BIOMETRICS 55, 746–753
September 1999

Treatment Comparisons Based on Two-Dimensional Safety and Efficacy Alternatives in Oncology Trials

Peter F. Thall* **and Su-Chun Cheng†**

Department of Biostatistics, Box 237, M. D. Anderson Cancer Center,
1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
**email:* rex@odin.mdacc.tmc.edu
†*Current address:* Department of Statistics, Texas A&M University,
College Station, Texas 77843, U.S.A.

SUMMARY. In addition to their desired anticancer effects, most cancer treatments may also cause transient toxicity, permanent organ damage, or death. A critical question in comparing an experimental treatment to a standard is how much increase in an adverse event rate is an acceptable trade-off for achieving a targeted improvement in efficacy, or vice versa. We consider settings where one may characterize patient outcome as a bivariate (efficacy, safety) variable and quantify treatment effect as a corresponding two-dimensional parameter. A set of target parameters, each representing a clinically meaningful improvement over the standard, are elicited from the physician. Each target is a two-dimensional generalization of the usual one-dimensional shift parameter. We define the alternative hypothesis in the two-dimensional effect space as the convex hull of the sets of parameters that are at least as desirable as each target point. The rejection region is obtained by shifting the alternative toward (0,0) to achieve a given type I error, with sample size computed to achieve a given power at the targets. The method is illustrated by application to two cancer chemotherapy trials.

KEY WORDS: Bivariate data; Cancer chemotherapy; Clinical trials; Hypothesis testing; Safety monitoring.

1. Introduction

An inherent problem in cancer therapeutics is that, in addition to their desired anticancer effects, treatments may also cause transient toxicity, permanent organ damage, or death. In many oncology trials, an experimental treatment may have not only a greater efficacy than the standard therapy but also a higher adverse event rate. For example, a higher chemotherapy dose for treatment of soft tissue sarcoma increases the probabilities of both tumor shrinkage and life-threatening toxic effects on the kidneys and nervous system. In chemotherapy of acute leukemia, the first goal is to achieve complete remission (CR), as this is a necessary precursor to long-term survival. Unfortunately, experimental treatments for leukemia often increase the rates of both CR and myelosuppression. Allogeneic bone marrow or peripheral blood stem cell transplantation, in which cells from a matched donor are infused into the patient, carries the risk of graft-versus-host disease, which may be fatal. Because the brain is thought to be a site of lung cancer metastasis, prophylactic irradiation of the brain may be prescribed for certain lung cancer patients. In this case, loss of brain function may be the price paid for a decreased risk of brain cancer.

Each of these examples illustrates the antagonistic relationship between efficacy and safety in cancer therapeutics. If the primary goal of a clinical trial is to improve efficacy by some targeted amount, then a critical question is how much of an increase in the risk of a severe adverse event is an acceptable trade-off for achieving the targeted improvement in efficacy. The analogous question is how much drop in efficacy, if any, is acceptable to achieve a targeted improvement in safety.

In this paper, we address these considerations by formulating hypotheses in terms of a multidimensional parameter characterizing efficacy and safety outcomes together, rather than regarding one as the primary endpoint and the other as the secondary. We consider a class of testing problems motivated by comparative clinical trials where the patient outcome can be characterized by a bivariate (efficacy, safety) variable and specific clinical goals can be quantified in terms of a two-dimensional treatment effect parameter $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$, where Δ_1 accounts for efficacy and Δ_2 for safety. We propose a geometric method for constructing two-sample tests tailored to these clinical goals. This method first requires the specification of a set of target points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ in $\boldsymbol{\Delta}$-space that constitute clinically meaningful improvements over the null hypothesis $\mathbf{0} = (0,0)$. The target points are elicited from the physician. Each $\boldsymbol{\xi}$ is a two-dimensional generalization of the usual one-dimensional shift used to construct tests based on a single parameter. We define the alternative hypothesis in

this two-dimensional effect space to be the convex hull of the sets of $\boldsymbol{\Delta}$ that are at least as desirable as the target points. The rejection region of the test, defined in terms of a consistent estimator $\widehat{\boldsymbol{\Delta}}$ of $\boldsymbol{\Delta}$, is obtained by shifting the boundary of the alternative toward $\mathbf{0}$ to achieve a given type I error. Sample size is computed similarly to the usual method for achieving a given power, with the important difference being that here the power figures at all K-targeted alternatives are considered.

In recent years, there have been many proposals for constructing tests based on multiple outcomes in clinical trials (O'Brien, 1984; Wei and Lachin, 1984; Pocock, Geller, and Tsiatis, 1987; Tang, Gnecco, and Geller, 1989; Wei, Lin, and Weissfeld, 1989; Wei, Su, and Lachin, 1990; Su and Lachin, 1992). Tests based on the consideration of the two-dimensional structure of the space of parameters characterizing safety and efficacy have been proposed by Jennison and Turnbull (1993) for randomized trials and, for single-arm phase II trials, by Bryant and Day (1995) and Conaway and Petroni (1995, 1996). We are similarly motivated by the geometry of the two-dimensional parameter space. Cook and Farewell (1994) and Cook (1994, 1996) dealt with the sequential testing problem for safety and efficacy by defining bivariate error spending functions, thereby extending the fundamental idea of Lan and DeMets (1983). Williams (1996) addressed the sequential monitoring problem in the case of multiple time-to-event outcomes. A general approach to group-sequential trials accommodating multivariate outcomes and covariates has been given by Jennison and Turnbull (1997).

The remainder of the paper is organized as follows. In Section 2, we formally define the proposed method. Numerical computation is discussed in Section 3. Section 4 presents two illustrative applications. We close with a discussion in Section 5.

