Published in final edited form as:

Pharm Stat. 2012 ; 11(4): 310–317. doi:10.1002/pst.1520.

Impact of safety monitoring on error probabilities of binary efficacy outcome analyses in large phase III group sequential trials

Yanqiu Weng, DDS, MS, Wenle Zhao, PhD, and Yuko Palesch, PhD

Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, SC

Abstract

In phase III clinical trials, some adverse events may not be rare or unexpected and can be considered as a primary measure for safety, particularly in trials of life-threatening conditions, such as stroke or traumatic brain injury. In some clinical areas, efficacy endpoints may be highly correlated with safety endpoints, yet the interim efficacy analyses under group sequential designs usually do not consider safety measures formally in the analyses. Furthermore, safety is often statistically monitored more frequently than efficacy measures. Since early termination of a trial in this situation can be triggered by either efficacy or safety, the impact of safety monitoring on the error probabilities of efficacy analyses may be non-trivial if the original design does not take the multiplicity effect into account. We estimate the actual error probabilities for a bivariate binary efficacy-safety response in large confirmatory group sequential trials. The estimated probabilities are verified by Monte Carlo simulation. Our findings suggest that type I error for efficacy analyses decreases as efficacy-safety correlation or between-group difference in the safety event rate increases. In addition, while power for efficacy is robust to misspecification of the efficacy-safety correlation, it decreases dramatically as between-group difference in the safety event rate increases.

Keywords

bivariate binary response; efficacy; group sequential test; phase III clinical trial; type I error; type II error; safety

1. Introduction

In some large phase III trials, particularly for life-threatening conditions, safety and efficacy endpoints may be highly correlated. For example, in acute neurological trials (*e.g.*, stroke and traumatic brain injury), mortality is often included in the primary efficacy measure, such as the modified Rankin Scale (mRS) [1, 2] or the Glasgow Outcome Scale (GOS) [3]. Furthermore, serious adverse events may not be rare or unexpected in these studies (*e.g.*, 15% early death rate and 13% congestive heart failure rate in acute stroke patients receiving 25% human serum albumin [4]). In such trials, interim efficacy analyses under the group sequential (GS) design often are not considered frequent enough to sufficiently monitor safety as well. In the currently ongoing large (maximum sample size is 1,100) multi-center randomized controlled trial, the Albumin in Acute Stroke (ALIAS) Part II Trial [4], three

interim efficacy analyses are planned at equally-spaced information intervals. The primary efficacy outcome is the binary “good” or “bad” outcome using scores on the mRS and the NIH (National Institutes of Health) Stroke Scale [5] assessed at 3 months from randomization. These relatively infrequent efficacy analyses are insufficient for monitoring safety, particularly if a difference in mortality between two groups exists. Therefore, statistical monitoring for safety is generally proposed more frequently than for efficacy.

GS designs are currently the most popular statistical approach to monitor efficacy in phase III clinical trials [6–8]. For safety monitoring, descriptive statistics (*e.g.*, mean, proportion, risk ratio) with or without formal statistical guidelines are implemented in practice. Some statistical safety monitoring guidelines are often pre-specified for life-threatening conditions with an adverse event of specific interest [9]. In the ALIAS Part II Trial, in addition to the 3 interim efficacy analyses, 11 safety analyses using repeated 99% confidence intervals (CIs) for the risk ratio (RR) of early (within 30 days of randomization) death are incorporated. These safety assessments are performed after every 100 subjects, with an expectation of observing up to 15 deaths. However, since the efficacy and safety parameters are correlated but monitoring boundaries for efficacy and safety are separately constructed, it is unknown how much impact safety monitoring would have on the error probabilities of the efficacy analyses. Also unclear is whether error probability estimation for the efficacy analyses is robust to any misspecification of the safety profiles, such as the efficacy-safety correlation and the safety event rates.

For exploratory phase II trials, a variety of designs have been adopted to account for the multiplicity effect between efficacy and safety evaluations [10–12]. For large confirmatory phase III randomized concurrently-controlled GS trials involving multivariate responses, a majority of the proposed methods are based on a global statistic [13, 14] or based on controlling for a study-wise error rate [15, 16]. One problem with the global method is that it cannot provide the exact marginal error probability for each outcome, making the study results sometimes difficult to interpret clinically, especially when the primary interest is to separately ascertain the efficacy and safety profiles of the test treatment. To address this problem, Cook and Farewell developed a method to calculate the marginal and joint error probabilities for a bivariate continuous efficacy-toxicity response for GS designs [17, 18], but comparable research for a bivariate binary response is sparse. In this paper, we develop a method to calculate the marginal and joint error probabilities for a bivariate binary efficacy-safety response in large GS trials based on multivariate normal approximation. The estimated probabilities are verified by Monte Carlo simulation.

