Published in final edited form as:

Comput Stat Data Anal. 2009 January 1; 53(3): 586–595. doi:10.1016/j.csda.2008.07.031.

Drop-the-Losers Design: Binomial Case

Michael W. Sill^{a,1} and Allan R. Sampson^b

^a The GOG Statistical and Data Center – Roswell Park Cancer Institute – Elm & Carlton Streets – and – The Dept. of Biostatistics – University at Buffalo, Buffalo NY USA

^b Department of Statistics, University of Pittsburgh, Pittsburgh PA, USA

Abstract

Drop-the-losers designs were introduced for normal distributions as a method of combining phase II and III clinical trials together under a single protocol with the purpose of more rapidly evaluating drugs by eliminating as much as possible the delays that typically occur between the two phases of clinical development. In the design, the sponsor would administer k treatments along with a control in the first stage. During a brief interim period, efficacy data would be used to select the best treatment (with a rule to deal with ties) for further evaluation against the control in a second stage. At the end of the study, data from both stages would be used to draw inferences about the selected treatment relative to the control with adjustments made for selection in between the two stages. Because the inferences are model based, exact confidence intervals can be determined for the parameter of interest. In the present case, the parameter of concern is the probability of a beneficial response that is dichotomous in nature.

Keywords

Phase II clinical trial; phase III clinical trial; inference after selection; ranking and selection; conditional tests; seamless designs; maximum; 2 stage designs; UMPU; UMPCU; bias correction; binomial distributions

1. Introduction

This paper explores an inferential technique for Drop-the-losers designs (also termed Follow-the-leader designs) in the binomial distribution setting. The utility and difficulties with using these designs for combining phase II and phase III studies was introduced and discussed in some detail for normal data in Sampson and Sill (2005), Bauer (2005), Bretz et al. (2005), Kieser (2005), and Wittes (2005). Furthermore, Chow and Chang (2007, Sec. 1.5.7) have noted the need to obtain results for Drop-the-losers designs for distributions such as the binomial which “match Sampson and Sill’s (2005) results.” One of the greatest attractions to using this design is perhaps the enabling of pharmaceutical sponsors to reduce the total length of time drugs are in clinical trials by shortening the period of time between the selection of the most promising drug (dose or regimen) from phase II studies and the execution of the critical phase III study. This is done by explicitly specifying relatively simple decision rules before the study

¹Corresponding author. GOG Statistical and Data Center, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263. Tel.: (716) 8455702; fax: (716) 8458393. E-mail address: msill@gogstats.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

starts such as selecting the drug (or regimen) that yields the best response rate among a set of k experimental treatments. Such decision rules are reasonable when the targeted drugs are relatively benign and are expected to be used for the treatment of fairly common and mild disorders. By restricting decision criteria in this way, it is possible to model their effects on the statistics used to make inferences on the parameter of interest and make proper adjustments accordingly. Some of advantages to combining all of the data in an analysis of efficacy are the ability to use estimators with smaller variance, create tests of hypotheses with greater power, and construct confidence intervals that are less broad than can be done with phase III data alone. However, considerable care needs to be taken when combining phase II data which stem from a selection rule based on efficacy information in a formal analysis of drug activity. For example, with the Drop-the-losers design where treatment selection is based on that treatment which yields the highest proportion of patients responding to therapy, the maximum likelihood estimator (MLE) of the probability of response for the selected treatment (after the phase II portion of the study is complete) is the sample response rate. This follows because the probabilities of response for all of the treatments are unknown and possibly distinct. Unfortunately, there is a serious risk of heavy bias in the MLE because it is based on the maximum of k observations. This difficulty led to extensive research into the problem of “estimation after selection” in order to find and characterize alternative estimators with reduced bias. A summary of the literature on this subject can be found in Cohen and Sackrowitz (1982) and Sill and Sampson (2007).

The potential for bias in the MLE (or similar natural estimators) creates significant difficulties with incorporating it into analyses of inference in ways that mirror the problems with estimation. For example, the estimated effect of the selected treatment tends to be most exaggerated in cases where the true probabilities of response among the k possible treatments are identical. In such situations, the sample response rate associated with the selected treatment is likely greater than the common true probability of response due to chance variability and the selection rule applied. Likewise, testing procedures that do not correct for this problem will naively use the sample response rate to erroneously reject the null hypothesis (when it is true) with size greater than its advertised level of significance. Additionally, confidence intervals will have probabilities of coverage that are lower than advertised because the bias of the MLE will yield lower bounds on the interval that are greater in value than expected. On the other extreme, if in fact there is a vastly superior treatment to the other treatment, this will be selected almost always for continuation into the second stage. Therefore, the sample response rate will be largely unbiased as it varies around the true probability of response. Additionally, tests of hypotheses and confidence intervals will have accurate operating characteristic and probabilities of coverage because the same treatment is nearly always selected. In this case, a data analyst would make few errors from an assumption of no selection.

