Contemporary Clinical Trials 83 (2019) 81–87

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Bridging blinded and unblinded analysis for ongoing safety monitoring and evaluation

Li-An Lin[a,*], Yilei Zhan[b], Hal Li[c], Shuai Sammy Yuan[c], Greg Ball[a], William Wang[c]

[a] Clinical Safety Statistics, Merck & Co., Inc., Rahway, NJ, USA
[b] Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ, USA
[c] Clinical Safety Statistics, Merck & Co., Inc., North Wales, PA, USA

ARTICLE INFO

Keywords:
Safety monitoring and evaluation
Blinded analysis
Unblinded analysis
IND safety reporting
Multiplicity

ABSTRACT

In order to better characterize the safety profile of investigational new drugs (INDs) during clinical development, more interest and attention have been paid to ongoing safety monitoring and evaluation. The 2015 US FDA IND safety reporting draft guidance compels sponsors to periodically evaluate unblinded safety data. However, maintaining the trial blind is necessary to avoid jeopardizing the validity of study findings. In this article, we propose an innovative new approach which includes analyzing both blinded and unblinded data. The proposed two-stage framework incorporates periodic analyses of blinded safety data to detect and flag adverse events that may have potential risk elevation related to experimental treatment, as well as planned unblinded analyses to quantify associations between the drug and adverse events, and to determine thresholds for referring adverse events for medical review and safety reporting.

1. Introduction

Evaluating patient safety is an integral and essential part of the clinical trial process. Safety monitoring and evaluation from ongoing clinical trials has a direct impact on the safety and clinical care of patients enrolled during and after completion of these trials. Goals of ongoing safety evaluation from clinical trials include early detection of important safety signals, protecting patients from unnecessary risks, and developing the safety profile of the experimental treatment. The regulatory landscape for safety monitoring of health care products has changed considerably in recent years. The US FDA published a final rule amending the safety reporting requirements under 21 CFR part 312 and 21 CFR part 320 in September 2010. The final rule has placed strong emphasis on expediting reports of serious events with a reasonable possibility of being associated with the experimental treatment, so that safety evaluations are not confounded with excessive noise and product safety can be assessed more meaningfully [1]. In particular, the 2012 final guidance on IND safety reporting (Safety Reporting Requirements for Investigational New Drug (IND) and Bioavailability/Bioequivalence (BA/BE) Studies) recommends that sponsors conduct ongoing safety evaluations, including periodic review and analysis of the entire safety database, not only for IND safety reporting purposes, but also to update the investigator brochure (IB), protocol, and consent forms with new safety information [2].

Moreover, the 2015 draft guidance (Safety Assessment for IND Safety Reporting) provides recommendations on the composition and role of a safety assessment committee (SAC), how it differs from a data monitoring committee (DMC), and how and when to unblind safety data. Specifically, the US FDA recommends “unblinding to allow a comparison of event rates and detection of numerical imbalances across treatment groups to identify important safety information.” [3]. It is generally agreed that unblinding during ongoing clinical trials, by an independent expert committee (such as a DMC or SAC), is needed to identify and evaluate important imbalances in serious adverse events while studies are ongoing [3,4]. With randomized and blinded clinical trials, it is also necessary to maintain the blind of study personnel to treatment assignment in order to avoid jeopardizing the validity of study findings. Auspiciously, ongoing analysis of blinded safety data can be used to inform relationships between the drug and adverse events. With an expected range of event rates from internal and/or external data sources, we can assess whether there is a possible risk elevation due to experimental treatment. Furthermore, evidence from a blinded safety analysis could be used to evaluate the need for performing an unblinded safety analysis [5,6]. Therefore, it is feasible and potentially advantageous to combine the strength of both periodic blinded analyses and planned unblinded analyses for monitoring and evaluation of accumulating safety data.

In this article, we propose a two-stage framework for ongoing safety

* Corresponding author.
E-mail address: li.an.lin@merck.com (L.-A. Lin).

https://doi.org/10.1016/j.cct.2019.06.022
Received 19 December 2018; Received in revised form 26 April 2019; Accepted 28 June 2019
Available online 29 June 2019

monitoring and evaluation. In the first stage, we periodically analyze blinded safety data to detect and flag adverse events that may have potential risk elevation related to experimental treatment. In the second stage, we conduct planned unblinded analyses to quantify associations between the drug and adverse events, and to determine thresholds for referring adverse events for medical review and possible safety reporting. The remainder of this article is organized as follows: In Section 2, we describe the proposed two-stage framework for ongoing safety monitoring and evaluation. Extensive simulations are presented in Section 3 to evaluate the performance of the proposed approach, in comparison to a single-stage approach with only an unblinded analysis. Practical considerations and related issues are discussed in Section 4.

