050	0.046
	0.6	0.3	< 0.001	0.002	0.045	0.055	0.054
	0.9	0.45	0.001	0.002	0.031	0.039	0.036

sensitivities for the tests. Here sensitivity is defined as the proportion of correct decisions for the two true signals. For the permutation methods and the proposed method, sensitivity is $\sum_{j=1}^{N}\{I(\widehat{p}_{j,1} < 0.05) + I(\widehat{p}_{j,21} < 0.05)\}/(2N)$, where $\widehat{p}_{j,k}$ is the estimated adjusted p-value for the kth AE from the jth replicate, respectively, and N is the total number of replicates. For the methods using the Bonferroni correction, sensitivity is $\sum_{j=1}^{N}\{I(\widetilde{p}_{j,1} < 0.05/M) + I(\widetilde{p}_{j,21} < 0.05/M)\}/(2N)$, where $\widetilde{p}_{j,k}$ is the un-adjusted p-value for the kth AE from the jth replicate. As expected, the Bonferroni correction methods perform very poorly with very low powers and sensitivities especially when the number of AEs is large. For settings mimicking *Example One*, the performance of the proposed test and the permutation tests are comparable; however, the proposed multiplier method is much more powerful than its permutation counterparts under the simulation setting mimicking *Example Two* with sample size $n = 100$, in which there are a large number of AEs with rare events. These results suggest that the permutation tests may lose power compared to the proposed method when there are rare events and sample size is small. The powers/sensitivities appear to be low under some simulation settings due to either small sample sizes or small effect sizes. As the sample size or effect size increases, we achieve better powers/sensitivities.

For every method under consideration, we also calculated the specificity, which is defined as the proportion of correct decisions for the $M - 2$ AEs such that the null hypothesis is true. For the permutation methods and the proposed method, specificity is defined as $\sum_{j=1}^{N}\sum_{1\le k\le M, k\neq 1, k\neq 21} I(\widehat{p}_{j,k} \ge 0.05)/\{(M-2)N)\}$; for the methods using the Bonferroni correction, specificity is defined as $\sum_{j=1}^{N}\sum_{1\le k\le M, k\neq 1, k\neq 21} I(\widetilde{p}_{j,k} \ge 0.05/M)/\{(M-2)N)\}$. In all simulation settings, the specificities for every method are greater than 0.999 suggesting low false positive rates.

We conduct additional simulation studies under different settings of the correlation structure of the AE terms. The results are presented in Web Appendix B. The conclusions from the new simulations remain the same.

4 | APPLICATION

We apply the proposed method to two real clinical trials. In both examples, we generated 100,000 Rademacher sequences

TABLE 2 Powers for detecting at least one AE at the nominal significance level of 0.05 based on 1,000 replicates.

n	ρ	r	Bonf. (Fisher)	Bonf. (Score)	Perm. (Fisher)	Perm. (Score)	Proposed
Simulation settings mimicking *Example One*							
100	0	0	0.034	0.091	0.200	0.201	0.211
	0.3	0	0.036	0.094	0.189	0.178	0.190
	0.6	0	0.032	0.072	0.183	0.190	0.190
	0.9	0	0.035	0.076	0.183	0.181	0.190
	0.3	0.15	0.029	0.077	0.186	0.188	0.193
	0.6	0.3	0.029	0.070	0.177	0.179	0.180
	0.9	0.45	0.028	0.072	0.161	0.147	0.152
200	0	0	0.161	0.229	0.384	0.378	0.379
	0.3	0	0.159	0.221	0.372	0.363	0.358
	0.6	0	0.140	0.217	0.372	0.359	0.361
	0.9	0	0.151	0.198	0.344	0.340	0.337
	0.3	0.15	0.147	0.219	0.374	0.366	0.364
	0.6	0.3	0.151	0.215	0.364	0.353	0.350
	0.9	0.45	0.159	0.221	0.360	0.354	0.356
Simulation settings mimicking *Example Two*							
100	0	0	0.009	0.055	0.450	0.465	0.577
	0.3	0	0.006	0.043	0.459	0.477	0.592
	0.6	0	0.007	0.047	0.458	0.480	0.592
	0.9	0	0.006	0.048	0.500	0.541	0.633
	0.3	0.15	0.006	0.054	0.457	0.484	0.589
	0.6	0.3	0.009	0.049	0.458	0.486	0.575
	0.9	0.45	0.007	0.041	0.503	0.561	0.614
200	0	0	0.327	0.631	0.947	0.981	0.973
	0.3	0	0.320	0.608	0.946	0.978	0.972
	0.6	0	0.322	0.613	0.958	0.978	0.972
	0.9	0	0.353	0.615	0.966	0.984	0.980
	0.3	0.15	0.330	0.597	0.948	0.975	0.967
	0.6	0.3	0.326	0.593	0.940	0.964	0.958
	0.9	0.45	0.310	0.576	0.953	0.970	0.970