The difference in estimation and inference between these two extremes can be rapidly summarized with plots of normalized likelihood functions as advocated by Blume (2002) with a simple example (see Figure 1). If the selected treatment demonstrated 20 responses out of 50 cases and was vastly superior to all of the other treatments, then it is appropriate to use the likelihood function based on the simple binomial distribution that is traced in Figure 1 with a solid line. In that case, the correct MLE is 40% with a 95% confidence interval bounded by 26% and 55%. On the other hand, if the selected treatment is from, say 5 treatments with identical probabilities of response, then the appropriate likelihood function (as derived from the distribution of the maximum of 5 binomially distributed random variables) is the one traced with a dashed line in Figure 1. In that case, the actual MLE is 33% and the 95% confidence interval is bounded by 23% and 41%.

As can be seen from the example presented above, the true constellation of response rates can have a considerable influence on the results of an analysis. Clearly, in realistic settings the

exact probabilities of response are unknown for all of the treatments assessed in the study, so it is not possible to know which extreme is most appropriate (if one wanted to analyze the data by one of these techniques). In most cases, the truth will be somewhere between these two extremes, so a procedure that makes relative adjustments based on the available evidence as to which constellation is more likely would be a desirable property.

Finally, this potentially biased phase II information needs to be suitably adjusted and combined with the unbiased phase III data to create a pooled statement with regard to an estimate, hypothesis test, or confidence interval. A significant advance to the solution of these problems was given by Cohen and Sackrowitz (1989) who developed an unbiased estimator for the univariate normal distribution with σ known and unknown as well as for the gamma distribution. Their estimator is said to be conditionally UMVU (i.e. uniformly minimum variance conditionally unbiased or UMVCU). That is, after conditioning on the order of the first stage outcomes, e.g. $\{\mathbf{X}: X_1 > X_2 \dots > X_k\}$, their estimator was UMVU. Their estimator was also unconditionally unbiased (as long as the conditioning sets for which their estimators were based formed a partition of the sample space), but lacked the UMVU property unconditionally. In spite of this problem, their estimator appears to perform quite competitively against the MLE unconditionally. Tappin (1992) extended this methodology to the estimation of the probability of response for binomial distributions. She demonstrated that when randomized decision rules were used to break ties, no UMVCUE existed. On the other hand, when a preference between treatments was stated before the trial started and used to break ties, then a UMVCUE existed. Sampson and Sill (2005) used a similar methodology provided by Cohen and Sackrowitz (1989) to develop a uniformly most powerful conditionally unbiased test (UMPCU) for the univariate normal distribution. However, as with many tests that are based on conditioning arguments (e.g. Fisher's Exact Test), this approach is a bit conservative unconditionally (see discussion by Bauer (2005)). We would therefore recommend using more powerful and flexible procedures such as those outlined in Bauer and Kieser (1999) for the purpose of hypothesis testing. Nevertheless, if the Drop-the-losers design is followed, then the methods outlined in Sampson and Sill (2005) and here can be used to construct exact confidence intervals with the advertised probabilities of coverage. Such analyses can be used to supplement the primary analysis, which typically involve simple hypothesis testing. Sill and Sampson (2007) extended the method of Cohen and Sackrowitz (1989) to a bivariate normal distribution with the covariance matrix known. Treatment selection was made with a surrogate variable but the primary analysis focused on another variable whose determination was delayed (for example).

The problem of using surrogates for treatment selection is very common in clinical trials. Characterization of the unbiased estimator developed in that paper demonstrated particular competitiveness with the MLE when the sample size of the first stage was small relative to the second stage which corresponded naturally to phase II and phase III trials.

Although the method is not recommended for hypothesis testing but rather for use in constructing confidence intervals, it is easier to compare procedures in the setting of hypothesis testing (for example by comparing power and true size). Therefore, much of the discussion in this paper will involve the topic of hypothesis testing, and the results should be readily translated into those involving confidence intervals through the duality property of hypothesis testing and confidence intervals (e.g. see DeGroot (1975), pp 481–3).