In Section 2, we present a method to calculate the marginal and joint error probabilities for a bivariate binary efficacy-safety response in large phase III trials where the safety event rates are non-trivial and correlated to the efficacy measure. In Section 3, we demonstrate the computation procedure based on a hypothetical example developed from the ALIAS Part II Trial. In Section 4, Monte Carlo simulations are carried out to verify the proposed method. The relationships between the error probabilities of efficacy analyses and safety profiles are shown in Section 5. Limitations and extension of the proposed method are discussed in Section 6.

2. Multivariate normal approximation applied to the estimate of a bivariate binary efficacy-safety response at interim analysis in group sequential designs

2.1 Bivariate normal approximation to estimate a bivariate binary efficacy-safety response

Consider a GS trial comparing a new treatment (A) to a control (B). $K - 1$ interim analyses ($K > 1$) and one final analysis are planned to monitor a bivariate binary efficacy-safety response, where the efficacy outcome is success or failure and the safety outcome is occurrence of a specific safety event. Assume subjects are equally accrued at each interim stage. Let n_i denote the size of i th sequential group ($i = 1, 2, \dots, K$), namely the number of subjects enrolled in each arm between $(i-1)$ th and i th time point (time 0 represents the beginning of the study). Let p_{jE} and p_{jS} denote the true efficacy rate and safety rate in treatment j , $j = A, B$. In addition, n_i is selected so that the expected values of $n_i p_{jE}$ and $n_i p_{jS}$ are greater than 5.

Let Y_{ijE} and Y_{ijS} denote the number of successes and the number of safety events from n_i subjects in treatment j . By normal approximation to binomial distribution,

$$\begin{aligned} Y_{ijE} &\sim N\left(n_i p_{jE}, n_i p_{jE} (1 - p_{jE})\right) \\ Y_{ijS} &\sim N\left(n_i p_{jS}, n_i p_{jS} (1 - p_{jS})\right) \end{aligned}, (i=1, 2, \dots, K),$$

Denote Δ_{iE} and Δ_{iS} the differences in the numbers of efficacy and safety events between two treatments from the i th sequential group,

$$\begin{aligned} \Delta_{iE} &= Y_{iAE} - Y_{iBE} \\ \Delta_{iS} &= Y_{iAS} - Y_{iBS} \end{aligned}. \quad (1)$$

Δ_{iE} and Δ_{iS} can be consider as the estimates of treatment effect on efficacy and safety from n_i subjects. Due to independence of the two treatment groups, it follows that

$$\Delta_{iE} \sim N(n_i(p_{AE} - p_{BE}), n_i p_{AE} q_{AE} + n_i p_{BE} q_{BE}) \quad \Delta_{iS} \sim N(n_i(p_{AS} - p_{BS}), n_i p_{AS} q_{AS} + n_i p_{BS} q_{BS})$$

where $q_{jE} = 1 - p_{jE}$, $q_{jS} = 1 - p_{jS}$. The covariance between Δ_{iE} and Δ_{iS} is given by:

$$\begin{aligned} \text{cov}(\Delta_{iE}, \Delta_{iS}) &= \text{cov}(Y_{iAE} - Y_{iBE}, Y_{iAS} - Y_{iBS}) \\ &= \text{cov}(Y_{iAE}, Y_{iAS}) - \text{cov}(Y_{iAE}, Y_{iBS}) - \text{cov}(Y_{iBE}, Y_{iAS}) + \text{cov}(Y_{iBE}, Y_{iBS}) \\ &= \text{cov}(Y_{iAE}, Y_{iAS}) - \text{cov}(Y_{iAE}, Y_{iBS}) - \text{cov}(Y_{iBE}, Y_{iAS}) + \text{cov}(Y_{iBE}, Y_{iBS}) \quad (2) \\ &= \text{cov}(Y_{iAE}, Y_{iAS}) + \text{cov}(Y_{iBE}, Y_{iBS}) - \text{cov}(Y_{iAE}, Y_{iBS}) - \text{cov}(Y_{iBE}, Y_{iAS}) \end{aligned}$$