2. Methods

2.1. General framework

Many adverse events are anticipated to occur in patients during the course of a clinical trial. These events are recognized on the basis of prior experiences (both clinical and non-clinical studies) with the product under investigation and with related products. The safety monitoring and evaluation process needs to distinguish between events that are adverse drug reactions (i.e., events likely caused by the investigational product) from those that likely would have occurred in the absence of the product (i.e., events that commonly occur in the disease population) [7]. The 2012 US FDA final rule describes two situations that would require an aggregate assessment to judge a causal association between the drug and an adverse event, which could require the sponsor to notify the FDA and all participating investigators in an IND safety reporting process: a. whether there is an unexpected adverse reaction occurring more frequently in the experimental treatment group than in a concurrent or historical control group (21 CFR 312.32(c)(1)(i)(C)); b. whether there is an increased rate of occurrence of a previously recognized adverse reaction over that listed in the IB or protocol (21 CFR 312.32(c)(1) (iv)) [1]. The proposed two-stage framework in this article is designed to evaluate all the aggregated adverse events in the ongoing clinical trials, with both blinded and unblinded data. The primary objective of the first stage blinded analysis is to provide early alerts on whether the adverse event rates in the ongoing clinical trials are higher than the expected rates which are derived from historical information. In the second stage, the flagged potential adverse reactions (from the first stage) plus well-established adverse reactions (listed in the IB or protocol) will be further evaluated in the unblinded analysis. The primary objective of the second stage unblinded analysis is to develop a procedure for testing associations between the drug and adverse events and establishing a threshold for tackling multiplicity that can properly balance the false positive and false negative error rates. Fig. 1 displays a flow chart of the proposed two-stage framework for safety monitoring and evaluation.

In the following sections, we will illustrate the proposed two-stage framework in a single randomized clinical trial; however, the proposed framework can be generalized to aggregate analysis across multiple trials in a clinical program. Suppose we have $m^{(0)}$ independent adverse events to be monitored simultaneously in a randomized clinical trial.

Table 1
Blinded vs. unblinded data.

Observed data	Control arm	Treatment arm
Unblinded	$X_{i1} \sim \text{Binomial}(n_1, p_{i1})$	$X_{i2} \sim \text{Binomial}(n_2, p_{i2})$
Blinded	$Y_i = X_{i1} + X_{i2}$	

Without loss of generality, suppose there are only two arms: experimental treatment and control. The randomization ratio between experimental arm and control is r. Among all the $m^{(0)}$ events, the proportion of adverse events with and without elevated risk are π_1 and $\pi_0 = 1 - \pi_1$, respectively. Assume a group of adverse events have been recognized as adverse reactions listed in the IB or protocol, based on pre-clinical studies, completed trials, or other information. Denote the proportion of previously recognized adverse reactions among all adverse events as q. For the ith adverse event, we assume the incidence rate is p_{i1} for the control arm, and p_{i2} for the active treatment arm. Assuming the total number of subjects enrolled to the current study is n, we only observe $\{Y_i : i = 1, 2, \cdots, m^{(0)}\}$, the total number of incidences across trial arms among n subjects when the data are blinded. While for the unblinded analysis, we can observe full data $\{n_1, n_2, (X_{i1}, X_{i2}) : i = 1, 2, \cdots, m^{(0)}\}$. Table 1 illustrates the observed data when blinded and when unblinded.

2.2. First stage: blinded analysis

The blinded data,without treatment information, can be obtained by the sponsor while the study is ongoing. With background event rate information from data sources external to the current study, sponsors can conduct blinded comparisons. Analyses of pooled (blinded) data can provide early alerts and additional assurance to sponsors before the planned unblinded analyses are carried out [8]. Boundaries for the blinded analyses are set up as alerts, they are not intended for immediately stopping the trial. Events that cross a boundary are flagged as potential adverse reactions that would warrant further evaluation.

Statistical methods have been developed for ongoing blinded safety monitoring [5,6,9–14]. Continuous safety monitoring is a dynamic process and Bayesian methods provide an excellent framework for monitoring of blinded safety data. They allow sponsors to take advantage of information originating from multiple sources, both internal and external to the trial. For example, they provide flexibility for incorporating historical knowledge of the safety profile into the decision making process. The sequential probability ratio test (SPRT) [10] and other likelihood-based methods [11] have also been developed for blinded safety monitoring. We apply a Bayesian safety monitoring method for the first stage blinded analysis, originally proposed by Ball [9]. Other well developed blinded safety monitoring methods are also feasible for our two-stage framework, but will not be illustrated in this article.

Assume the pooled number of events for adverse event i across both trial arms, $\{Y_i : i = 1, 2, \cdots, m^{(0)}\}$, follows a Binomial$(n, \theta_i)$ distribution, where θ_i is the weighted rate of event i for both experimental and control arms combined under the randomization ratio r. A prior is

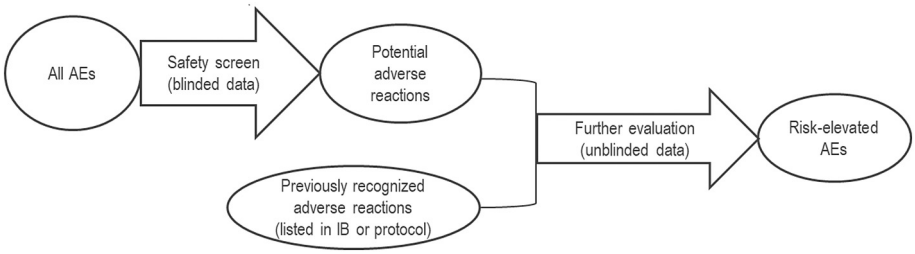

Fig. 1. The general framework of the proposed two-stage approach.

imposed on θ_i under the Bayesian framework, assuming the prior distribution for θ_i is Beta(α_{0i},β_{0i}). Note $\alpha_{0i}/(\alpha_{0i} + \beta_{0i})$ can be viewed as the mean rate of event i from the prior database and $\alpha_{0i} + \beta_{0i}$ can be viewed as the prior sample size. This is the classical Beta-Binomial model where the posterior distribution of θ_i is Beta($\alpha_{0i} + Y_i$,$\beta_{0i} + n - Y_i$). Suppose a critical pooled rate of event i is determined to be θ_{ci}. The posterior probability of the pooled rate being greater than the critical pooled rate is thus

$$P(\theta_i > \theta_{ci} \mid Y_i = y_i) = \int_{\theta_{ci}}^{1} \text{Beta}(\alpha_{0i} + Y_i, \beta_{0i} + n - Y_i)d\theta_i, \quad (1)$$

where Beta ($\alpha_{0i} + Y_i$,$\beta_{0i} + n - Y_i$) is the probability density function of Beta distribution.