2. Drop-the-losers design for binomial distributions

In Stage A, which corresponds to phase II trials, the sponsor would administer k experimental treatments, $\tau_1, \dots, \tau_k$ along with the control treatment, τ_0 , to n_A patients per treatment. Data on patient responses would be collected after Stage A. The total number of responses to treatment

τ_i would be X_i with the assumption that $X_i \sim \text{Bin}(n_A, \pi_i)$ where π_i is the probability of response for treating a person on this treatment. The sponsor would like to select the treatment with the highest response rate (i.e. the treatment associated with $\max\{\pi_1, \dots, \pi_k\}$) for continuation into the second stage, called Stage B. Since the true response rates are unknown, the sponsor selects the treatment associated with the highest observed number of responses, τ_M where $X_M = \max\{X_1, \dots, X_k\}$. If τ_M is not unique because of ties, then we would recommend selecting one treatment based on a preference established before the trial opened to enrollment. This problem is discussed further in Section 7, but for ease of explanation, we will assume that τ_M is unique.

Once a particular treatment is selected for continuation, it would be administered to n_B additional patients along with the control treatment. At the end of Stage B, the number of responses to the selected treatment within this phase of the study, Y , would be determined. We note that Y has a binomial distribution, i.e. $Y \sim \text{Bin}(n_B, \pi_M)$. The control treatment would be administered to a total of n_0 patients, and the number of responses in the control, Y_0 , would also have a binomial distribution, i.e. $Y_0 \sim \text{Bin}(n_0, \pi_0)$.

3. Naïve inference and simple corrections for inference

The naïve procedure assumes that no treatment selection is taking place and to erroneously presume that $X_M \sim \text{Bin}(n_A, \pi_M)$. In that case, Fisher's Exact Test could be used to test $H_0: \pi_M = \pi_0$. Defining $Z = X_M + Y$:

$$P(Z|Z+Y_0=t) = \frac{\binom{n_0}{t-Z} \binom{n_A+n_B}{Z}}{\binom{n_0+n_A+n_B}{t}}.$$

Note that Z is restricted according to: $\max\{0, t-n_0\} \leq Z \leq \min\{n_A+n_B, t\}$. When the alternative hypothesis is $H_1: \pi_M > \pi_0$, the null hypothesis is rejected at the α level of significance when Z is greater than the $(1-\alpha)$ quantile of the distribution. A Bonferroni like correction has been proposed to counter the deficiencies of the naïve procedure. It uses the naïve distribution but will not reject the null hypothesis unless Z is greater than the $(1-\alpha/k)$ quantile of the null distribution. Another approach, termed the Stage B procedure, drops the Stage A data and utilizes only the Stage B data. The procedure has the advantage of being inferentially "unbiased" but fails to incorporate all of the relevant information about drug efficacy. The Stage B procedure also uses Fisher's Exact Test according to the following distribution:

$$P(Y|Y+Y_0=t) = \frac{\binom{n_0}{t-Y} \binom{n_B}{Y}}{\binom{n_0+n_B}{t}}.$$

4. The UMPCU approach

Another approach to assessing the selected treatment against the control is with a method similar to Cohen and Sackrowitz (1989). They derive in their paper a 2-stage UMVCUE by conditioning on the order of the first stage outcomes for the k normally distributed variables such as $Q = \{X: X_1 > X_2 > X_3 > \dots > X_k\}$. This conditioning simplifies the estimation problem considerably by determining (with certainty) the treatment selected for continuation, which brings the distribution of statistics into the exponential family. A UMVCUE of the parameter of interest, μ_1 , is then found by Rao-Blackwellizing on the conditionally unbiased estimator,

Y (see Casella and Berger (1990), p316). For the purpose of statistical inference, the use of the exponential family of distributions is exploited again with a theorem provided by Lehmann (1959, p134–39). In particular, a uniformly most powerful (UMP) test for the parameter of interest can be constructed with a distribution found by conditioning on the actual values of the sufficient statistics for the nuisance parameters within the problem (i.e. $X_2 = x_2, X_3 = x_3, \dots, X_k = x_k$ for $\pi_2, \dots, \pi_k$). This second conditioning has the additional benefit of causing the nuisance parameters to explicitly drop from the distribution. The main part of Lehman's theorem (Theorem 3, p136) states that the approach is unconditionally UMP, but the test is properly labeled a conditionally UMP procedure in this context because the method requires a conditioning on the order of the first stage outcomes such as the event Q or $X_1 > \max\{X_2, \dots, X_k\}$.

The objective of the following paragraphs is to use this theory to derive a UMPCU test for a parameter that expresses a deviation between π_1 and π_0 .