for the proposed method and 100,000 permuted samples for the permutation methods.

4.1 | Example one: safety analysis for a vaccine trial

We first illustrate the proposed method using the safety data collected from a vaccine trial. The data were published by Mehrotra and Heyse (2004a). The trial randomized healthy toddlers aged 12–18 months into a quadrivalent vaccine containing measles, mumps, rubella, and varicella (MMRV) administrated on day 0, or a tri-valent vaccine containing measles, mumps and rubella (MMR) administrated on day 0 followed by varicella (V) on day 42. The primary safety comparison was between the combination vaccine MMRV and the separate administration of varicella (control group), i.e., adverse experiences observed in the MMRV group during days 0–42 and that observed following the varicella vaccination (days 42–84). A total of 280 subjects were included in the data set (148 in the MMRV group, and 132 in the control group). Overall, there were 40 different AE terms involving 8 BSs.

Figures 3(a) and 3(b) in Web Appendix C present the heat maps of Spearman's correlation matrices of the raw outcome data and the proposed test statistics across the 40 AEs, respectively. It appears that the proposed method can capture the correlations across all AEs well. By using the proposed method, the adjusted p-value is 0.834 at the AE with the smallest p-value. Similar results were obtained from the permutation methods. The adjusted p-values are 0.855 and 0.767, respectively, corresponding to the permutation methods with the Fisher's exact test and the score-type test. Figure 3(c) in Web Appendix C displays the adjusted p-values and no significant signals were detected using either of the three methods.

4.2 | Example two: safety analysis for an anti-depression trial

We now apply the proposed method to an anti-depression clinical trial for substance-p antagonist aprepitant (Keller et al., 2006). In this trial, 468 patients aged 18 years or older with a documented diagnosis of major depression disorder were randomized in approximately equal ratios to receive aprepitant

TABLE 3 Sensitivity based on 1,000 replicates.

n	ρ	r	Bonf. (Fisher)	Bonf. (Score)	Perm. (Fisher)	Perm. (Score)	Proposed
Simulation settings mimicking *Example One*							
100	0	0	0.014	0.038	0.080	0.080	0.084
	0.3	0	0.017	0.040	0.081	0.076	0.082
	0.6	0	0.014	0.031	0.079	0.082	0.082
	0.9	0	0.016	0.034	0.084	0.083	0.086
	0.3	0.15	0.012	0.033	0.080	0.080	0.083
	0.6	0.3	0.014	0.034	0.081	0.080	0.083
	0.9	0.45	0.014	0.036	0.084	0.077	0.080
200	0	0	0.082	0.114	0.194	0.188	0.192
	0.3	0	0.078	0.109	0.194	0.185	0.184
	0.6	0	0.069	0.103	0.188	0.181	0.182
	0.9	0	0.074	0.098	0.178	0.173	0.172
	0.3	0.15	0.073	0.110	0.198	0.190	0.190
	0.6	0.3	0.076	0.110	0.193	0.187	0.185
	0.9	0.45	0.080	0.113	0.198	0.194	0.194
Simulation settings mimicking *Example Two*							
100	0	0	0.004	0.026	0.242	0.256	0.330
	0.3	0	0.003	0.021	0.249	0.264	0.340
	0.6	0	0.004	0.023	0.248	0.264	0.351
	0.9	0	0.003	0.024	0.272	0.306	0.376
	0.3	0.15	0.002	0.026	0.254	0.270	0.347
	0.6	0.3	0.004	0.024	0.259	0.278	0.347
	0.9	0.45	0.004	0.020	0.294	0.337	0.384
200	0	0	0.179	0.390	0.776	0.844	0.827
	0.3	0	0.175	0.372	0.766	0.839	0.819
	0.6	0	0.178	0.374	0.774	0.829	0.814
	0.9	0	0.194	0.382	0.784	0.840	0.828
	0.3	0.15	0.182	0.370	0.775	0.839	0.826
	0.6	0.3	0.187	0.376	0.778	0.830	0.816
	0.9	0.45	0.178	0.372	0.807	0.844	0.838