Since Y_{iAE} and Y_{iAS} (also, Y_{iBE} and Y_{iBS}) are correlated binomial random variables, let ρ_A and ρ_B be the phi correlation coefficients between efficacy and safety in treatments A and B, their covariance are:

$$\text{cov}(Y_{iAE}, Y_{iAS}) = \rho_A n_i \sqrt{p_{AE} q_{AE} p_{AS} q_{AS}} \quad \text{cov}(Y_{iBE}, Y_{iBS}) = \rho_B n_i \sqrt{p_{BE} q_{BE} p_{BS} q_{BS}} \quad (3)$$

From Equations 2 and 3, we get:

$$\text{cov}(\Delta_{iE}, \Delta_{iS}) = \rho_A n_i \sqrt{p_{AE} q_{AE} p_{AS} q_{AS}} + \rho_B n_i \sqrt{p_{BE} q_{BE} p_{BS} q_{BS}}$$

Therefore, the estimates of treatment effects on efficacy and safety from n_i subjects compose a random vector $(\Delta_{iE}, \Delta_{iS})'$, which is distributed as bivariate normal, $BVN(\boldsymbol{\mu}_i, \Sigma_i)$, with

$$\boldsymbol{\mu}_i = \begin{pmatrix} n_i(p_{AE} - p_{BE}) \\ n_i(p_{AS} - p_{BS}) \end{pmatrix}, \quad (4)$$

$$\Sigma_i = \begin{pmatrix} n_i(p_{AE} q_{AE} + p_{BE} q_{BE}) & \rho_A n_i \sqrt{p_{AE} q_{AE} p_{AS} q_{AS}} + \rho_B n_i \sqrt{p_{BE} q_{BE} p_{BS} q_{BS}} \\ \rho_A n_i \sqrt{p_{AE} q_{AE} p_{AS} q_{AS}} + \rho_B n_i \sqrt{p_{BE} q_{BE} p_{BS} q_{BS}} & n_i(p_{AS} q_{AS} + p_{BS} q_{BS}) \end{pmatrix}. \quad (5)$$

2.2 Joint distribution of efficacy and safety estimates from interim analyses

For a GS design with K analyses for both efficacy and safety, the estimates of treatment effect on efficacy and safety form a vector $\tilde{\Delta}$ with length of $2K$:

$$\tilde{\Delta} = \left(\Delta_{1E}, \Delta_{1S}, (\Delta_{1E} + \Delta_{2E}), (\Delta_{1S} + \Delta_{2S}), \dots, \sum_{i=1}^K \Delta_{iE}, \sum_{i=1}^K \Delta_{iS} \right), \quad (6)$$

Since $\Delta_{1E}, \Delta_{2E}, \dots, \Delta_{KE}$ are mutually independent, as are $\Delta_{1S}, \Delta_{2S}, \dots, \Delta_{KS}$, using matrix algebra, the vector in Equation 6 has a multivariate normal (MVN) distribution $\tilde{\Delta} \sim MVN(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ with

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 \\ \vdots \\ \sum_{i=1}^K \boldsymbol{\mu}_i \end{pmatrix}, \quad (7)$$

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_1 & \Sigma_1 & \cdots & \Sigma_1 \\ \Sigma_1 & \Sigma_1 + \Sigma_2 & \cdots & \Sigma_1 + \Sigma_2 \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_1 & \Sigma_1 + \Sigma_2 & \cdots & \sum_{i=1}^K \Sigma_i \end{pmatrix}, \quad (8)$$

where μ_i and Σ_i are defined in Equation 4–5. From $\tilde{\cdot}$, we can further develop a standardized vector $\mathbf{Z} = (Z_{1E}, Z_{1S}, Z_{2E}, Z_{2S}, \dots, Z_{KE}, Z_{KS})$, with Z_{kE} and Z_{kS} represent the standardized test statistics for efficacy and safety at time k , ($k=1, 2, \dots, K$).

2.3 Calculation of the stopping probabilities for a GS trial

Assume a trial using boundary $\mathbf{c}_E = (c_{1E}, c_{2E}, \dots, c_{KE})$ to monitor efficacy and boundary $\mathbf{c}_S = (c_{1S}, c_{2S}, \dots, c_{KS})$ to monitor safety. c_{kE} and c_{kS} are the critical values for the standardized test statistics Z_{kE} and Z_{kS} . Then, for a one-sided test

$$\begin{aligned} H_0: P_{AE} &\leq P_{BE} \\ H_a: P_{AE} &> P_{BE} \end{aligned},$$

the marginal stopping probabilities for efficacy at time k is

$$\Pr(\text{stop for efficacy at time } k) = \Pr\left(Z_{kE} \geq c_{kE} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right), \quad (9)$$

and the marginal stopping probabilities for safety at time k is

$$\Pr(\text{stop for safety at time } k) = \Pr\left(Z_{kS} \geq c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right). \quad (10)$$