We follow to an extent the notation of Cohen and Sackrowitz (1989). For simplicity, we assume the following ordered outcome after Stage A, $Q = \{X: X_1 - 1 \geq X_2 \geq X_3 \geq \dots \geq X_k\}$. This event implies that τ_1 is selected for continuation and that τ_1 its Stage A number of responses is the unique maximum among the k treatments. The binomial distribution belongs to the exponential family, so comparisons between two populations usually involve inference on the natural parameter, the log odds:

$$\theta_i = \ln \left(\frac{\pi_i}{1 - \pi_i} \right).$$

In general we will be interested in testing the following hypotheses: $H_0: \theta_1 - \theta_0 \leq \Delta_{10}$ versus $H_1: \theta_1 - \theta_0 > \Delta_{10}$. The first step is to consider the joint distribution of $\mathbf{X}, Y|Q$:

$$f_Q(\mathbf{X}, Y) = K(\pi)^{-1} I_Q(\mathbf{X}) \binom{n_A}{X_1} \pi_1^{X_1} (1 - \pi_1)^{n_A - X_1} \binom{n_B}{Y} \pi_1^Y (1 - \pi_1)^{n_B - Y} \Pi,$$

where $K(\pi)$, with $\pi = (\pi_1, \pi_2, \dots, \pi_k)$, is the probability of observing event Q , $I_Q(\mathbf{X})$ is the indicator function on the set Q , and Π is:

$$\Pi = \prod_{i=2}^k \binom{n_A}{X_i} \pi_i^{X_i} (1 - \pi_i)^{n_A - X_i}.$$

Noting that $Z = X_1 + Y$, we can rewrite the distribution in the following form:

$$f_Q(\mathbf{X}, Z) = K(\pi)^{-1} I_Q(\mathbf{X}) \binom{n_A}{X_1} \binom{n_B}{Z - X_1} \pi_1^Z (1 - \pi_1)^{n_A + n_B - Z} \Pi.$$

The statistics, $(Z, X_2, \dots, X_k) = (Z, \mathbf{X}^*)$ are conditionally sufficient for $\pi = (\pi_1, \pi_2, \dots, \pi_k)$, so we wish to integrate out X_1 from the above distribution by summing over its support. Note that the binomial assumption and Q require $X_2 + 1 \leq X_1 \leq n_A$. Remembering that $X_1 = Z - Y$ also requires $Z - n_B \leq X_1 \leq Z$. Therefore,

$$f_Q(\mathbf{X}^*, Z) = K(\pi)^{-1} \cdot \Pi \cdot I_Q^*(\mathbf{X}^*) \cdot \pi_1^Z (1 - \pi_1)^{n_A + n_B - Z} \cdot \sum_{X_1 = \max\{X_2 + 1, Z - n_B\}}^{\min\{n_A, Z\}} \binom{n_A}{X_1} \binom{n_B}{Z - X_1},$$

where $I_Q^*(\mathbf{X}^*)$ is the indicator function for $(X_2, \dots, X_k)$ on $Q = \{\mathbf{X}^*: X_2 \geq \dots \geq X_k\}$. The above distribution can be used to derive the conditional distribution for drawing inferences about π_1 . For details see Sill (2000). Noting that Y_0 is independent of $(Z, \mathbf{X}^*)$, the joint distribution of $Z, \mathbf{X}^*, Y_0|Q$ can be written as:

$$f_Q(\mathbf{X}^*, Z, Y_0) = f_Q(\mathbf{X}^*, Z) \cdot \binom{n_0}{Y_0} \pi_0 (1 - \pi_0)^{n_0 - Y_0}.$$

Observing that Y_0 is sufficient for π_0 (and again that Z is sufficient for π_1), we write part of the distribution in exponential form:

$$K^*(1 - \pi_1)^{n_A + n_B} \binom{n_0}{Y_0} (1 - \pi_0)^{n_0} h(Z) \exp \left\{ \ln \left(\frac{\pi_1}{1 - \pi_1} \right) Z + \ln \left(\frac{\pi_0}{1 - \pi_0} \right) Y_0 \right\},$$

where $K^* = K(\pi)^{-1} \cdot \Pi \cdot I_Q^*(\mathbf{X}^*)$ and:

$$h(Z) = \sum_{X_1 = \max\{X_2 + 1, Z - n_B\}}^{\min\{n_A, Z\}} \binom{n_A}{X_1} \binom{n_B}{Z - X_1}.$$

We rewrite the exponential term, $Z\theta_1 + Y_0\theta_0 = Z(\theta_1 - \theta_0) + (Y_0 + Z)\theta_0$. Guided by the right hand side, we define for ease of exposition $T = Y_0 + Z$. Therefore Z is sufficient for $\Delta_1 = \theta_1 - \theta_0$. The distribution of Z and T is:

$$f_Q(\mathbf{X}^*, Z, T) = K^*(1 - \pi_1)^{n_A + n_B} \binom{n_0}{T - Z} (1 - \pi_0)^{n_0} h(Z) \exp \{Z\Delta_1 + T\theta_0\}.$$