A simple alerting rule is $P(\theta_i > \theta_{ci} | Y_i = y_i) > \theta_T$, where θ_T is the probability threshold boundary. θ_T is chosen to satisfy desired operating characteristics, such as false positive and false negative error rates. Moreover, the quantity of θ_T represents how certain we want the posterior probability of $P(\theta_i > \theta_{ci} | Y_i = y_i)$ to be before an alert is considered. Using simulations, we can evaluate operating characteristics with different values of θ_T. A set of θ_T can be selected, based on the purpose of the threshold boundaries. For example, we can choose one threshold boundary to flag potential adverse reactions and another one to refer events to the DMC (or SAC) for an unblinded review.

Proper specification of priors and critical rates is crucial for good performance of the blinded analysis [9]. The more certainty there is about prior knowledge for event rates, the more confidence there will be that the blinded analysis will make accurate inferences about the rate of adverse events in the trial population. The multidisciplinary team members should work together to discuss prior information for incorporation into statistical models. One source of prior knowledge is from internal studies, such as pre-clinical studies, early phase clinical trials, and reference safety information (RSI) for the investigational drug. Determining prior information from external sources requires careful examination of heterogeneity and potential bias; however, each source of data has its own strengths and limitations.

2.3. Second stage: unblinded analysis

Potential adverse reactions, flagged by the first stage blinded analysis, would compel further evaluation, via unblinded analysis, for quantifying the strength of evidence of any associations. For ongoing trials, the trial blind must be maintained, such that a DMC (or SAC) would be the only group conducting unblinded analyses. Therefore, a process for communication between the DMC and the study team must be established to avoid unintentional unblinding. At the end of a trial, the sponsor is able to access treatment information and conduct unblinded analyses. In this process, many adverse events are analyzed simultaneously, which makes multiplicity adjustment necessary. The ICH E9 guidance has recognized the investigation of safety and tolerability as a multidimensional problem [15]. One approach to address the safety and tolerability implications is by applying descriptive statistical methods to the data, supplemented by the calculation of confidence interval estimates of treatment differences. However, special care should be taken when interpreting putative statistically significant findings without appropriate multiplicity adjustment. Many articles have pointed out the potential of too many false safety signals if the multiplicity problem is not adequately addressed [16–18]. An increasingly common approach to address this multiplicity issue is to control the false discovery rate (FDR), which is loosely defined as the expected proportion of false positives among all significant hypotheses. The key objective in the context of evaluating potential adverse reactions is to find as many true signals as possible, while controlling the possibility of false positive findings to a certain level.

In the literature of multiplicity control, Benjamini and Hochberg (BH) developed the first FDR controlling method, and showed that it provides large gains in power over family wise error rate controlling

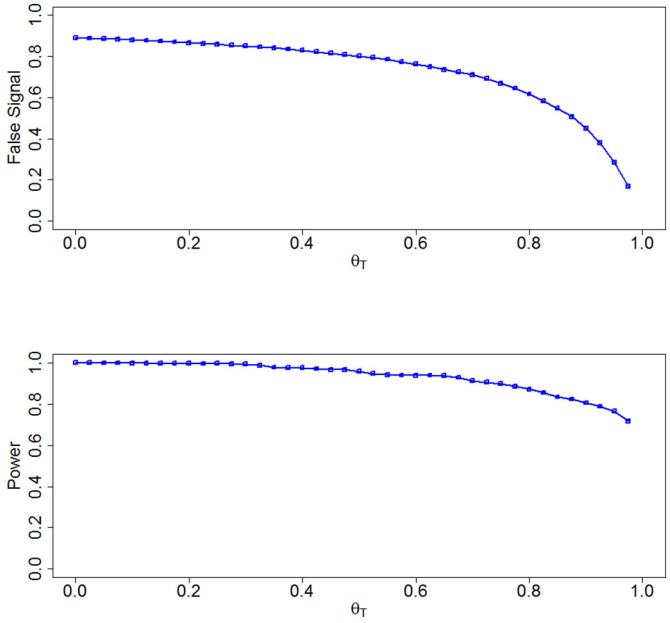

Fig. 2. Power and rate of false signals for first stage blinded analysis. Upper panel shows the mean rate of false signals, the mean proportion of false alerts among all the alerts. Bottom panel shows the corresponding power, the mean proportion of alerts among all risk-elevated events.

methods [19]. The BH p-value step-up procedure can provide strong but potentially conservative control of FDR. Storey has proposed a modified version of the FDR called the “positive false discovery rate” (pFDR) [20]. This procedure provides strong, but less conservative, control of FDR [21]. We will apply the pFDR method to control multiplicity for our second stage unblinded analysis.