Furthermore, the probability of stopping for both efficacy and safety at time k is:

$$\Pr(\text{stop for efficacy and safety at time } k) = \Pr\left(Z_{kE} \geq c_{kE} \cap Z_{kS} \geq c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right). \quad (11)$$

In above and below equations, we let $[Z_{0E} < c_{0E}] = [Z_{0S} < c_{0S}] = \Omega$. The overall marginal stopping probabilities for efficacy and safety are then given by:

$$\Pr(\text{stop for efficacy}) = \sum_{k=1}^K \Pr\left(Z_{kE} \geq c_{kE} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right), \quad (12)$$

$$\Pr(\text{stop for safety}) = \sum_{k=1}^K \Pr\left(Z_{kS} \geq c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right), \quad (13)$$

and the overall probability of stopping for both efficacy and safety reasons is:

$$\Pr(\text{stop for efficacy and safety}) = \sum_{k=1}^K \Pr\left(Z_{kE} \geq c_{kE} \cap Z_{kS} \geq c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}\right). \quad (14)$$

In many phase III trials of life-threatening conditions, the primary efficacy outcome may not be assessed for several months or even years after treatments are administered. Meanwhile, efficacy assessments can be suspended by the primary safety outcome – early death.

Therefore, the true type I error for efficacy in this circumstance can be expressed as a joint probability that a study stops for efficacy but not for safety concerns, which is:

$$\Pr(\text{true type I error}) = \sum_{k=1}^K \Pr \left(Z_{kE} \geq c_{kE} \cap Z_{kS} < c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}; H_0 \right). \quad (15)$$

Likewise, the true power is:

$$\Pr(\text{true power}) = \sum_{k=1}^K \Pr \left(Z_{kE} \geq c_{kE} \cap Z_{kS} < c_{kS} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell E} < c_{\ell E} \cap \bigcap_{\ell=0}^{k-1} Z_{\ell S} < c_{\ell S}; H_a \right). \quad (16)$$

Furthermore, we can use Equations 12 and 14 to derive the joint probabilities in Equations 15 and 16 with the relationships that:

$$\text{True type I error} = \Pr(\text{stop for efficacy}; H_0) - \Pr(\text{stop for efficacy and safety}; H_0),$$

$$\text{True power} = \Pr(\text{stop for efficacy}; H_a) - \Pr(\text{stop for efficacy and safety}; H_a).$$

3. An illustrative example and computation

We demonstrate our method using an example derived from a currently ongoing phase III trial, the Albumin in Acute Stroke (ALIAS) Part II Trial [4]. For ease of exposition, the original study design has been modified to some extent. For instance, for the hypothetical ALIAS trial, the sample size has been change from 1,100 to 1,200, and the number of safety looks also has been changed from 11 to 12. The phase III trial aims to investigate a neuroprotective drug in acute ischemic stroke. Eligible patients are randomized to either the albumin treatment or saline treatment (control). The primary efficacy endpoint is binary, *i.e.*, treatment success or failure at 3 month from randomization. Assuming the success rate in the control group is 40%, a sample size of 1,200 is computed to adequately detect a 10% difference with 93% power in the primary efficacy analyses under a one-sided O'Brien-Fleming type GS boundary [19], with three equally spaced interim analyses (overall type I error of 0.025). Death within 30 days from randomization is considered the primary measure for safety. Although it does not govern the sample size, the safety outcome is sequentially monitored during the study. Suppose the Data and Safety Monitoring Board (DSMB) requests 12 safety analyses to examine the excess deaths after every 100 subjects are assessed using repeated one-sided tests with p-value of 0.01. The structure of the GS design is illustrated in Figure 1:

The critical values for the test statistics at k th analysis can be calculated by:

$$c_{kE} = \Phi^{-1}(1 - \alpha_E(k)) \quad c_{kS} = \Phi^{-1}(1 - \alpha_S(k))$$

where $\alpha_E(k)$ and $\alpha_S(k)$ represent the nominal type I error for efficacy and safety analyses at time k under the Lan-DeMets alpha spending guideline [20], and $\Phi^{-1}(X)$ is X percentile of the standard normal distribution. The nominal type I errors for this study at each stage are presented in Table 1.