2. Constructing Two-Dimensional Tests

Let T and C index the experimental treatment and the standard control, respectively, and let Y_1 and Y_2 denote the efficacy and safety outcomes, respectively. Our method requires the specification of a two-dimensional treatment effect parameter $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$ such that Δ_1 and Δ_2 are real-valued T-versus-C efficacy and safety effects, respectively, with $\mathbf{0} = (0,0)$ the null point corresponding to no treatment difference. Thus, positive values of Δ_1 and Δ_2 correspond to superior efficacy and superior safety, respectively, with T compared with C, whereas negative values correspond to superiority of C over T. If the average behavior of $\mathbf{Y} = (Y_1, Y_2)$ under treatment i is characterized by the parameter $\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2})$ for $i = T, C$, then one may define the effects as $\Delta_j = g(\theta_{T,j}) - g(\theta_{C,j})$ for $j = 1, 2$, for an appropriate transformation g. For the bivariate binary $\mathbf{Y}$, where $\theta_{i,j} = \Pr(Y_{i,j} = 1)$, g may be the identity function, the variance stabilizing transformation $g(\theta) = \sin^{-1}(\theta)^{1/2}$, or a link function such as the logit, probit, or complementary log–log. Nonnegative-valued Y_j's may motivate the use of $\Delta_j = \log(\theta_{T,j}) - \log(\theta_{C,j})$, with $\theta_{i,j}$ the mean or median of Y_j under treatment i or the more general parameter $\Delta_j = \Pr(Y_{C,j} \leq Y_{T,j}) - \Pr(Y_{T,j} \leq Y_{C,j})$, which is defined without reference to any intermediate θ_j's.

Our aim is to construct one-sided tests of whether T is superior to C that quantify specific clinical goals in terms of both the efficacy effect Δ_1 and the safety effect Δ_2. Our construction of the alternative hypothesis $\Omega_a \subset R^2$ first requires the specification of one or more *target parameters* $\boldsymbol{\xi} = (\xi_1, \xi_2)$, with the requirement that at least one entry of each target $\boldsymbol{\xi}$ is positive. The relationship between each $\boldsymbol{\xi}$ and (0,0) here is a two-dimensional generalization of that in the one-parameter case between a targeted improvement $\xi > 0$, where the test's power is computed, and the null effect, 0. We will refer to a target parameter for which either $\xi_1 > 0 > \xi_2$ or $\xi_1 < 0 < \xi_2$ as a *trade-off target*. In the first case, the drop ξ_2 in safety is the trade-off for the improvement ξ_1 in efficacy. In the second case, the drop ξ_1 in efficacy is the trade-off for the improvement ξ_2 in safety. A more optimistic target specifies either an improvement in efficacy with no drop in safety, specifically $\xi_1 > 0 = \xi_2$, or an improvement in safety with no drop in efficacy, given by $\xi_1 = 0 < \xi_2$. The most optimistic type of target has both $\xi_1 > 0$ and $\xi_2 > 0$. In practice, the targets are elicited from the physician so that they quantify the clinical goals of the trial. We have found it most natural to elicit the targets in terms of parameters with which the physician is familiar, such as $\Pr(Y_j = 1)$ in the binary outcome case or median(Y_j) for time-to-event outcomes, and to then transform them to real-valued (ξ_1, ξ_2), as described earlier for mathematical convenience.

Given the targets $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$, we construct the alternative Ω_a as follows. For each $k = 1, \ldots, K$, define the set $A(\boldsymbol{\xi}_k) = \{\boldsymbol{\Delta} : \Delta_1 \geq \xi_{k,1} \text{ and } \Delta_2 \geq \xi_{k,2}\}$ of points that are at least as desirable as $\boldsymbol{\xi}_k$. The alternative Ω_a is then defined to be the convex hull of $\cup_{k=1}^{K} A(\boldsymbol{\xi}_k)$. That is, we require that if Δ_1 and Δ_2 are desirable alternatives, then so is $\lambda\Delta_1 + (1-\lambda)\Delta_2$ for each $0 < \lambda < 1$. If one or more of the $\boldsymbol{\xi}_k$'s is already inside the convex hull defined by the other target points, then any such $\boldsymbol{\xi}_k$'s are superfluous and may be dropped for simplicity without altering Ω_a. In most applications, $K = 1, 2$, or 3. The targets must be defined so that the null (0,0) is outside Ω_a; otherwise, the construction makes no sense. The boundary $\partial\Omega_a$ of Ω_a is a polygonal line with the $\boldsymbol{\xi}_k$'s at its vertices, illustrated by the solid line in Figure 1.

We construct a test of Ω_a versus (0,0) as follows. Assume that a consistent estimator $\widehat{\boldsymbol{\Delta}}$ of $\boldsymbol{\Delta}$ is available and that, for a trial with $2n$ patients randomized equally between T and

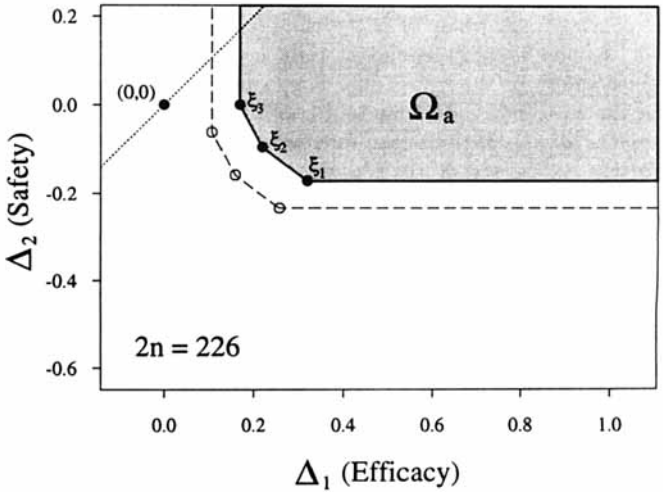

Figure 1. Alternative hypothesis and test boundary of Design 1 for the trial of ifosfamide in soft tissue sarcoma.

C, $(2n)^{1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$ is asymptotically bivariate normal with mean (0,0) and covariance matrix Σ, denoted $(2n)^{1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \sim AN(\mathbf{0}, \Sigma)$. Also assume that a consistent estimator $\widehat{\Sigma}$ of Σ is available. The test is defined in terms of $\widehat{\boldsymbol{\Delta}}$. The rejection region is obtained by shifting Ω_a the necessary distance toward (0,0) along the 45° line $\{\boldsymbol{\Delta} : \Delta_1 = \Delta_2\}$, given by the dotted line in Figure 1, to achieve a specified type I error probability α. Formally, the rejection region is $R(c_\alpha) = \Omega_a - (c_\alpha, c_\alpha)$, where c_α is the largest value such that $\Pr\{\widehat{\boldsymbol{\Delta}} \in R(c_\alpha) \mid \boldsymbol{\Delta} = \mathbf{0}\} \leq \alpha$. The test accepts Ω_a and concludes that T is superior to C if $\widehat{\boldsymbol{\Delta}} \in R(c_\alpha)$. Note that the null hypothesis may be defined as $\{\boldsymbol{\Delta} : \Delta_j \leq 0, j = 1,2\}$ without altering the test. The boundary of the rejection region is illustrated by the dashed line in Figure 1.