Without loss of generality, suppose that $m^{(1)}$ adverse events were found to be potential adverse reactions. Together with the already established adverse reactions listed in the IB or protocol, the total number of events to be analyzed at second stage is $m = m^{(1)} + m^{(0)} * q$. Here we list the essential computational steps involved in the testing procedure. Theoretical proofs and technical details can be found in Storey and Tibshirani's paper [20].

(1) Compute marginal p-values p_1, p_2, ⋯, p_m.

To measure an event rate difference between two groups, commonly used risk metrics are the risk difference, risk ratio, and odds ratio. Odds ratio is mathematically equivalent to risk ratio when events are rare. Risk difference tends to show conservative confidence interval coverage and low statistical power for rare events [22]. Due to the low event rate of many adverse events (especially serious events), we perform the statistical tests on the risk ratio instead of the risk difference; however, the testing procedure described in this section is also applicable for other risk metrics. Marginal p-values are obtained by performing Miettinen- Nurminen (MN) test [23]. The MN test is an asymptotic method where the variance is estimated by maximizing the conditional likelihood, which is commonly used for reporting confidence intervals and p-values of between-treatment group comparisons, especially for late-stage randomized trials.

(2) Estimate π_0 by

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda: i = 1,\cdots,m\}}{m(1-\lambda)}, \quad (2)$$

where $\lambda \in [0,1]$. $\hat{\pi}_0(\lambda)$ is a conservative estimate of π_0 depending on the level of λ. Choices of λ can be made with a smoothing spline or the bootstrap method.

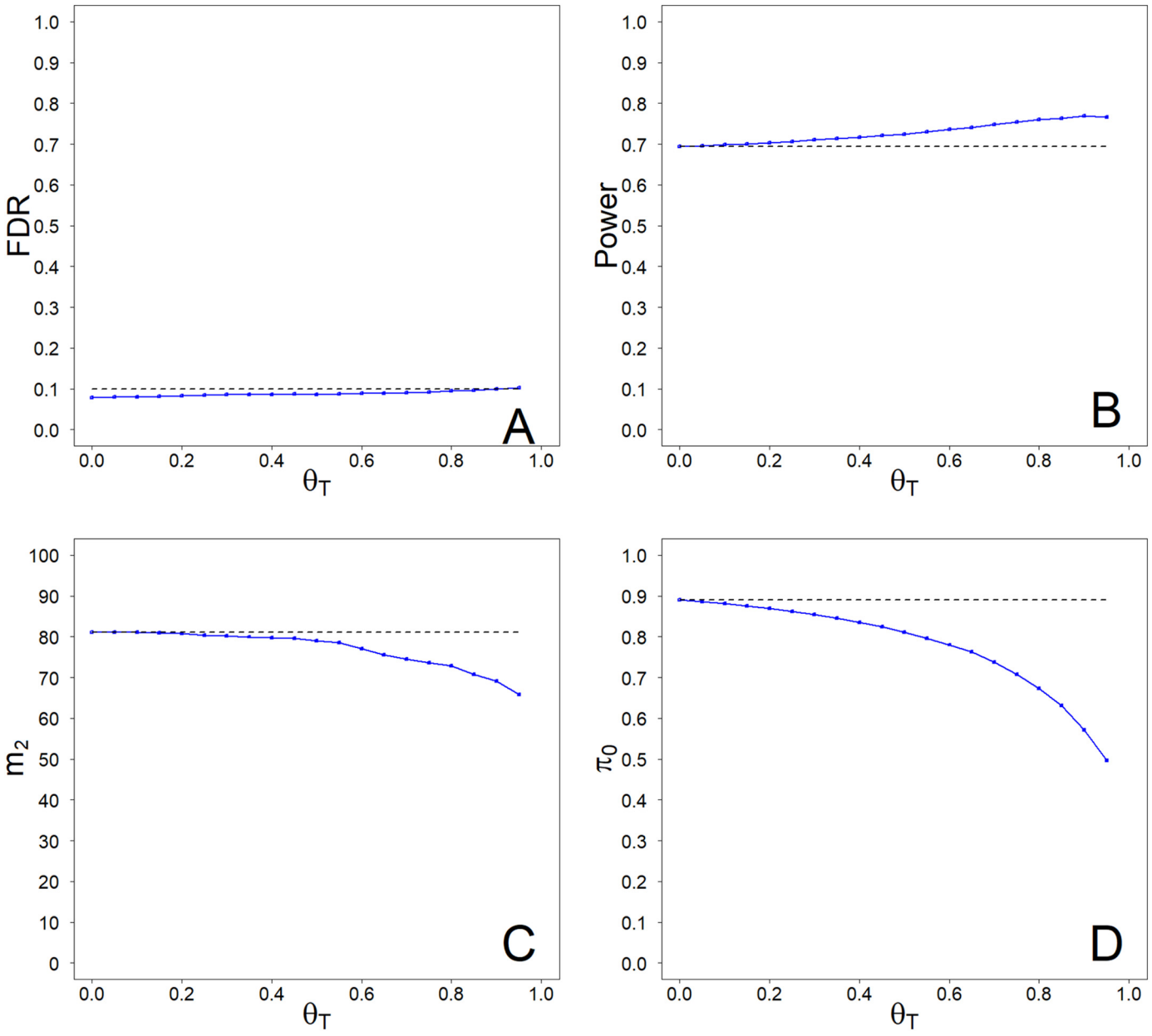

Fig. 3. Power and rate of false discovery for the whole procedure. Panel (A) shows the empirical FDR, the mean proportion of false discoveries among all the discoveries; Panel (B) the corresponding power, the mean proportion of true discoveries among all risk-elevated events. Panel (C) shows the number of identified true risk-elevated events m_2 for the second stage unblinded analysis, Panel (D) the corresponding proportion of non-risk events π_0. Two-stage approach, solid line; single-stage approach, dashed line.