Let δ_E denote the true differences in efficacy rates between two treatments. Let δ_S denote the true differences in mortality rates. Let ρ_j be their correlation coefficient in treatment j , $j = A, B$. In addition, assume the true efficacy and mortality rates in the control are 40% and 15%. Using the proposed method, we calculate the true type I errors (*i.e.*, when $\delta_E = 0$) and true power (*i.e.*, when $\delta_E = 0.10$) for efficacy analyses in the following scenarios:

Scenario 1: $\delta_S = 0$, $\rho_A = \rho_B = 0$,

Scenario 2: $\delta_S = 0$, $\rho_A = \rho_B = 0.3$,

Scenario 3: $\delta_S = 0.03$, $\rho_A = \rho_B = 0.3$.

The stopping probabilities under null and alternative hypotheses for these scenarios are presented in Tables 2 and 3.

As shown in Tables 2 and 3, the true type I errors for the efficacy analyses are all lower than what is expected ($\alpha = 0.025$), and the true power are also lower than the expected power ($1 - \beta = 0.93$).

Furthermore, to assess the relationship between the correlation coefficient ρ_j and marginal power for efficacy and safety, we relax the value ρ_j in Scenario 3, and scan the marginal stopping probabilities as ρ_j changes from -0.3 to 0.3 . The range of ρ_j , according to Prentice [21], is determined by the marginal probabilities of two responses. For this study,

$$\rho \in \left[\max \left\{ -(p_E p_S / q_E q_S)^{1/2}, -(q_E q_S / p_E p_S)^{1/2} \right\}, \min \left\{ (p_E q_S / q_E p_S)^{1/2}, (p_S q_E / q_S p_E)^{1/2} \right\} \right].$$

Thus, when $p_E = 0.40$, $p_S = 0.15$, $\rho \in [-0.343, 0.514]$. As shown in Table 4, as ρ_j increases, the marginal stopping probability for efficacy increases, while the marginal stopping probability for safety decreases. This result is different from the conclusion in Cook and Farewell's paper [17] that, the smallest plausible ρ will guarantee the minimum power requirement for both efficacy and safety analyses.

In addition, we explore the relationship between the sample size and marginal power. We let $\delta_E = 0.10$, $\delta_S = 0.03$, $\rho_A = \rho_B = 0$, and allow the total sample size vary from 1,200 to 2,400. As shown in Table 5, as sample size increases, the marginal stopping probability for efficacy first increases then decreases, while the marginal stopping probability for safety keeps increasing. This is also different from the recommendation from Cook and Farewell [17] that choosing a larger sample size will satisfy the minimum power requirement for both efficacy and safety analyses.

4. Simulation study

To assess the validity of the proposed method, we perform Monte Carlo simulations in 5 scenarios with various efficacy and safety parameters (*i.e.*, p_{jE} , p_{jS} and ρ_j), and apply 100,000 runs for each simulation. We use the same study design as those in Section 3. In the simulation, crossing of the boundary is judged according to the comparison between the p -value from the interim analysis and the nominal alpha in Table 1. The study will be stopped as soon as either the efficacy or safety boundary is crossed. The stopping probabilities from

the simulation are then compared with those probabilities resulting from the proposed method in Table 6.

As shown in Table 6, the approximation results consistently match the simulation results, indicating that the proposed method is valid for power calculation for bivariate binary responses in GS designs.

5. Relationship between error probabilities for efficacy analyses and safety profiles

For trials with an efficacy endpoint and a safety endpoint, it is always desired that the true error probabilities for the efficacy analyses be robust to modest misspecification of the safety profiles. Here, the safety profiles refer to the safety-efficacy correlation (ρ_j) and the between-treatment differences in safety rates (δ_S), both of which are unknown in advance. We evaluate the robustness of these relationships in this section.

5.1 Relationship between type I errors for efficacy, ρ_j and δ_S

We set the total sample size as 1,200, assume the efficacy and safety rates in control as 0.40 and 0.15, and use the similar statistical boundaries from in Table 1. To obtain the type I error for efficacy, we set δ_E as 0. By applying different ρ_j ($\rho_j \in [-0.3, 0.3]$) and δ_S ($\delta_S \in [0, 0.1]$), we obtain the true type I errors under different scenarios and plot them in Figure 2. As shown in Figure 2, for all $\rho_j \in [-0.3, 0.3]$, the true type I errors are bounded by the points from the situations when $\rho_A = \rho_B = -0.3$ and $\rho_A = \rho_B = 0.3$, which indicate that true type I error for efficacy analyses decreases as ρ_j increases. Meanwhile, the type I error is driven by δ_S – as δ_S increases, the type I error decreases rapidly. A 5% increase in δ_S results in a 50% decrease in the true type I error for the efficacy analysis.