To verify that the power function $\phi(\boldsymbol{\Delta}) = \Pr\{\widehat{\boldsymbol{\Delta}} \in R(c_\alpha) \mid \boldsymbol{\Delta}\}$ of the test is nondecreasing in each of its two arguments, we proceed similarly to Jennison and Turnbull (1993) by first defining the standardized variables $Z_j = (\widehat{\Delta}_j - \Delta_j)/\widehat{\sigma}_j$, $j = 1,2$, where $\sigma_j^2 = \text{var}(\widehat{\Delta}_j)$. Thus, (Z_1, Z_2) is approximately bivariate normal with standard normal marginals. Denote the boundary of $R(c_\alpha)$ by $\partial R(c_\alpha)$. It follows from the definitions of Ω_a and $R(c_\alpha)$ that, for any correlation between Z_1 and Z_2, $\phi(\boldsymbol{\Delta}) = \Pr\{\exists\ (d_1, d_2) \in \partial R(c_\alpha)\ \ni\ \widehat{\Delta}_j \geq d_j,\ j = 1,2 \mid \boldsymbol{\Delta}\} = \Pr\{\exists\ (d_1, d_2) \in \partial R(c_\alpha)\ \ni\ Z_j \geq (d_j - \Delta_j)/\sigma_j,\ j = 1,2\}$ is nondecreasing in each Δ_j. Because, in general, $\phi(\boldsymbol{\xi}_1), \ldots, \phi(\boldsymbol{\xi}_K)$ may take on different values, in practice, the sample size $2n$ may be chosen either to achieve a desired power ϕ^* at a particular target $\boldsymbol{\xi}_k$ or to ensure that $\min_{1 \leq k \leq K} \phi(\boldsymbol{\xi}_k) \geq \phi^*$. We have found it useful to collaborate with the physician when examining $\phi(\boldsymbol{\xi}_1), \ldots, \phi(\boldsymbol{\xi}_K)$ over a range of n, so that $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ may be appropriately modified to obtain a realistic test and sample size.

Two common clinical trial settings where the method may be applied are those where both entries of $\mathbf{Y}$ are either binary or nonnegative valued. In the binary case, Y_1 is the indicator of the efficacy event, such as ≥50% tumor shrinkage, whereas Y_2 indicates that the adverse event, such as toxicity, did not occur. An important special case is when the occurrence of both the efficacy and adverse events is impossible, such as when the adverse event is death. Hence, there are three possible patient outcomes rather than four. The second setting typically arises when Y_1 is the time to relapse or death, subject to the usual independent right censoring, and Y_2 quantifies safety or quality of life. For example, if the treatment is known to cause damage to a specific organ, Y_2 might be a quantitative index of organ function. The distribution theory for these data structures is given in the Appendix.

3. Computation

For convenience and simplicity, we illustrate the method in the case of bivariate binary outcomes with the effects defined as $\Delta_j = \sin^{-1}(\theta_{T,j})^{1/2} - \sin^{-1}(\theta_{C,j})^{1/2}$ for $j = 1,2$, to stabilize the binomial variances. To facilitate presentation, we discuss parameter and trade-off values in the probability domain—as we do when communicating with physicians while developing a design—with the targets denoted by $\mathbf{p}_1, \ldots, \mathbf{p}_K$ to avoid confusing them with their corresponding transformed values $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$. As noted in the Appendix, in this case, $(2n)^{1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$ is asymptotically bivariate normal with mean $\mathbf{0}$, both variances 1, and correlation $\rho_\Delta = (\rho_C + \rho_T)/2$, where ρ_C and ρ_T are the respective correlations between Y_1 and Y_2 under treatments C and T. Given the null probability $\theta_{\mathbf{C}}$, ρ_C, the targets $\mathbf{p}_1, \ldots, \mathbf{p}_K$, and the sample size $2n$, the rejection region $R(c_\alpha)$ is uniquely determined by the type I error α, as $\Pr\{\widehat{\boldsymbol{\Delta}} \in R(c_\alpha) \mid \boldsymbol{\Delta} = \mathbf{0}\}$ is monotonically increasing in α. Thus, $R(c_\alpha)$ is easily obtained from a monotone search in c.

In general, the power function $\phi(\boldsymbol{\Delta})$ depends on $\boldsymbol{\Delta}$, n, and the correlation ρ_Δ. Thus, given $R(c_\alpha)$ and n, to compute $\phi(\boldsymbol{\xi}_k)$ for $k = 1, \ldots, K$, a value of ρ_Δ must be chosen at each $\boldsymbol{\xi}_k$. Although the intuitive choice $\rho_\Delta = \rho_C = \rho_T$ for all $\boldsymbol{\xi}_k$ might seem reasonable, in general, for given marginal probabilities $(\theta_{i,1}, \theta_{i,2})$ for $i = T, C$, not all values of ρ_Δ in $[-1, 1]$ are feasible. We thus chose to characterize association in terms of a common odds ratio $\psi = \psi_C = \psi_T$, as this may be carried out for all $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_T$ corresponding to a consistent probability distribution on the 2×2 table of outcomes. Therefore, ρ_Δ generally varies with $\boldsymbol{\Delta}$, with the provision that $\rho_\Delta = 0$ if $\psi = 1$.

All numerical calculations were carried out in S-plus on a Sun SPARC Station 20. We used the ACM Transactions on Mathematical Software Fortran algorithms 706 (Berntsen and Espelid, 1992) and 725 (Drezner, 1993) to compute approximations to bivariate normal probabilities, defined as integrals over convex sets obtained as unions of rectangles and triangles. Given α, the sample size $2n$, the null $\theta_{\mathbf{C}}$, the targets $\mathbf{p}_1, \ldots, \mathbf{p}_K$, and a common odds ratio ψ, the S-plus program computes the rejection region $R(c_\alpha)$, the correlations $\rho(\boldsymbol{\xi}_k)$, and the power figures $\phi(\boldsymbol{\xi}_k)$ for each $k = 1, \ldots, K$. The sample size is then sequentially modified until $\min_{1 \leq k \leq K} \phi(\boldsymbol{\xi}_k) \geq \phi^*$ for a desired power ϕ^*. Computations for the case of nonnegative-valued outcomes are carried out similarly, based on the parameterization and distribution theory discussed in the Appendix.