(3) Estimate $FDR(t)$ by

$$\widehat{FDR}(t) = \frac{m\hat{\pi}_0 \cdot t}{S(t)}, \tag{3}$$

for some threshold $t \in (0,1]$, where $F(t) = \#\{p_i \le t : i \text{ is correspond true null}\}$, and $S(t) = \#\{p_i \le t : i = 1, \cdots, m\}$. Strictly speaking, the FDR is defined as $FDR(t) = \mathbb{E}\left[\frac{F(t)}{S(t)} \mid S(t) > 0\right]\mathbb{P}\{S(t) > 0\}$. However, since the number of tests m is large, $\mathbb{P}\{S(t) > 0\} \approx 1$. Therefore, alternatively we compute a similar error measure called positive false discovery rate (pFDR), which is defined as $pFDR(t) = \mathbb{E}\left[\frac{F(t)}{S(t)} \mid S(t) > 0\right]$.

(4) Estimate q-value by

$$\hat{q}_i = \hat{q}(p_i) = \min_{t \ge p_i} \widehat{FDR}(t). \tag{4}$$

For a given test, the q-value is defined as the minimum pFDR at which false positives occurred. It is a measure of evidence of the strength of an observed statistic with respect to the pFDR. The q-value gives each adverse event its own individual measure of evidence. In addition, the q-value can be interpreted as a Bayesian version of the p-value, which is the posterior probability of false positives among all adverse events as or more extreme than the one that was observed.

Following the above test procedure, a q-value threshold can be determined for possible association between drug and adverse events. If a possible association between the drug and an adverse event was to be identified, the clinical team would use medical judgment to assess the possible causal relationship and judge whether or not an aggregate safety report should be made.

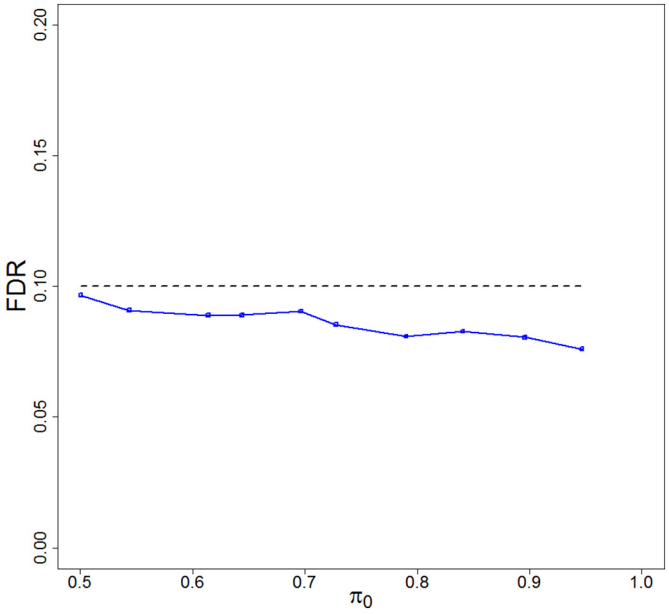

Fig. 4. π_0 vs. empirical FDR level. Empirical FDR level, solid line; Nominal FDR level, dashed line.

3. Numerical results

We illustrate our two-stage safety monitoring procedure with a hypothetical trial inspired by a phase III trial. This hypothetical trial randomized 4000 patients into treatment and control arms with a 1:1 randomization ratio. The priors and critical pooled rates were based on the RSI from previously completed trials. Out of 1000 adverse events (in preferred term) simulated, 50 events were previously recognized as adverse reactions.

3.1. Data generating

The true event rates were based on the phase III trial that inspired this research. We used Beta distributions to generate true event rates where $p_{i1} \sim \text{Beta}(0.4,5)/3$ for previously recognized adverse reactions, and $p_{i1} \sim \text{Beta}(0.2,5)/3$ for all other adverse events. The power of testing procedures to detect a safety signal varies by the event ratio between treatment groups. To assess the performance of the proposed procedure, we assume 100 adverse events are true safety signals with event ratio $p_{i2}/p_{i1} = 3$. However, not all events are observable in the simulated trials due to the low event rates in the simulation. We have excluded events with zero occurrence from our analysis since these events would not be reportable in real clinical trials. In a safety analysis, the threshold for FDR needs to properly balance the false discovery and false non-discovery error rates. In our illustration, we control the FDR level at $\alpha = 0.1$. All data generation and computation were implemented in *R*.

3.2. Simulation results

We conducted simulations to evaluate the performance, under various scenarios, of the proposed two-stage approach and to compare it with the traditional single-stage approach, which only has an unblinded analysis (using method described in Section 2.3) at the end of the trial. To simplify our illustration, only one blinded analysis and one unblinded analysis are conducted for the proposed two-stage approach. A total of 2000 independent simulations are performed for each scenario and method.