160 mg, paroxetine HCl 20 mg, or placebo for a period of 8 weeks. Adverse experiences that occurred during this treatment period were included in the primary safety analysis for the comparisons. In the data set, 220 patients were in the test treatment group and 248 patients were in the control group. Of the 468 patients in the study, 364 patients reported at least one AEs. A total of 351 different AE teams were reported in 21 BSs. In this data set, for many AE terms, the number of observed events is very small. In the placebo arm, no patients experienced any event for 133 AE terms and only one event was observed for 146 AE terms; in the treatment arm, there were no observed events for 114 AE terms and there was only one observed event for 145 AE terms.

Figures 5(a) and 5(b) in Web Appendix C present the heat maps of Spearman's correlation matrices of the raw outcome data and the proposed test statistics across the 351 AEs, respectively. It appears that most AEs are independent. The proposed method was able to capture those AE pairs with noticeable correlations. All three methods detected the same signals at at AE #42 *Dry mouth* and AE #59 *Nausea*. The adjusted p-values were 0.017, 0.015, 0.013 for AE #42, and 0.031, 0.020, 0.020 for AE #59, corresponding to the proposed method, the permutation methods with the Fisher's exact test and the score-type test, respectively. Figure 5(c) in Web Appendix C displays the adjusted p-values. On the other hand, no tests using the Bonferroni correction detect any signals.

5 | DISCUSSION

This paper proposes an efficient method for drug safety signal detection from correlated AEs in clinical trials. The advantages of the proposed multiplier bootstrap method for controlling the FWER are twofold. First, the proposed method allows for arbitrary correlation structures of the test statistics within the same BS and across different BSs. Second, the proposed method is simulation-based and thus less sensitive to rare events than methods relying on large sample theories. Additionally, compared to the permutation method, the proposed method is computationally more efficient and appears to have superior power to detect the true signals with rare events and small samples based on empirical studies.

We have considered binary outcomes whether a subject experiences an AE from a certain BS. An underlying assumption is that all patients have the same exposure time.

Furthermore, patients with multiple occurrences of an AE are treated the same as patients who experience the AE only once. In many studies, patients may have different exposure times to the treatment and they may experience an AE multiple times. Failing to account for varying exposure times or utilize all the available information may result in loss of statistical power for detecting AE signals. As a future topic, we will consider several regression models including the Poisson model, the negative binomial model, and zero-inflated models and then apply the proposed Monte-Carlo method for multiple testing adjustment.

It is possible that the adverse reaction to a drug is impacted by some factors such as age, gender, ethnicity, etc. To account for the confounding effects of these factors, one may include them as covariates in a regression model. Following Diao et al. (2014), we can derive the efficient influence function for the parameter of interest, i.e., the treatment effect and then apply the proposed multiplier bootstrap method using Rademacher sequences. Since iterative procedures are needed to analyze each permuted data set, the permutation method can be computationally very intensive. On the other hand, the proposed multiplier bootstrap method only involves the analysis once for each AE and therefore is computationally much more efficient than the permutation method.

Recently, Segal et al. (2018) proposed an asymptotic approximation and a resampling algorithm for quickly estimating small permutation p-values in two-sample tests. The authors showed that their proposed algorithm is more computationally efficient than standard alternatives, particularly for extremely small p-values. In our application, due to the nature of rare events and small to moderate large sample sizes, in general the p-values are not small. It would be interesting to apply this method to our setting in a future research project.