4. Illustrations

4.1 *Higher Ifosfamide Dose for Soft Tissue Sarcoma*

A standard chemotherapeutic regimen for untreated metastatic soft tissue sarcoma is 10 g/m^2 of ifosfamide. This dose not only achieves ≥50% tumor shrinkage, the efficacy outcome, in 20% of patients, but it also causes life-threatening (grade 3 or 4) nephrotoxicity or neurotoxicity in 5% of patients. Thus, by denoting the indicators of the efficacy outcome by Y_1 and the absence of both of these toxicities by Y_2, $E(Y_{C,1}, Y_{C,2}) = (\theta_{C,1}, \theta_{C,2}) = (.20,.95)$ under this standard treatment. The clinician indicated that these efficacy and toxicity events occur independently; hence, $\psi = 1$. It was hypothesized that by increasing the dose to 16 g/m^2, the experimental treatment might improve $\theta_{C,1}$ without causing too much of a decrease in $\theta_{C,2}$. Specifically, the three target points $\mathbf{p}_1 = (.50,.85)$, $\mathbf{p}_2 = (.40,.90)$, and $\mathbf{p}_3 = (.35,.95)$ were specified by the clinician. The first target, $\mathbf{p}_1$, allows a drop of $\delta_2 = -.10$ in safety as a trade-off for an increase of $\delta_1 = .30$ in efficacy, $\mathbf{p}_2$ targets the smaller increments $\delta_2 = -.05$ and $\delta_1 = .20$, and $\mathbf{p}_3$ targets the smallest clinically meaningful improvement $\delta_1 = .15$, for which no drop in safety is acceptable ($\delta_2 = 0$). This is summarized in Table 1 as Design 1, which requires a sample size of $2n = 226$ to achieve size .05 and power .80. The alternative determined by these three targets is illustrated in Figure 1.

Table 1
Three designs for the trial of ifosfamide in soft tissue sarcoma[a]

Design		$\boldsymbol{\delta} = (\delta_1, \delta_2)$	$\boldsymbol{\theta}_C + \boldsymbol{\delta}$	$\boldsymbol{\xi}_k$	$2n$
1	$\mathbf{p}_1$	(0.30, −0.10)	(0.50, 0.85)	(0.322, −0.172)	226
	$\mathbf{p}_2$	(0.20, −0.05)	(0.40, 0.90)	(0.221, −0.096)	
	$\mathbf{p}_3$	(0.15, 0)	(0.35, 0.95)	(0.169, 0)	
2	$\mathbf{p}_1$	(0.30, −0.15)	(0.50, 0.80)	(0.322, −0.238)	232
	$\mathbf{p}_2$	(0.20, −0.10)	(0.40, 0.85)	(0.221, −0.172)	
	$\mathbf{p}_3$	(0.15, −0.05)	(0.35, 0.90)	(0.169, −0.096)	
3	$\mathbf{p}_1$	(0.30, −0.10)	(0.50, 0.85)	(0.322, −0.172)	486
	$\mathbf{p}_2$	(0.20, −0.05)	(0.40, 0.90)	(0.221, −0.096)	
	$\mathbf{p}_3$	(0.10, 0)	(0.30, 0.95)	(0.116, 0)	

[a] $(\theta_{C,1}, \theta_{C,2}) = (.20, .95)$ and $\xi_{k,j} = \sin^{-1}(\theta_{C,j} + \delta_j)^{1/2} - \sin^{-1}(\theta_{C,j})^{1/2}$ for the kth target, for $j = 1, 2$.

Table 2
Four designs for the trial of gemcy versus ara-C in AML[a]

Design		$\boldsymbol{\delta} = (\delta_1, \delta_2)$	$\boldsymbol{\theta}_C + \boldsymbol{\delta}$	$\boldsymbol{\xi}_k$	$2n$
1	$\mathbf{p}_1$	(0.20, −0.05)	(0.90, 0.57)	(0.258, −0.051)	334
	$\mathbf{p}_2$	(0, 0.25)	(0.70, 0.87)	(0, 0.295)	
2	$\mathbf{p}_1$	(0.20, −0.05)	(0.90, 0.57)	(0.258, −0.051)	436
	$\mathbf{p}_2$	(0, 0.20)	(0.70, 0.82)	(0, 0.226)	
3	$\mathbf{p}_1$	(0.20, −0.05)	(0.90, 0.57)	(0.258, −0.051)	744
	$\mathbf{p}_2$	(0.10, 0)	(0.80, 0.62)	(0.116, 0)	
	$\mathbf{p}_3$	(0, 0.25)	(0.70, 0.87)	(0, 0.295)	
4	$\mathbf{p}_1$	(0.20, 0.05)	(0.90, 0.67)	(0.258, 0.052)	240
	$\mathbf{p}_2$	(0.05, 0.20)	(0.75, 0.82)	(0.056, 0.226)	

[a] $(\theta_{C,1}, \theta_{C,2}) = (.70, .62)$ and $\xi_{k,j} = \sin^{-1}(\theta_{C,j} + \delta_j)^{1/2} - \sin^{-1}(\theta_{C,j})^{1/2}$ for the kth target, for $j = 1, 2$.

A number of other possible designs were also considered, two of which are summarized in Table 1 as Designs 2 and 3. Under Design 2, each slippage δ_2 in safety is .05 larger than the corresponding value under Design 1. This difference, however, has a trivial effect on the sample size. In contrast, although the only difference between Designs 3 and 1 is that the efficacy component $p_{3,1}$ of $\mathbf{p}_3$ is .05 closer to 0, this more than doubles the sample size.