Applying Lehmann (Theorem 3, p136) to find the conditional distribution of $Z | \mathbf{X}^*, T, Q$, we first note restrictions on Z , e.g. $X_2 + 1 \leq Z \leq n_A + n_B$. Also from the definition of T , we find $T - n_0 \leq Z \leq T$. To obtain the joint distribution of $\mathbf{X}^*, T$, we marginalize the joint distribution of $\mathbf{X}^*, Z, T$ over Z :

$$f_Q(\mathbf{X}^*, T) = \sum_{Z = \max\{X_2 + 1, T - n_0\}}^{\min\{n_A + n_B, T\}} f_Q(\mathbf{X}^*, Z, T).$$

The conditional distribution used to define the UMPCU test is found with $f_Q(\mathbf{X}^*, Z, T) / f_Q(\mathbf{X}^*, T)$. The conditional distribution simplifies considerably with $K^*(1 - \pi_1)^{n_A + n_B} (1 - \pi_0)^{n_0} \exp\{T\theta_0\}$ explicitly dropping from the expression to obtain

$$f_Q(Z|\Delta_1, \mathbf{X}^*, T) = C_N \binom{n_0}{T-Z} \exp\{Z\Delta_1\} = \sum_{X_1=\max\{X_2+1, Z-n_B\}}^{\min\{n_A, Z\}} \binom{n_A}{x_1} \binom{n_B}{Z-x_1}, \quad (1.1)$$

where:

$$C_N^{-1} \sum_{Z=\max\{X_2+1, T-n_0\}}^{\min\{n_A+n_B, T\}} \binom{n_0}{T-Z} \exp\{Z\Delta_1\} = \sum_{X_1=\max\{X_2+1, Z-n_B\}}^{\min\{n_A, Z\}} \binom{n_A}{x_1} \binom{n_B}{Z-x_1}, \quad (1.2)$$

that is, C_N is the normalization constant. Like other distributions derived with this technique, the conditional distribution depends only on X_2 and not $X_3, \dots, X_k$ nor k (see Cohen and Sackrowitz (1989), Tappin (1992), Sampson and Sill (2005), and Sill and Sampson (2007)).

5. Using the conditional distribution to draw inferences

The various approaches to drawing inferences will be compared in a hypothesis testing paradigm. Typical hypotheses of interest are $H_0: \Delta_1 \leq \Delta_{10}$ versus $H_1: \Delta_1 > \Delta_{10}$ and $H_0: \Delta_1 = \Delta_{10}$ versus $H_1: \Delta_1 \neq \Delta_{10}$. Ordinarily the null hypothesis is rejected for the one-sided alternative hypothesis when $Z_{obs} \geq Z_U$ where Z_U is found so that:

$$\alpha \geq \sum_{Z=Z_U}^{\min\{n_A+n_B, T\}} f_Q(Z|\Delta_{10}, \mathbf{X}^*, T).$$

To use the distribution for the construction of $(1 - \alpha) \cdot 100\%$ confidence intervals for Δ_1 , $[\Delta_L, \Delta_U]$, functions involving Δ_1 can be used as follows:

$$p_1(\Delta_1) = \sum_{Z=\max\{X_2+1, T-n_0\}}^{Z_{obs}} f_Q(Z|\Delta_1, \mathbf{X}^*, T),$$

and

$$p_2(\Delta_1) = \sum_{Z=Z_{obs}}^{\min\{n_A+n_B, T\}} f_Q(Z|\Delta_1, \mathbf{X}^*, T).$$

To find the lower bound of a one-sided confidence interval, Δ_L would be found such that $p_2(\Delta_L) = \alpha$. In this case, $\Delta_U = 1$. Finally, bounds on an interval for a 2-sided confidence interval, Δ_L and Δ_U , can be found by solving the following relationships $p_2(\Delta_L) = \alpha/2$ and $p_1(\Delta_U) = \alpha/2$ respectively.

6. Conditional power comparisons

The behavior of the four competing tests (the UMPCU, Naïve, Stage B, and Bonferroni) was compared with conditional power curves. For the purposes of discussion, a test will be said to have conditional α level of significance if the $P(\text{Reject } H_0) \leq \alpha$ when $\Delta_1 \in \Omega_0$ for all $\pi^* = (\pi_2,$

$\pi_3, \dots, \pi_k$) where Ω_0 is the null parameter space. The size of the procedure will be defined as $\sup_{\Delta_1 \in \Omega_0} \{P(\text{Reject } H)\}$ for a fixed π^* . For ease of interpretation, the power curves were plotted on the π_1 scale. The same qualitative conclusions would be obtained on the log odds scale.