Ongoing blinded analyses can detect potential safety signals while the trial is still blinded. Fig. 2 displays the mean rate of false signals (the mean proportion of false alerts among all alerts) and the corresponding power (the mean proportion of alerts among all risk-elevated events) in the first stage blinded analysis, with different levels of probability threshold boundaries θ_T. The simulation results show that power decreases as θ_T increases. However, the corresponding rate of false signals also decreases, and the rate of false signals decreases faster than power decreases. Extensive simulations need to be conducted to evaluate the balance between the rate of false signal and power. Ideally, to minimize the noise, the rate of false signals should be as low as possible. At the same time, to better capture an early safety signal, the power should be as high as possible. By evaluating the operating characteristics, the multidisciplinary team can decide which values of θ_T would be most suitable. For example, if $\theta_T = 0.9$ is selected, the corresponding rate of false signals and power are 0.45 and 0.81.

After the first stage blinded analysis, the potential adverse reactions together with previously recognized adverse drug reactions (50 events) are entered into the second stage unblinded analysis. In this process, the control of the FDR is theoretically guaranteed when applying the specific multiple testing procedure illustrated in the previous section. Fig. 3(A) shows that the empirical FDR of the whole procedure is well controlled under different levels of the probability threshold boundary θ_T in the first stage blinded analysis. These results support the claim that the two-stage multiple testing procedure provides strong control of FDR.

As shown in Fig. 3(C) and (D), when θ_T increases, the number of identified true risk-elevated events m_2 would not change much, as long as θ_T is not extremely large. However, the proportion of true nulls π_0 drops dramatically. Therefore, we consider the possibility that the slightly increasing trend in FDR might be due to the decrease in π_0. As illustrated in Fig. 4, we run additional simulations to show that with a

Table 2
The power under various scenarios.

		$\pi_0^{(0)} = 0.9$		$\pi_0^{(0)} = 0.95$	
		$q = 0.05$	$q = 0.15$	$q = 0.05$	$q = 0.15$
$m^{(0)} = 1{,}000$	$\theta_T = 0.5$	74.0% (4.3%)	71.8% (4.2%)	56.6% (5.1%)	72.1% (1.8%)
	$\theta_T = 0.7$	76.1% (7.3%)	73.7% (7.1%)	58.8% (9.5%)	72.9% (2.9%)
	$\theta_T = 0.9$	77.9% (10.1%)	75.2% (9.6%)	62.5% (16.7%)	73.9% (4.4%)
$m^{(0)} = 3{,}000$	$\theta_T = 0.5$	68.8% (3.9%)	69.3% (3.7%)	67.0% (2.9%)	70.3% (2.3%)
	$\theta_T = 0.7$	70.9% (7.2%)	70.9% (5.9%)	68.6% (5.1%)	71.4% (3.9%)
	$\theta_T = 0.9$	73.3% (11.1%)	72.0% (7.9%)	71.0% (8.9%)	72.6% (5.6%)
$m^{(0)} = 5{,}000$	$\theta_T = 0.5$	74.6% (3.8%)	74.9% (3.4%)	71.4% (4.2%)	71.2% (3.2%)
	$\theta_T = 0.7$	76.6% (6.5%)	76.6% (5.4%)	73.5% (7.2%)	772.% (5.2%)
	$\theta_T = 0.9$	78.3% (9.1%)	77.4% (6.9%)	76.4% (11.4%)	74.2% (7.4%)

The numbers in this table are power (the mean proportion of alerts among all risk-elevated events) and percentage improvement in power compared to the single-stage approach(in parentheses). $\pi_0^{(0)}$, percentage of non-risk events; q, percentage of adverse reactions listed in the IB or protocol; $m^{(0)}$, number of simulated adverse events; θ_T, probability threshold boundary for the first stage blinded analysis.