4.2 *Chemotherapy of Acute Leukemia*

Our second illustration arises from an experimental trial of the combination chemotherapy gemcitabine + cyclophosphamide (gemcy) versus the standard of cytosine arabinoside (ara-C)–based regimens for treatment of good-prognosis acute myelogenous leukemia (AML) or myelodysplastic syndrome patients. The efficacy outcome was the indicator Y_1 that the patient achieved CR, whereas "toxicity" was the event that the patient either died or suffered severe myelosuppression during the first 5 weeks. Thus, Y_2 was the indicator of no toxicity. The historical probabilities of these events with ara-C–based treatments were $\theta_{C,1} = .70$ and $\theta_{C,2} = .62$, with the odds ratio $\psi_C = 3.05$. The positive association between Y_1 and Y_2 may be expressed in more clinical terms by the fact that the CR rate was higher among patients not experiencing toxicity, specifically $\Pr(Y_{C,1} = 1 \mid Y_{C,2} = 1) = .790$, whereas $\Pr(Y_{C,1} = 1 \mid Y_{C,2} = 0) = .553$.

Four possible designs, summarized in Table 2 and illustrated in Figure 2, were considered for this trial. All tabulated sample sizes correspond to $\alpha = .05$ and $\phi^* = .80$. In Design 1, $\mathbf{p}_2$ targets a .25 increase in $\theta_{C,2}$ from .62 to .87, equivalent to a drop in toxicity probability from .38 to .13, with no change in efficacy. The other target, $\mathbf{p}_1$, allows a drop of .05 in safety as a trade-off for an increase of .20 in efficacy. Design 2 differs from Design 1 only in that $p_{2,2} = .20$ rather than .25; this increases the sample size from 334 to 436. Design 3 differs from Design 1 in that an additional third target is specified. This has the effect of greatly increasing the sample size, from 334 to 744, because the additional target is much closer to $\mathbf{0}$, as shown in Figure 2. Each of the Designs, 1, 2, and 3, specifies an alternative in which no drop in efficacy is desirable, regardless of how safe the experimental treatment might be. Similarly, each of these designs allows at most a small drop in safety, regardless of efficacy. In our experience constructing this type of design to compare an experimental cancer treatment with an established standard regimen, we have found that oncologists often place such absolute lower limits on efficacy, safety, or both. This attitude reflects the fact that either failing to achieve the efficacy outcome or experiencing the adverse outcome may have severe consequences for the patient. Design 4 is the most optimistic, in that each of its two targets specifies an improvement in both efficacy and safety. These four designs illustrate the general

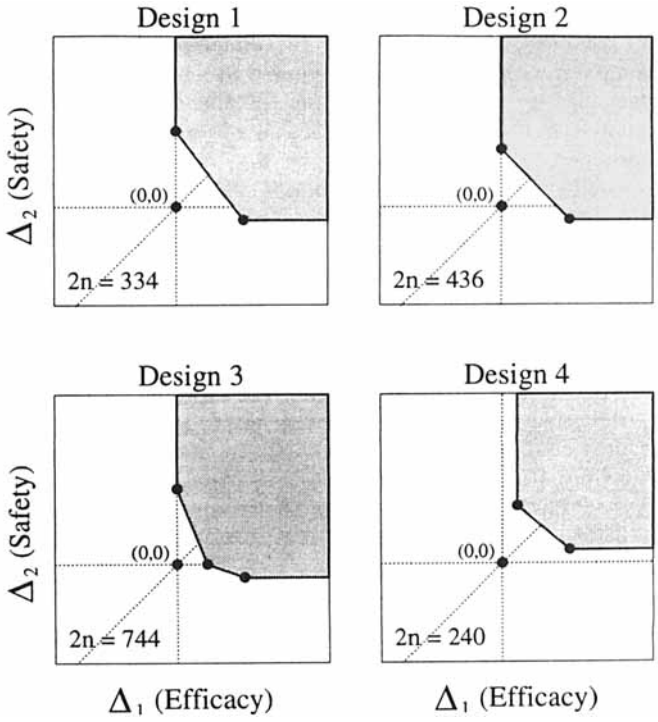

Figure 2. Alternative hypotheses of four designs for the gemcy versus ara-C trial in AML.

phenomenon that alternatives that are closer to (0,0) require larger sample sizes. Specifically, if two alternatives are nested such that $\Omega_{a1} \subset \Omega_{a2}$, then the test based on Ω_{a2} requires a larger sample size. Thus, the sample size for Design 3 is larger than that for Design 1 because $\Omega_{a1} \subset \Omega_{a3}$. Moreover, the increase is very large because Ω_{a3} is much closer to (0,0). This is because the power $\phi(\boldsymbol{\xi})$ at each target $\boldsymbol{\xi}$ is the volume of a bivariate normal distribution with mean $\boldsymbol{\xi}$ that is over the two-dimensional rejection region $\Omega_a - (c_\alpha, c_\alpha)$, and we determine the sample size to ensure $\phi(\boldsymbol{\xi}) \geq .80$ at all specified targets.

To examine the sensitivity of the method to association between Y_1 and Y_2, we computed the sample size for Design 1 for a range of hypothetical ψ_C values different from the actual null value $\psi_C = 3.05$. Given the marginals $\theta_{C,1}$ and $\theta_{C,2}$, fixing ψ_C is equivalent to fixing $\pi_{C,11} = \Pr(Y_{C,1} = 1 \text{ and } Y_{C,2} = 1)$. Values of $\pi_{C,11}$, the corresponding ψ_C, and the sample size are given in Table 3. We obtained the range $.32 \leq \pi_{C,11} \leq .62$ in Table 3 by first fixing the marginal probabilities at their null values $\theta_{C,1} = .70$ and $\theta_{C,2} = .62$. A practical implication of Table 3 is that the design with $2n = 334$ has less than the nominal power to detect alternatives with $\psi > 3.05$, or with $\pi_{C,11} > 0.49$, and greater than the nominal power to detect alternatives with $\psi < 3.05$. Given a desirable alternative $(\theta_{T,1}, \theta_{T,2}) \in \Omega_a$, larger values of $\pi_{T,11}$ are of course more desirable. Thus, if the clinician wishes to specify not only the two-dimensional target points but also the magnitude of the third parameter, $\pi_{T,11}$, then the sample size may be determined to achieve a given power for this three-dimensional alternative. We have used the null value ψ_C in computations given in Tables 1 and 2 as a reasonable compromise.