The results in Figure 2 through 6 and Table 1 were obtained from 10,000 simulations at each point in the design and parameter space using SAS 8.02. The power curves were constructed from 13 sample points of π_1 in the parameter space for π_1 . In all of the cases examined, the total sample size in each arm that continues through Stage B is 100 patients. The tests were compared using a one-sided level of significance of 5% (i.e. $H_0: \Delta_1 \leq 0$ versus $H_1: \Delta_1 > 0$, with advertised $P(\text{Type I Error}) \leq 0.05$). The purpose of using a 5% level of significance was arbitrary. Limited investigations at the more conventional level of 2.5% indicated the same qualitative conclusions. Figure 2 looks at a case with $k = 2$ where both experimental arms are of relatively comparable performance. In this particular case, the interim analysis is performed after 25 patients are accrued to each of the experimental arms (i.e. $n_A = 25$). The first notable feature of the figure is the fact that the Naïve procedure is the most powerful of all. This feature should be tempered by the fact that the size of Naïve test is approximately 6% even though it is advertised at 5%. Of the three other procedures, the UMPCU test is the most powerful with an estimated size of 4% (theoretically, it is exactly 5% but due to data discreteness, it can deviate considerably in realized designs).

In Figure 3, the probability of response for the rejected treatment, π_2 , is 10% while the selected treatment has a probability of response near 50%. This is the case where the selected treatment is vastly more effective than the other treatments. In this case, the Naïve and UMPCU procedures have nearly identical power curves. These two procedures are the most powerful ones in the set.

Figure 4 shows an example that demonstrates the fact that the Naïve and Bonferroni procedures are not conditional α -level tests. The Naïve and Bonferroni procedures have sizes of about 10% and 6% respectively. Of the two procedures that are truly α -level, the UMPCU procedure is slightly more powerful. Although the Bonferroni procedure is not an α -level test conditionally, it is still an α -level test unconditionally. The unconditional probability of selecting treatment τ_1 when the other treatment is truly outperforming it by 20% is relatively small, so the conditional observations are peculiar because it is usually the case that $X_1/n_A > X_2/n_A \gg \pi_1$ which gives a substantial boost to the apparent efficacy of the selected treatment.

Figures 2, 5, and 6 show the impact of delaying the time to select the treatment from 25% to 50% and 75% (respectfully) of the total experimental arm size. The effects seem to be fairly modest in the Naïve, UMPCU, and Bonferroni tests but have an obviously negative effect on the Stage B test.

Table 1 examines the sizes of the selected tests as the design parameters change over a range of values. For the case where $k = 2$, all of these tests are relatively insensitive to the timing of the interim analysis as long as $\pi_2 \leq \pi_1 \in \Omega_0 = \{\pi_1: \pi_1 - \pi_0 \leq 0\}$. When the selected treatment is inferior to the rejected treatment, the size of the Naïve and Bonferroni tests become highly dependent on the timing of the interim analysis, increasing drastically as the decision to select is delayed. For the UMPCU test, on the other hand, the simulations showed a resilience to the interim timing. The size of the Naïve and Bonferroni tests have a tendency to increase as k increases as suggested by the last row of the table.

7. Dealing with ties in Stage A

Because the responses are discrete random variables, it is possible that several treatments could yield responses equivalent to $\max\{X_1, \dots, X_k\}$ so that τ_M is not unique. This fact leads to special problems. If the investigators are interested in only one treatment for continuation into Stage

B, then they must rely on ancillary decision rules to make their selection. The effect of some of these rules on the estimator of the probability of response was investigated in Tappin (1992). One rule involved a random selection from the tied treatments. Another algorithm used the smallest index selection rule (e.g. if $X_2 = X_4 = X_7 = \max\{X_1, \dots, X_k\}$, then τ_2 was selected since $2 = \min\{2, 4, 7\}$). She proved that no UMVUE existed for the first rule while the second yielded a UMVUE.

If investigators wish to apply the proposed design and use the UMPCU method for inference, then we suggest establishing a preference for regimens before the study is carried out. For example, if the treatment regimens involve the same drug but different schedules of administration, then regimens with more convenient schedules should be given higher preference. Then if two regimens yield the same maximum number of responses, the more convenient regimen will be selected for continued investigation. Such a priori decision rules can make clinical sense and enable investigators to apply the inferential techniques discussed in this paper and others.