The proposed method is developed for drug safety data from clinical trials in which the individual-level data are available and there are two treatment groups. In some clinical trials, there are more than two treatment groups or multiple dose levels. The proposed method can be extended to incorporate such scenarios. Suppose that there are $K+1$ treatment groups including the control group. The variable X_i takes value 0 for the control group and value k for the kth treatment or dose level, $k = 1, \ldots, K$. We consider the following logistic regression model for each AE with the control group as the reference level:

$$\log \frac{P(Y_{ibj}=1|X_i)}{P(Y_{ibj}=0|X_i)} = \alpha_{bj} + \sum_{k=1}^{K} \beta_{bjk} I(X_i = k), \quad i = 1, \ldots, n; b = 1, \ldots, B; j = 1, \ldots, k_b,$$

where $I(\cdot)$ is the indicator function. We are interested in testing the null hypotheses $\beta_{bj1} = \cdots = \beta_{bjK} = 0$ for all $b = 1, \ldots, B$ and $j = 1, \ldots, k_b$. The efficient score for β_{bjk} for the ith subject is then $U_{ibjk} = (Y_{ibj} - \pi_{bj})\{I(X_i = k) - n_k/n\}$, where π_{bj} is the AE rate under the null hypothesis, and $n_k = \sum_{i=1}^{n} I(X_i = k)$ is the number of patients in the kth treatment group. Write $\mathbf{U}_{ibj} = (U_{ibj1}, \ldots, U_{ibjK})^T$, $\mathbf{U}_{bj} = \sum_{i=1}^{n} \mathbf{U}_{ibj}$, $\mathbf{V}_{bj} = \sum_{i=1}^{n} \mathbf{U}_{ibj}\mathbf{U}_{ibj}^T$, and $W_{bj} = \mathbf{U}_{bj}^T \mathbf{V}_{bj}^{-1} \mathbf{U}_{bj}$. For fixed b and j, W_{bj} converges to a chi-square distribution with degrees of freedom K under the null hypothesis. Replacing the unknown parameter π_{bj} with its estimator under the null hypothesis $\widetilde{\pi}_{bj}$, we obtain $\widetilde{U}_{ibjk}, \widetilde{\mathbf{U}}_{ibj}, \widetilde{\mathbf{U}}_{bj}$, $\widetilde{\mathbf{V}}_{bj}$, and $\widetilde{W}_{bj}$. Following Section 2, under the null hypothesis, given the observed data, $\widehat{W}_{bj} \equiv \{\sum_{i=1}^{n} \widetilde{\mathbf{U}}_{ibj} G_i\}^T \widetilde{\mathbf{V}}_{bj}^{-1} \{\sum_{i=1}^{n} \widetilde{\mathbf{U}}_{ibj} G_i\}$ also converges to a chi-square distribution with degrees of freedom K for a sequences of Rademacher random variables $G_i, i = 1, \ldots, n$. Furthermore, it can be shown that $\{\widetilde{W}_{bj}, b = 1, \ldots, B; j = 1, \ldots, k_b\}$ and $\{\widehat{W}_{bj}, b = 1, \ldots, B; j = 1, \ldots, k_b\}$ (given the observed data) have the same limiting joint distribution. The same multiplier bootstrap procedure described in Section 2 then can be used to control for the FWER.

ACKNOWLEDGEMENTS

The authors thank the Editor, the Associate Editor, and two anonymous referees for their valuable comments that have improved the presentation of the paper.

ORCID

Guoqing Diao http://orcid.org/0000-0001-7304-9591
Donglin Zeng http://orcid.org/0000-0003-0843-9280

REFERENCES

Ahmed, I., Thiessard, F., Miremont-Salamé, G., Begaud, B., and Tubert-Bitter, P. (2010). Pharmacovigilance data mining with methods based on false discovery rates: A comparative simulation study. *Clinical Pharmacology & Therapeutics* 88, 492–498.

Ahmed, I., Thiessard, F., Miremont-Salame, G., Haramburu, F., Kreft-Jais, C., Be'gaud, B., and Tubert-Bitter, P. (2012). Early detection of pharmacovigilance signals with automated methods based on false discovery rates. *Drug Safety* 35, 495–506.

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

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

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

Chen, M., Zhu, L., Chiruvolu, P., and Jiang, Q. (2015). Evaluation of statistical methods for safety signal detection: A simulation study. *Pharmaceutical Statistics* 14, 11–19.

Diao, G., Hanlon, B., and Vidyashankar, A. N. (2014). Multiple testing for high dimensional data. *Perspectives on Big Data Analysis: Methodologies and Applications, Contemporary Mathematics, American Mathematical Society* 622, 95–108.