4.3 *Comparison to One-Dimensional Tests*

An important practical issue is how the sample size required to conduct a test under this two-dimensional formulation compares with what might be required by a more conventional one-sided test formulated if one of the two outcomes is ignored. In the ifosfamide trial, if safety is ignored and a usual one-sided .05-level test of $\theta_{T,1} = \theta_{C,1}$ versus $\theta_{T,1} > \theta_{C,1}$ is conducted based on Y_1, sample sizes of $2n = 60$, 128, 216, or 460 are required to achieve power .80 at alternatives $\theta_{T,1} = .50$, .40, .35, or .30, respectively. Some care must be taken while making such comparisons, however, as they may not always be appropriate. Although each target $\boldsymbol{\xi}$ is meaningful as a two-dimensional alternative compared with (0,0), projecting a given $\boldsymbol{\xi}$ onto one of the two one-dimensional subspaces may not be meaningful. For example, in Design 1 of the gemcy trial, if the lower-right target $\boldsymbol{\xi}_1 = (.258, -.051)$ is projected onto its second component and this is used as the alternative point for constructing a test of $\Delta_2 = 0$ versus $\Delta_2 < 0$ based on Y_2 alone, then the required sample size for a .05-level test with power .80 at $\xi_{1,2} = -.051$ is $2n = 2382$. This is a consequence of the fact that, although (.258,−.051) is a reasonable two-dimensional alternative to (0,0), its second component is so close to 0 that, if considered alone, it is not a practically meaningful alternative to 0 in the Δ_2 subspace. The point is that this two-dimensional target is determined by allowing a small drop in safety as a trade-off for a large improvement in efficacy, and thus considering one entry of $\boldsymbol{\xi}$ without the other destroys this essential feature. Similarly, if one ignores safety and constructs a one-dimensional test of $\Delta_1 = 0$ versus $\Delta_1 > 0$ to achieve power .80 at $\xi_{1,1} = .258$ based on Y_1 alone, this requires only $2n = 94$ patients. Thus, although it may be the case that one entry of a given target is a meaningful alternative to 0 in its one-dimensional subspace, this will not always be the case. The simplest illustration of this phenomenon is a target having one of its two entries equal to 0.

5. Discussion

The primary goal of our proposed method is to construct comparative tests aimed at alternatives that explicitly quantify both efficacy and safety. We find this preferable to the common practice of basing a formal test on efficacy alone while informally monitoring adverse events. The method relies on eliciting the two-dimensional target alternatives from the physician. We have found that oncologists are quite comfortable designing trials this way because it conforms quite naturally to their clinical perspective.

In rapidly fatal diseases, a central issue in specifying the targets is the extent to which failure to achieve the efficacy outcome is associated with early mortality. In acute leukemia, a patient who fails to achieve CR with the first treatment regimen is much less likely to achieve CR with a second treatment, and hence is more likely to die sooner. In other cancers, failure to achieve a response may not have such severe consequences in that a second round of treatment may be nearly as likely as the first to produce a remission. Thus, a higher risk of a severe adverse outcome is more likely to be considered an acceptable trade-off for a given improvement in efficacy in trials of rapidly fatal diseases. This general consideration will be reflected in the way the physician specifies the targets.

To apply the method in settings with nonnegative-valued outcomes $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2})$ for $i = T, C$, if the effects are defined as $\Delta_j = \Pr(Y_{C,j} \leq Y_{T,j}) - \Pr(Y_{T,j} \leq Y_{C,j})$, as suggested in Section 2 and discussed in the Appendix, given the distribution theory for $\widehat{\boldsymbol{\Delta}}$, the main practical issue is eliciting the targets. Depending upon the domain with which the physician is most comfortable, this may be done directly in terms of probabilities that determine the Δ_j's or, alternatively, assuming a Weibull distribution $\Pr(Y_{i,j} > x) = \exp[-(x/\lambda_{i,j})^{\tau_j}]$ for each j with common shape parameter $\tau_j = \tau_{T,j} = \tau_{C,j}$, the effects are given by $\Delta_j = (\lambda_{T,j}^{\tau_j} - \lambda_{C,j}^{\tau_j})/$

Table 3
Effect of association on sample size in Design 1 for the AML trial

$\pi_{C,11}$	0.62	0.57	0.53	0.49	0.45	0.410	0.370	0.32
ψ_C	∞	21.90	7.27	3.05	1.38	0.606	0.224	0
$2n$	412	386	360	334	306	276	244	200

$(\lambda_{T,j}^{\tau_j} + \lambda_{C,j}^{\tau_j})$. If the targets are elicited in terms of medians $\tilde{\mu}_{T,j}$ and $\tilde{\mu}_{C,j}$, then Δ_j may be obtained from the equality $\lambda_{i,j}^{\tau_j} = \tilde{\mu}_{i,j}^{\tau_j}/\log(2)$.

In certain applications, the test may be derived by moving the alternative toward (0,0) in a manner slightly different from moving it along the 45° line. If Ω_a is defined in such a way that there exists a line $L(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)^{\perp}$ through (0,0) that is orthogonal to a line $L(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ on the boundary of Ω_a connecting the two target points $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, then a test may be defined by moving Ω_a toward $\mathbf{0}$ along $L(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)^{\perp}$. Formally, the rejection region for this test is $R = \Omega_a - (c_{\alpha,1}, c_{\alpha,2})$, where $(c_{\alpha,1}, c_{\alpha,2})$ lies on $L(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)^{\perp}$ and $(c_{\alpha,1}^2 + c_{\alpha,2}^2)^{1/2}$ is the largest value such that $\Pr\{\widehat{\boldsymbol{\Delta}} \in R \mid \boldsymbol{\Delta} = \mathbf{0}\} \leq \alpha$. Although this construction is not possible for the ifosfamide trial, it may be applied to the alternatives in Designs 1, 2, or 4 of the gemcy trial, for which the respective sample sizes are 370, 444, and 252. Each of these values is larger than the corresponding value in Table 2, although the difference is nontrivial only for Design 1, where an additional 36 patients would be required.