To apply this method for the case where $k = 2$, treatment τ_1 is selected in the event that $Q_1 = \{\mathbf{X}: X_1 \geq X_2\}$ occurs. The corrected distribution that has the UMPCU property can then be obtained by substituting X_2 in place of $X_2 + 1$ in (1.1) and (1.2). The only cases where τ_2 is selected is when its response rate is strictly greater than the preferred treatment's response rate, so the corrected distribution would be based on $Q_2 = \{\mathbf{X}: X_2 > X_1\}$. In this case, (1.1) and (1.2) would be used with proper relabeling of X_2 for X_1 and vice versa.

The situation is a little more complex for the case where $k > 2$. Assuming treatment preferences are established in such a way as to be equivalent to the smallest index selection rule, treatment τ_1 is selected as long as $X_1 \geq \max\{X_2, \dots, X_k\}$, so the corrected distribution would be based on $Q_1 = \{\mathbf{X}: X_1 \geq \max\{X_2, \dots, X_k\}\}$. This distribution can be found by substituting $\max\{X_2, \dots, X_k\}$ in place of $(X_2 + 1)$ in (1.1) and (1.2). Treatment τ_2 is selected as long as $X_2 \geq \max\{X_3, \dots, X_k\}$ and $X_2 > X_1$. Therefore the lower bound of X_2 is $\max\{X_3, \dots, X_k, X_1 + 1, Z - n_B\}$. The corrected distribution could be obtained from (1.1) and (1.2) if $\max\{X_3, \dots, X_k, X_1 + 1\}$ was substituted for $(X_2 + 1)$ with proper relabeling. If treatment τ_3 is selected, it would require a response $X_3 \geq \max\{X_4, \dots, X_k\}$, $X_3 > X_2$, and $X_3 > X_1$, so the corrected distribution could be obtained by substituting $\max\{X_4, \dots, X_k, X_1 + 1, X_2 + 1\}$ for $(X_2 + 1)$ in (1.1) and (1.2) with proper relabeling. The pattern continues in this fashion through τ_k . It is noted that τ_k would be selected only if its response was uniquely greater than all others so the distribution derived in section 4 could be used with $\max\{X_1 + 1, X_2 + 1, \dots, X_{k-1} + 1\}$ substituted in place of $(X_2 + 1)$.

8. Discussion

It is often considered acceptable or reasonable to favor a biased estimator over an unbiased estimator when other characteristics are considered such as the mean squared error (MSE) and shown to favor the biased estimator. The ideas behind such justifications are that the biased estimator, while having a tendency to overestimate (or underestimate) the parameter, tends to be nearer to the true parameter by some commonly accepted measure of "closeness" (e.g. MSE). Therefore, an unbiased estimator has to perform quite comparably to a relatively good (though biased) competitor before the competitor is abandoned. To put it another way for the problem of estimation, small differences in expectation from the estimator and the parameter are acceptable if its variance around the parameter is reasonably small.

However, when it comes to inference, this is not the case at all. Any degree of bias in an inferential procedure is widely considered unacceptable no matter how powerful it is in the alternative parameter space. For example, it is conceivable that extensive debate would ensue from an important study that published a result with an advertised level of significance of say

5% but in reality may have had only 8% or 10% level of significance. This rigidity stems in part from the consequences of decisions based on inference. In the arena of clinical trials, these include substantial changes in clinical practice and increases in the demand for new drugs worth potentially billions of dollars. Therefore, it is imperative to have an accurate view of what the data are saying about the parameter of interest before a decision is made. If information is being incorporated into an overall statement of efficacy that is known (or suspected) to be biased in some way, then particular care needs to be taken to assure it does not distort or corrupt the final analysis in the slightest way. Therefore, even though the naïve procedure often yields tests with true sizes only slightly larger than its advertised level of significance in realistic situations (e.g. the case where $\pi_2 = 0.50$ in Table 1), this error would necessitate a correction before the design could be applied in practice.

The use of confidence intervals (CIs) significantly enhances our understanding of the range of possible values of the parameter. They can be used to judge the best and worst case scenarios by assuming the parameter is equal to the lower or upper bounds of a CI, and calculating the effects of these assumptions on important events of interest. Such an approach can be useful in an analysis for risk management, for example.

Additionally, CIs can be used to shed light on a study that failed to demonstrate superior efficacy according to the primary analysis. Often studies are designed for the detection of superior efficacy because the sample size requirements are generally not as large as those that are designed to demonstrate non-inferiority. However, if a study was not able to show that the new drug had greater effectiveness in comparison to the standard of care, the question naturally turns toward the less ambitious one of non-inferiority. CIs can be used to quickly assess non-inferiority by comparison of the lower bound of the interval, Δ_L , to the non-inferiority margin. If $\Delta_L > \Delta_{10}$, then the new drug may be considered non-inferior. However, if $\Delta_L < \Delta_{10}$, then non-inferiority to the standard cannot be established.