Diao, G., Lin, D., and Zou, F. (2004). Mapping quantitative trait loci with censored observations. *Genetics* 168, 1689–1698.

Diao, G. and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics* 194, 781–783.

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

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

DuMouchel, W. and Pregibon, D. (2001). Empirical bayes screening for multi-item associations. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 67–76. ACM.

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

Food and Drug Administration (2008). Guidance for industry: Diabetes mellitus-evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. [Online; accessed May 2018].

Hu, N., Huang, L., and Tiwari, R. C. (2015). Signal detection in FDA AERS database using dirichlet process. *Statistics in Medicine* 34, 2725–2742.

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

Huang, L., Zalkikar, J., and Tiwari, R. C. (2011). A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *Journal of the American Statistical Association* 106, 1230–1241.

Huang, L., Zalkikar, J., and 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.

Huang, L., Zheng, D., Zalkikar, J., and Tiwari, R. (2017). Zero-inflated poisson model based likelihood ratio test for drug safety signal detection. *Statistical Methods in Medical Research* 26, 471–488.

Jiang, Q. and Snapinn, S. (2016). Analysis of safety data. In *Cancer Clinical Trials*, pages 125–150. Chapman and Hall/CRC.

Keller, M., Montgomery, S., Ball, W., Morrison, M., Snavely, D., Liu, G., Hargreaves, R., Hietala, J., Lines, C., Beebe, K., et al. (2006). Lack of efficacy of the substance p (neurokinin1 receptor) antagonist aprepitant in the treatment of major depressive disorder. *Biological Psychiatry* 59, 216–223.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34, 2593–2656.

Lin, D. (2005). An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21, 781–787.

MedDRA (2018). Introductory Guide MedDRA Version 21.0. https://www.meddra.org/sites/default/files/guidance/file/intguide_21_0_english.pdf. Accessed: 2019-01-22.

Mehrotra, D. V. and 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. and Heyse, J. F. (2004a). Multiplicity considerations in clinical safety analyses. *Statistical Methods in Medical Research* 13, 227–238.

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

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* 4, 213–226.

Nam, K., Henderson, N. C., Rohan, P., Woo, E. J., and Russek-Cohen, E. (2017). Logistic regression likelihood ratio test analysis for detecting signals of adverse events in post-market safety surveillance. *Journal of Biopharmaceutical Statistics* 27, 990–1008.

Norén, G. N., Bate, A., Orre, R., and Edwards, I. R. (2006). Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine* 25, 3740–3757.

Odani, M., Fukimbara, S., and Sato, T. (2017). A Bayesian meta-analytic approach for safety signal detection in randomized clinical trials. *Clinical Trials* 14, 192–200.

Price, K. L., Amy Xia, H., Lakshminarayanan, M., Madigan, D., Manner, D., Scott, J., Stamey, J. D., and Thompson, L. (2014). Bayesian methods for design and analysis of safety trials. *Pharmaceutical Statistics* 13, 13–24.

Rothman, K. J., Lanes, S., and Sacks, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and Drug Safety* 13, 519–523.

Segal, B. D., Braun, T., Elliott, M. R., and Jiang, H. (2018). Fast approximation of small p-values in permutation tests by partitioning the permutations. *Biometrics* 74, 196–206.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 479–498.

Tao, T. (2009). Talagrand's concentration inequality. https://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/. Accessed: 2018-05-30.

Wang, W., Whalen, E., Munsaka, M., Li, J., Fries, M., Kracht, K., Sanchez-Kam, M., Singh, K., and Zhou, K. (2018). On quantitative methods for clinical safety monitoring in drug development. *Statistics in Biopharmaceutical Research* 10, 85–97.

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

Zhao, Y., Yi, M., and Tiwari, R. C. (2018). Extended likelihood ratio test-based methods for signal detection in a drug class with application to FDA's adverse event reporting system database. *Statistical Methods in Medical Research* 27, 876–890.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article, including Web Appendices, Tables, and Figures referenced in Sections 1, 3, and 4.

How to cite this article: Diao G, Liu GF, Zeng D, Wang W, Tan X, Heyse JF, Ibrahim, JG. Efficient methods for signal detection from correlated adverse events in clinical trials. *Biometrics.* 2019;75:1000–1008. https://doi.org/10.1111/biom.13031