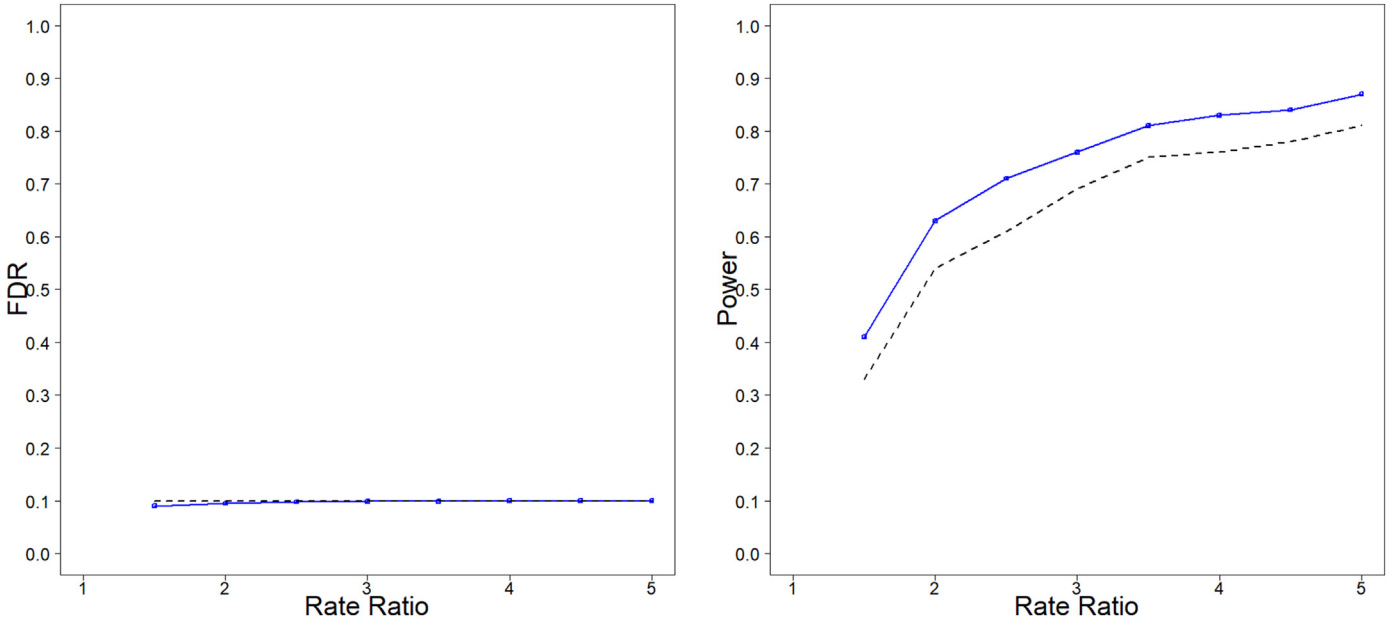

Fig. 5. Power and rate of false discovery under various scenarios. Left panel shows the empirical FDR, the mean proportion of false discoveries among all the discoveries. Right panel shows the corresponding power, the mean proportion of true discoveries among all risk-elevated events.Two-stage approach, solid line; single-stage approach, dashed line.

fixed number of true risk-elevated events, empirical FDR decreases as π_0 increases.

The proposed two-stage approach is compared with a single-stage approach which directly applies the multiple testing procedure on unblinded data of all observed adverse events. This approach corresponds to the particular two-stage approach with the probability threshold boundary $\theta_T = 0$. Fig. 3(B) illustrates the positive relationship between θ_T and empirical power. Simulation results show that the power increases as θ_T increases. When $\theta_T = 1$, this approach excludes all potential adverse reactions for unblinded analysis except the previously recognized adverse reactions. In this case, the power may be decreased for some scenarios. Both Fig. 2 and Fig. 3 could be useful for deciding the most appropriate choice of θ_T. Simulation results with more scenarios are summarized in Table 2 and Fig. 5.

4. Discussion

In the process of characterizing the product safety profile and identifying possible causal relationships between experimental treatment and adverse events while studies are ongoing and blinded, we propose a two-stage framework for safety monitoring and evaluation. The two-stage approach allows us to explore early safety signals and flag potential adverse reactions during the first stage blinded analysis. The second stage unblinded analysis enables us to validate and report adverse reactions, as appropriate. The proposed two-stage approach not only bridges the ongoing blinded and unblinded analyses, it also provides a systematic approach for updating safety information. The simulation studies show that the proposed framework provides strong control of FDR among multiple safety endpoints and can increase the power to detect safety signals, compared with a traditional single stage approach.

The proposed two-stage framework is a structured approach where potential adverse reactions are flagged from the complete set of adverse events at the first stage blinded analysis and then assessed for association at the second stage unblinded analysis. Many adverse events will be observed during the course of a clinical trial; most of them can be anticipated due to the disease being treated or population being studied. It has been argued that adjusting for multiplicity in safety evaluations can lead to a level of “cheating” by adding many safety parameters which are known to be unaffected by treatment but will increase the number of dimensions [18]. With our structured approach, the first stage blinded analysis is deployed to flag potential adverse reactions and reduce the dimension for multiplicity adjustment.

To infer associations between the drug and adverse events using blinded data, we need to ascertain the expected range of event rates from data sources that are external to the current study. It is challenging to use external data to determine suitable event rates in the trial. A multidisciplinary team is needed to ensure the validity of the blinded analyses. Regardless, treatment comparisons in blinded analyses can only be considered as exploratory. A blinded analysis alone cannot be used for IND safety reporting or updating safety information; an unblinded analysis would be needed.

At the second stage unblinded analysis, associations between the drug and adverse events are quantified and appropriate thresholds are determined for possible reporting of safety signals by taking advantage of information in the multiple blinded and unblinded analyses. Without multiplicity adjustment, there is potential for an excess of false positive findings which could complicate the safety profile of an experimental treatment [18]. However, the level of false discovery rate to be controlled cannot be predefined in safety analysis; it has to be considered under the specific trial setting to minimize the chance of missing rare but potentially important safety endpoints [24]. The balance between false discovery and false non-discovery error rates warrants further investigations.

To analyze and characterize adverse events, a three tiered approach has been recommended by the Safety Planning, Evaluation and Reporting Team (SPERT) [16]. In this system: Tier 1 events are those events with specific hypotheses being tested formally; Tier 2 events include common events without formal hypotheses; and Tier 3 events occur infrequently (must rely on medical judgment, statistical analysis is not informative). The framework we describe here applies most directly to Tier 2 events, which could benefit from statistical analysis and a systematic process. However, the principle of our proposed two-stage framework, bridging blinded and unblinded analyses is also applicable to Tier 1 and Tier 3 events.

Our proposed approach has some limitations that we view as future research opportunities. For example, we ignore the grouping of adverse events in this article. The grouping of adverse events may be defined by

body systems, standardized MedDRA queries (SMQs) (that represent a variety of safety topics of regulatory interest), or other characteristics when appropriately based on the underlying mechanism of action for the experimental treatment [18]. Adverse events are now routinely coded with Medical Dictionary for Regulatory Activities (MedDRA) terms in a hierarchical structure. Biological relationships, reflected in this intrinsic medical coding structure, indicate that adverse events in the same body systems are more likely to be similar [17]. How to incorporate the hierarchical structure of adverse events in the proposed two-stage framework is an interesting topic for future research.