5.2. Relationship between true power for efficacy, ρ_j and δ_S

Using the similar ranges of ρ_j and δ_S , we estimate the true power for efficacy under five effect sizes of efficacy, *i.e.* $\delta_E = \{0, 0.05, 0.10, 0.15, 0.20\}$. The results from our work suggest that true power for the efficacy analyses increases as δ_E increases, while it decreases as δ_S increases. Figure 3 presents 5 shaded bands representing the true power under 5 different sizes of δ_E . The width of each band represents the range of variations in power as ρ_j varies from -0.3 to 0.3 (the mid-line within each band represents the points when $\rho_A = \rho_B = 0$). As shown in Figure 3, most parts of these bands are narrow, especially when $\delta_E = 0.10$ and $\delta_S = 0.05$, indicating that power is robust to change in ρ_j if the true effect size does not have a dramatic deviation from the original assumptions (*i.e.*, $\delta_E = 0.10$, $\delta_S = 0.0$). However, Figure 3 also indicates that power for efficacy drops rapidly as δ_S increases. For instance, when δ_E is 0.20, a 5% increase in δ_S leads to more than 20% loss of power.

In addition to the stopping probability for efficacy, we also plot the stopping probability for safety in Figure 4. It shows that, no matter how optimal the treatment effect is for efficacy, the probability of stopping for safety always rises steeply as δ_S increases. These results can explain the findings in Figure 3 – due to the increase in stopping probability for safety, the trial is less likely to be stopped for efficacy. In other words, the safety endpoint and efficacy

endpoint behave like two competing risks for termination of the study during sequential monitoring. In addition, in Figure 5, we plot the conditional probability of stopping for efficacy given not stopping for safety. From the figure, we can see that the conditional probability always increases as δ_E increases. Moreover, the conditional probability is not so robust to variations of ρ_j when δ_E is small (0 or 0.05) or δ_S is greater (0.05~0.10).

6. Discussion

In this paper, we use a multivariate normal approximation approach to estimate the marginal and joint stopping probabilities for a bivariate binary efficacy-safety response in large confirmatory GS trials. Since normal approximation is valid when the product of the sample size and event rate is greater than 5, the proposed method requires the expected number of events for the first interim analyses to be greater than 5. This requirement is quite reasonable for large confirmatory trials where a specific safety outcome that is being monitored is not rare. When the sample size is small or the adverse event is rare, the proposed method might not be suitable. But in those situations, safety monitoring using descriptive statistics is generally preferred over formal statistical guidelines, and the impact of safety monitoring on the error probabilities of efficacy analyses might be less. Further exploration in the exact joint distribution of bivariate binary responses will provide a more complete answer for small trials with rare safety events.

For ease of exposition, we have presented a simple GS design with equal increments in information for each interim look and use one-sided tests for both efficacy and safety analyses. In practice, this method can be applied in those designs with arbitrary timing of interim analyses and a mixture of one-sided and two-sided tests for safety and efficacy endpoints. The true power for efficacy in this paper is defined as the probability of having a favorable efficacy outcome without a safety concern (Equation 16). We use this joint probability because, in many phase III trials with early death as the primary safety outcome, when a subject reaches the safety endpoint, efficacy observations have to be suspended even if the study drug is truly efficacious for that subject. We recognize this definition is not always applicable and could vary depending on study purposes and diseases of interest. But as shown in Equations 12 and 13, marginal power for efficacy and safety can be easily attained from the proposed method if they are of more interest.

Although the joint distribution of test statistics for bivariate normal responses in GS designs has been studied by Cook and Farewell [17], we believe it is still informative and imperative to make an exploration in the bivariate binary responses. Since the variance-covariance is arbitrarily assumed for the bivariate continuous responses, the method and relevant conclusions in Cook and Farewell's paper [17] might not be applicable for bivariate binary responses. For example, Cook and Farewell state that, when the correlation of efficacy and safety, ρ , is unknown, using the smallest acceptable ρ will guarantee the greater power for both efficacy and safety analyses. However, our study suggests that marginal power for a bivariate binary response can sometimes decrease as ρ increases, which has been shown in Table 4. In addition, Cook and Farewell [17] recommend that, "if there are power requirements for both the efficacy and toxicity analyses there will be two group sizes, g_1 and g_2 , determined. Taking $\max(g_1, g_2)$ will satisfy the more stringent requirement and provide a

more powerful analysis for the remaining outcome". From Table 5, we can see that the assumption also may not hold for binary data. These discrepancies might be explained by the fact that the covariance matrix of the test statistics for a bivariate binary response is constrained by its means (event rates), while the covariance matrix for a bivariate continuous response is often determined from the external information or is arbitrarily assumed [17, 18, 22].