A class of trade-off alternatives that cannot be accommodated by our procedure is illustrated by the following hypothetical case, suggested by a referee as a reasonable possibility that might arise in practice. Suppose that the clinician specifies the two target points $\boldsymbol{\xi}_1 = (1,-2)$, an improvement of 1 unit in efficacy with a drop of 2 units in safety to achieve it, and $\boldsymbol{\xi}_2 = (-1,2)$, an improvement of 2 units in safety with a drop of 1 unit in efficacy. Because the null hypothesis (0,0) is on the line connecting these two targets, i.e., on the boundary of the convex hull Ω_a of $A(\boldsymbol{\xi}_1) \cup A(\boldsymbol{\xi}_2)$, it is impossible to construct the test. Similarly, the slight modification obtained by specifying $\boldsymbol{\xi}_1 = (1,-2.1)$ produces an alternative hypothesis with (0,0) in its interior. A similar situation may arise if the clinician either has limited knowledge of the mechanism whereby T may provide an improvement over C or is simply being optimistic. In general, such settings may be accommodated by defining the alternative to be $\Omega_a^o = \cup_{k=1}^{K} A(\boldsymbol{\xi}_k)$ rather than the convex hull Ω_a of Ω_a^o. We have not examined the behavior of tests based on Ω_a^o here because we have not yet encountered an application of the type hypothesized above. However, if this second approach is used to construct an alternative Ω_a^o based on the three targets given in the case of Design 1 for the ifosfamide trial (Table 1), then the resulting sample size would be 246, as compared with 226 when Ω_a is the alternative. Although the alternative certainly should include $A(\boldsymbol{\xi})$ for each specified target $\boldsymbol{\xi}$, whether the additional points contained in Ω_a, but not in Ω_a^o, should be included in the alternative is both a philosophical and a practical issue. The use of Ω_a is based on the belief that all convex combinations of desirable alternatives should also be included in the set of alternatives, whereas the use of Ω_a^o accommodates a broader class of problems.

Given that safety is an essential consideration, it follows that the trial design must accommodate situations where T either has unacceptably low efficacy or is unsafe compared with C. A group-sequential design with early termination and acceptance of the null hypothesis in either of these cases provides an additional level of safety. It may also be desirable to stop the trial early with the rejection of the null hypothesis when there is strong evidence that $\boldsymbol{\Delta} \in \Omega_a$. This motivates a group-sequential trial in which the three possible interim decisions are to stop and reject Ω_a, stop and accept Ω_a, or continue to the next stage, with either acceptance or rejection of Ω_a at the end. It seems reasonable that such a group-sequential version of the test may be constructed using the general theory of Jennison and Turnbull (1997). We are currently developing a group-sequential version of the procedure.

As noted in the analysis of the gemcy trial design's sensitivity to ψ_C, a more general approach to the bivariate binary case would be to specify the three-dimensional alternatives in terms of $(\theta_1, \theta_2, \pi_{11})$. A simple way to do this would be to first construct the 2×2 table of null probabilities, then elicit the marginals of the target, and then ask the physician to specify how the four null probabilities should be adjusted to obtain the targeted marginals. This approach may prove useful in settings where a physician has a causal explanation for how an experimental treatment should affect the four elementary outcome probabilities.

ACKNOWLEDGEMENTS

We thank the two referees for their constructive comments and suggestions.

RÉSUMÉ

En plus de leur effet sur la maladie, la plupart des traitements du cancer provoquent des effets indésirables, comme des toxicités, des dommages irréparables sur certains organes ou des décès. Un des problèmes majeurs dans la comparaison d'un nouveau traitement avec un traitement standard est de savoir dans quelle mesure l'augmentation d'effets indésirables est acceptable en regard d'une certaine amélioration de l'efficacité. Nous considérons comme critère de jugement la variable bi-dimensionnelle (efficacité, tolérance) et nous quantifions l'effet du traitement par un paramètre en deux dimensions. Un ensemble de valeurs cibles pour ce paramètre est défini par les cliniciens, ces valeurs correspondant a un gain clinique significatif du nouveau traitement sur le traitement standard. Chacune de ces valeurs est une généralisation bi-dimensionnelle de la différence attendue dans une analyse classique. Nous définissons l'alternative comme l'enveloppe convexe des valeurs cibles dans le plan. La zone de rejet est alors obtenue en comparant l'éloignement de l'aternative au point $(0,0)$, en tenant compte de l'erreur de type I, avec un nombre de sujets calculé qui tienne compte de la puissance. Nous illustrons cette méthode avec deux essais de chimiothérapie.

REFERENCES

Berntsen, J. and Espelid, T. O. (1992). DCUTRI: An algorithm for adaptive cubature over a collection of triangles. *Transactions on Mathematical Software* **18,** 329–342.

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51,** 1372–1383.

Conaway, M. R. and Petroni, G. R. (1995). Bivariate sequential designs for phase II clinical trials. *Biometrics* **51,** 656–664.

Conaway, M. R. and Petroni, G. R. (1996). Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* **52,** 1375–1386.

Cook, R. J. (1994). Interim monitoring of bivariate responses using repeated confidence intervals. *Controlled Clinical Trials* **15,** 187–200.

Cook, R. J. (1996). Coupled error spending functions for parallel bivariate sequential tests. *Biometrics* **52,** 442–450.

Cook, R. J. and Farewell, V. T. (1994). Guidelines for monitoring efficacy and toxicity response in clinical trials. *Biometrics* **50,** 1146–1152.

Drezner, Z. (1993). Computation of the multivariate normal integral. *Transactions on Mathematical Software* **19,** 546.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49,** 741–752.

Jennison, C. and Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92,** 1330–1341.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential monitoring boundaries for clinical trials. *Biometrika* **70,** 659–663.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40,** 1079–1087.

Pocock, S. J., Geller, N. J., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43,** 487–498.

Su, J. Q. and Lachin, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* **48,** 1033–1042.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association* **84,** 776–779.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association* **73,** 339–347.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79,** 653–661.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

Wei, L. J., Su, J. Q., and Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* **77,** 359–364.

Williams, P. L. (1996). Sequential monitoring of clinical trials with multiple survival endpoints. *Statistics in Medicine* **15,** 2341–2357.

Received January 1998. *Revised May* 1998.
Accepted October 1998.