The US FDA 2015 IND safety reporting draft guidance acknowledges that a two-step approach could be feasible for improving the overall quality of safety reporting and for complying with requirements for aggregate IND safety reports [3]. The proposed two-stage framework for ongoing safety monitoring and evaluations is designed to complement and interface with other existing processes for review of safety data in clinical development programs, such as DMCs. Potential safety concerns identified by the safety monitoring team could be referred to a DMC (or an SAC as proposed in the US FDA 2015 IND safety reporting draft guidance) for an unblinded assessment. Existing communication channels with the DMC could be used, in accordance with the DMC charter. The proposed two-stage approach provides a quantitative framework to partner with clinical safety judgment and to have a thoughtful process where there is a system in place to assess imbalances in safety events between treatment groups while protecting the trial integrity. Ultimately, it is the multidisciplinary team that must make informed judgments about the “reasonable possibility” of causal relationships.

References

[1] US Department of Health and Human Services, Food and Drug Administration, Investigational new drug safety reporting requirements for human drug and biological products and safety reporting requirements for bioavailability and bioequivalence studies in humans, Fed. Regist. 75 (188) (2010) 59935.

[2] US Department of Health and Human Services, Food and Drug Administration, Safety Reporting Requirements for INDs and BA/BE Studies, Guidance for Industry and Investigators, (2012).

[3] US Department of Health and Human Services, Food and Drug Administration, Safety Assessment for IND Safety Reporting, Draft Guidance for Industry and Investigators, (2015).

[4] P. Archdeacon, C. Grandinetti, J.M. Vega, D. Balderson, J.M. Kramer, Optimizing expedited safety reporting for drugs and biologics subject to an investigational new drug application, Ther. Innov. Regul. Sci. 48 (2) (2014) 200–207.

[5] P.M. Schnell, G. Ball, A bayesian exposure-time method for clinical trial safety monitoring with blinded data, Ther. Innov. Regul. Sci. 50 (6) (2016) 833–838.

[6] A.L. Gould, W.B. Wang, Monitoring potential adverse event rate differences using data from blinded trials: the canary in the coal mine, Stat. Med. 36 (1) (2017) 92–104.

[7] P.A. Lachenbruch, J. Wittes, Sentinel event methods for monitoring unanticipated adverse events, Adv. Stat. Methods Health Sci. (2007) 61–74.

[8] L. Zhu, B. Yao, H.A. Xia, Q. Jiang, Statistical monitoring of safety in clinical trials, Stat. Biopharm. Res. 8 (1) (2016) 88–105.

[9] G. Ball, Continuous safety monitoring for randomized controlled clinical trials with blinded treatment information: part 4: one method, Contemp. Clin. Trials 32 (2011) S11–S17.

[10] B. Yao, L. Zhu, Q. Jiang, H. Xia, Safety monitoring in clinical trials, Pharmaceutics 5 (1) (2013) 94–106.

[11] J. Herson, Safety Monitoring, Wiley-Blackwell, 2014, pp. 293–318 Ch. 11.

[12] G. Ball, P.M. Schnell, Blinded safety signal monitoring for the fda ind reporting final rule, Stat. Applic. Clin. Trials Personalized Med. Financ. Bus. Analytics (2016) 201.

[13] A.L. Gould, Control charts for monitoring accumulating adverse event count frequencies from single and multiple blinded trials, Stat. Med. 35 (30) (2016) 5561–5578.

[14] S. Mukhopadhyay, B. Waterhouse, A. Hartford, Bayesian detection of potential risk using inference on blinded safety data, Pharm. Stat. 17 (6) (2018) 823–834.

[15] ICH E9 Expert Working Group, ICH harmonised tripartite guideline: statistical principles for clinical trials, Stat. Med. 18 (1999) 1905–1942.

[16] B.J. Crowe, H.A. Xia, J.A. Berlin, D.J. Watson, H. Shi, S.L. Lin, J. Kuebler, R.C. Schriver, N.C. Santanello, G. Rochester, et al., Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team, Clin. Trials 6 (5) (2009) 430–440.

[17] H. Amy Xia, H. Ma, B.P. Carlin, Bayesian hierarchical modeling for detecting safety signals in clinical trials, J. Biopharm. Stat. 21 (5) (2011) 1006–1029.

[18] D.V. Mehrotra, J.F. Heyse, Use of the false discovery rate for evaluating clinical safety data, Stat. Methods Med. Res. 13 (3) (2004) 227–238.

[19] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. (1995) 289–300.

[20] J.D. Storey, et al., The positive false discovery rate: a bayesian interpretation and the q-value, Ann. Stat. 31 (6) (2003) 2013–2035.

[21] J.D. Storey, J.E. Taylor, D. Siegmund, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, J. Roy. Stat. Soc. 66 (1) (2004) 187–205.

[22] M.J. Bradburn, J.J. Deeks, J.A. Berlin, A. Russell Localio, Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events, Stat. Med. 26 (1) (2007) 53–77.

[23] O. Miettinen, M. Nurminen, Comparative analysis of two rates, Stat. Med. 4 (2) (1985) 213–226.

[24] A.L. Gould, Unified screening for potential elevated adverse event risk and other associations, Stat. Med. 37 (18) (2018) 2667–2689.