Like a majority of the GS designs for superiority tests, the method introduced in this paper assumes that the DSMB complies with the statistical guidelines throughout the trial, which may not always be the case in practice. In fact, decision making by the DSMB, especially when it is considering multiple endpoints simultaneously, is based on many facets other than statistical evidence. An extended approach incorporating random effects in the decision making processes of DSMBs is therefore needed to make power estimations more practical and realistic for GS designs.

Finally, for a trial with a bivariate efficacy-safety response, it is always desirable that the estimated error probabilities for the efficacy analyses be robust to modest misspecification of the safety profiles because the exact safety profiles are usually unavailable in advance. However, our findings suggest that joint power as well as marginal power for efficacy analyses are very sensitive to variations in the safety profiles. For instance, in the example, when the effect size for efficacy is 10%, marginal power can decrease by 7% if the between-group difference in safety rates changes from 0 to 2%; when the effect size for efficacy is 20%, a 5% difference in safety event rates could lead to 20% loss of marginal power for the efficacy analyses. Such a decline in power has been shown to be related to the dramatic increase in the stopping probability for safety in Section 5.2. Consequently, if a new treatment has optimal efficacy while accompanied by a slightly increased adverse event rate, which can be common in life-threatening conditions, the use of two univariate GS boundaries for a bivariate response might be problematic. This is because applying two separate boundaries ignores the multiplicity effect between efficacy and safety, leading to an underpowered study. Because of these relationships, it is imperative that investigators use the bivariate GS design to plan their trials if both efficacy and safety endpoints are of interest, and consider the various potential scenarios for efficacy and safety parameters before the study.

Acknowledgements

We would like to thank Dr. Stacia DeSantis for helpful discussions. We also thank the Editor and two referees for their insightful comments, which greatly improved our manuscript. This work was supported by a National Institute of Neurological Disorders and Stroke (NINDS) grant, U01 NS054630.

Reference

1. Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scottish medical journal*. 1957; 2:200–215. [PubMed: 13432835]
2. Bonita R, Beaglehole R. Recovery of motor function after stroke. *Stroke; a journal of cerebral circulation*. 1988; 19:1497–1500.
3. Jennett B, Bond M. Assessment of outcome after severe brain damage. *Lancet*. 1975; 1:480–484. [PubMed: 46957]

4. Ginsberg MD, Palesch YY, Martin RH, Hill MD, Moy CS, Waldman BD, et al. The albumin in acute stroke (ALIAS) multicenter clinical trial: safety analysis of part 1 and rationale and design of part 2. *Stroke; a journal of cerebral circulation*. 2011; 42:119–127.
5. Brott T, Adams HP Jr, Olinger CP, Marler JR, Barsan WG, Biller J, et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke; a journal of cerebral circulation*. 1989; 20:864–870.
6. Food and Drug Administration. Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committee. 2006
7. Friedman, LM.; Furberg, C.; DeMets, DL. Fundamentals of clinical trials. 3rd ed.. New York: Springer; 1998.
8. Jennison, C.; Turnbull, BW. Group sequential methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC; 2000.
9. Whitehead J. On being the statistician on a Data and Safety Monitoring Board. *Statistics in medicine*. 1999; 18:3425–3434. [PubMed: 10611616]
10. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995; 51:1372–1383. [PubMed: 8589229]
11. Ivanova A, Qaqish BF, Schell MJ. Continuous toxicity monitoring in phase II trials in oncology. *Biometrics*. 2005; 61:540–545. [PubMed: 16011702]
12. Ray HE, Rai SN. An evaluation of a Simon 2-Stage phase II clinical trial design incorporating toxicity monitoring. *Contemporary clinical trials*. 2011; 32:428–436. [PubMed: 21256247]
13. Jennison C, Turnbull BW. Exact Calculations for Sequential t, chi-square and F tests. *Biometrika*. 1991; 78:133–141.
14. Tang DI, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*. 1993; 49:23–30. [PubMed: 8513104]
15. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics*. 1993; 49:741–752. [PubMed: 8241370]
16. Kosorok MR, Yuanjun S, DeMets DL. Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics*. 2004; 60:134–145. [PubMed: 15032783]
17. Cook RJ, Farewell VT. Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics*. 1994; 50:1146–1152. [PubMed: 7786995]
18. Cook RJ. Coupled error spending functions for parallel bivariate sequential tests. *Biometrics*. 1996; 52:442–450. [PubMed: 8672698]
19. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
20. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
21. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics*. 1988; 44:1033–1048. [PubMed: 3233244]
22. Todd S. An adaptive approach to implementing bivariate group sequential clinical trial designs. *Journal of biopharmaceutical statistics*. 2003; 13:605–619. [PubMed: 14584711]