Appendix

For bivariate binary outcomes, denote $\Pr(Y_{i,1} = x \text{ and } Y_{i,2} = y) = \pi_{i,xy}$, $x, y = 0, 1$, so that $\theta_{i,1} = \Pr(Y_{i,1} = 1) = \pi_{i,11} + \pi_{i,10}$ and $\theta_{i,2} = \Pr(Y_{i,2} = 1) = \pi_{i,11} + \pi_{i,01}$, $i = T, C$. For simplicity, temporarily suppress i and consider a single sample of n patients. Let $W = (W_{00}, W_{10}, W_{01}, W_{11})$ be the multinomial vector corresponding to $(\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$. The marginal efficacy and safety counts $X_1 = W_{11} + W_{10}$ and $X_2 = W_{11} + W_{01}$ are binomial with parameters (n, θ_1) and (n, θ_2), respectively, and $\text{cov}(X_1, X_2) = n(\pi_{11}\pi_{00} - \pi_{01}\pi_{10})$. The distribution of $\mathbf{Y}$ may be parameterized by the marginal probabilities θ_1, θ_2 and a third parameter accounting for association, which may be π_{11}, a conditional probability such as π_{11}/θ_1, the correlation $\text{cor}(Y_1, Y_2) = \rho = (\pi_{11} - \theta_1\theta_2)/\{\theta_1(1-\theta_1)\theta_2(1-\theta_2)\}^{1/2}$ or the odds ratio $\psi = \pi_{11}\pi_{00}/\pi_{10}\pi_{01}$.

Denote $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2) = (X_1, X_2)/n$. Because $(n)^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim AN(\mathbf{0}, \Sigma_\theta)$, where

$$\Sigma_\theta = \begin{pmatrix} \theta_1(1-\theta_1) & \pi_{11} - \theta_1\theta_2 \\ \pi_{11} - \theta_1\theta_2 & \theta_2(1-\theta_2) \end{pmatrix}$$

for any suitable transformation g, it follows by the delta method that $(n)^{1/2}\left(g(\widehat{\theta}_1) - g(\theta_1), g(\widehat{\theta}_2) - g(\theta_2)\right) \sim AN(\mathbf{0}, \Sigma_g)$, where

$$\Sigma_g = \begin{pmatrix} \{g'(\theta_1)\}^2\theta_1(1-\theta_1) & g'(\theta_1)g'(\theta_2)(\pi_{11} - \theta_1\theta_2) \\ g'(\theta_1)g'(\theta_2)(\pi_{11} - \theta_1\theta_2) & \{g'(\theta_2)\}^2\theta_2(1-\theta_2) \end{pmatrix} = \begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{12} & \sigma^{22} \end{pmatrix}.$$

Reintroducing the treatment index i and denoting the (a,b)th entry of $\Sigma_{i,g}$ by σ_i^{ab}, as $\widehat{\boldsymbol{\Delta}} = \left(g(\widehat{\theta}_{T,1}) - g(\widehat{\theta}_{C,1}), g(\widehat{\theta}_{T,2}) - g(\widehat{\theta}_{C,2})\right)$, it follows that $(2n)^{1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \sim AN(\mathbf{0}, \Sigma_\Delta)$, where

$$\Sigma_\Delta = 2\begin{pmatrix} \sigma_T^{11} + \sigma_C^{11} & \sigma_T^{12} + \sigma_C^{12} \\ \sigma_T^{12} + \sigma_C^{12} & \sigma_T^{22} + \sigma_C^{22} \end{pmatrix}.$$

For the variance stabilizing transformation $g(\theta) = \sin^{-1}(\theta)^{1/2}$, as $g'(\theta) = \{4\theta(1-\theta)\}^{-1/2}$, it follows that

$$\Sigma_\Delta = \begin{pmatrix} 1 & \rho_\Delta \\ \rho_\Delta & 1 \end{pmatrix},$$

where $\rho_\Delta = (\rho_C + \rho_T)/2$ and ρ_i denotes the correlation between safety and efficacy for $i = C, T$.

An important special case is when the adverse and efficacy events cannot both occur, so that $\pi_{10} = 0$. It follows that $\theta_1 = \pi_{11}$, $\theta_2 = \theta_1 + \pi_{01}$, $\rho = \theta_1(1-\theta_2)/\{\theta_2(1-\theta_1)\}$ and $\psi = \infty$. Thus, safety and efficacy are positively associated. The extreme case occurs when π_{01} is also 0, equivalently $\theta_1 = \theta_2$ and $\rho = 1$, which is the binary case when safety and efficacy are the same event.

To apply the method in settings with nonnegative-valued outcomes, we apply the general theory developed by Wei and Lachin (1984). For a J-variate vector of nonnegative-valued variables $\mathbf{Y} = (Y_1, \ldots, Y_J)$ subject to right censoring, Wei and Lachin (1984) constructed a J-variate statistic, which may be used to construct a variety of two-sample tests. For each $j = 1, \ldots, J$, define the T-versus-C effect $\Delta_j = \Pr(Y_{C,j} \leq Y_{T,j}) - \Pr(Y_{T,j} \leq Y_{C,j})$ and let U_j be a nonnegative-valued censoring variable with $\widetilde{Y}_j = \min(Y_j, U_j)$, and with $d_j = 1$ if $Y_j = \widetilde{Y}_j$ and 0 otherwise. The Wei–Lachin statistic based on data from $2n$ patients is $(X_{1,n}, \ldots, X_{J,n})$, where

$$X_{j,n} = (2n)^{-3/2} \sum_{i=1}^{n}\sum_{i'=1}^{n} \left\{ d_{C,j,i} I\left(\widetilde{Y}_{C,j,i} \leq \widetilde{Y}_{T,j,i'}\right) - d_{X,j,i'} I\left(\widetilde{Y}_{T,j,i'} \leq \widetilde{Y}_{C,j,i}\right) \right\},$$

and $I(\cdot)$ denotes the indicator function. Wei and Lachin showed that $(X_{1,n}, \ldots, X_{J,n}) \sim AN(\mathbf{0}, \Sigma)$ under the null hypothesis and provided a consistent estimator of Σ. Now, as in Thall and Lachin (1988), note that $\widehat{\Delta}_j = \{(2n)^{3/2}/n^2\}\, X_{j,n}$ is a consistent estimator of Δ_j. Applying this with $J = 2$, we have $(n)^{1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \sim AN(\mathbf{0}, 2^3\Sigma)$.