Figure 1.
Interim analysis plan

Figure 2.
True Type I error, $\rho_j \in [-0.3, 0.3]$

Figure 3.
True power bands, $\rho_j \in [-0.3, 0.3]$

Probability of stopping for safety

Figure 4.
Probability of study stopping for safety, $\rho_j \in [-0.3, 0.3]$

Probability of stopping for efficacy given not stopping for safety

Figure 5.
Probability of study stopping for efficacy given not stopping for safety, $\rho_j \in [-0.3, 0.3]$

Table 1

Nominal type I error at each interim analysis

Time <i>k</i>	1	2	3	4	5	6	7	8	9	10	11	12	Overall type I error
$\alpha_E(k)^*$	0	0	0.00001	0	0	0.0015	0	0	0.0092	0	0	0.0220	0.025
$\alpha_S(k)^{\dagger}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.0475

* 3 interim analyses for efficacy using the one-sided O'Brien-Fleming group sequential guideline;

[†] 12 interim looks for safety using p-value < 0.01.

Table 2

Stopping probabilities under H_0 *

Scenario	Parameter (δ_S, ρ)	Pr(stop for efficacy)	Pr(stop for safety)	Pr(stop for safety and efficacy)	True type I error
1	0, 0	0.0239	0.0475	0.0001	0.0239
2	0, 0.3	0.0224	0.0473	0.0002	0.0221
3	0.03, 0.3	0.0137	0.2915	0.0012	0.0125

* $H_0 : p_{AE} = p_{BE}; \delta_E = 0$

Table 3

Stopping probabilities under H_a^*

Scenario	Parameter (δ_S, ρ)	Pr(stop for efficacy)	Pr(stop for safety)	Pr(stop for safety and efficacy)	True power
1	0, 0	0.8990	0.0408	0.0023	0.8966
2	0, 0.3	0.9006	0.0383	0.0033	0.8973
3	0.03, 0.3	0.7601	0.2106	0.0273	0.7329

* $H_a : pAE > pBE; \delta E = 0.1$

Table 4

Correlation and marginal stopping probability

ρ	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
Marginal stopping probability for efficacy	0.752	0.752	0.753	0.753	0.754	0.756	0.760
Marginal stopping probability for safety	0.231	0.228	0.224	0.221	0.218	0.214	0.211

$\rho_A = \rho_B = \rho$

Table 5

Sample size and marginal stopping probability

Sample size	1200	1440	1680	1920	2160	2400
Marginal stopping probability for efficacy	0.754	0.766	0.767	0.762	0.755	0.746
Marginal stopping probability for safety	0.221	0.238	0.253	0.268	0.281	0.294

$p_A = p_B = 0$

Table 6

Comparison of estimates from the MVN approximation method and the simulation study

Scenario	Group: (p_E, p_S, ρ)	Probability	MVN approximation	Simulation (100,000 times)
1	A: (0.50, 0.18, 0)	True power	0.73250	0.73440
		Stop for safety	0.22100	0.21872
		Stop for safety and efficacy	0.02151	0.02155
2	A: (0.50, 0.18, 0.3)	True power	0.73217	0.73129
		Stop for safety	0.21569	0.21660
		Stop for safety and efficacy	0.02449	0.02486
3	A: (0.48, 0.19, 0)	True power	0.51330	0.51348
		Stop for safety	0.37027	0.36981
		Stop for safety and efficacy	0.02737	0.02751
4	A: (0.60, 0.25, 0.3)	True power	0.22893	0.22709
		Stop for safety	0.77102	0.77291
		Stop for safety and efficacy	0.11499	0.11423
5	A: (0.70, 0.25, 0.3)	True power	0.76350	0.76104
		Stop for safety	0.23631	0.23896
		Stop for safety and efficacy	0.05753	0.05795