The incorporation of biased phase II data into a phase III trial using Drop-the-losers designs tend to yield corrected lower bounds on CIs that are smaller in value than the naïve method, which is especially important if an assertion of non-inferiority is going to be made without bias.

The only dependence of the conditional distribution on the selection rule is through X_2 . At an intuitive level, this is a strange result. All of the information about $\pi_2, \dots, \pi_k$ and k that is required to keep the estimators and procedures unbiased is somehow transmitted through the single statistic, X_2 . Some of the benefits of this fact are simpler distributions, estimators, and inferential procedures. However, one cannot escape the thought that important information could be disregarded by this approach. This unease was expressed in Cohen and Sackrowitz's original work. Yet the method seems to perform adequately in some but not all design situations (e.g. when n_A is small relative to n_B).

The flexibility of the technique may be greater than the implications provided earlier. For example, it may not be necessary for all of the first stage sample sizes to be equal to n_A . As a matter of fact, there are absolutely no requirements with regard to the other distributions other than $X_1 > X_2$ because the entire distributions of these statistics drop from the problem. This fact opens the possibility that the method is considerably adaptable to many circumstances. Additional work needs to be conducted to determine the robustness of the method to deviations from the originally intended design. It seems likely that the method is reasonably robust to minor deviations in the first stage sample sizes. However, if the first stage sample sizes were determined for a selection design, then the probability of selecting the correct treatment is not robust against deviations from the planned n_A when the outcomes are discrete (e.g. the $P(X_1$

$> X_2$) depends on their respective sample sizes in addition to π_1 and π_2). These problems need to be resolved with additional research.

Extension of the method based on Sill and Sampson (2007) from estimation to inference on a primary variable is certainly possible by developing a UMPCU procedure when selection is done with a surrogate where both variables are normally distributed. Actual implementation of the method may be computationally intensive, making its use less attractive. Additional research may show that using the unbiased estimator for inference may be preferable in practical applications.

Because many clinical trials use a response type variable in the phase II setting (e.g. objective tumor response) but use a normally distributed primary variable in the phase III setting, it may be of interest to pursue estimation and inference in this specific situation. However, mathematical difficulties may lie in specifying a relationship between the probability of response, π , in the surrogate and the mean of the primary outcome, μ . Instead it may be simpler and suitable enough to use a normal approximation for the binomial distribution (provided sample sizes are sufficiently large).

9. Conclusions

Given the importance of maintaining the integrity of the operating characteristics of the design (i.e. having a design that accurately specifies probabilities of type I and type II errors), corrections to inferential procedures are necessary when Drop-the-losers designs are used. The UMPCU method provides a method of correction that is conditionally guaranteed to be more powerful than the Stage B method. However, like many conditional tests, it has been criticized as being a bit conservative. Yet, the UMPCU method can be inverted to form confidence intervals which can enrich an analysis of efficacy. Additionally, these confidence intervals can assess hypotheses related to questions of non-inferiority in cases where the primary analysis fails to reject in favor of superior effectiveness.

Acknowledgments

The authors wish to thank Dr. Mark Brady for his insights with regard to non-inferiority studies and the referees for their careful review and helpful suggestions, which led to an improved paper. The research of the first author was supported in part by National Science Foundation Grant No. DMS9704400 and a National Cancer Institute grant to the Gynecologic Oncology Group Statistical and Data Center (5 U10 CA37517). The research of the second author was supported in part by NIH Grant 45156.

References

- Bauer P. Discussion. *Biometrical J* 2005;47:269–270.
- Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in medicine* 1999;18:1833–1848. [PubMed: 10407255]
- Blume JD. Tutorial in biostatistics: likelihood methods for measuring statistical evidence. *Statist med* 2002;21:2563–2599.
- Bretz F, Strassburger K, Maurer W. Discussion. *Biometrical J* 2005;47:274–275.
- Casella, G.; Berger, RL. *Statistical inference*. Wadsworth, Inc.; Belmont, California: 1990.
- Chow, S-C.; Chang, M. *Adaptive design in clinical trials*. Chapman & Hall/CRC; FL: 2007.
- Cohen, A.; Sackrowitz, HB. A decision theory formulation for population selection followed by estimating the mean of the selected population. In: Gupta, SS.; Berger, JO., editors. *Statistical decision theory related topics–IV*. Vol. 2. Springer-Verlag; New York: 1988. p. 33–36.
- Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Statist probab lett* 1989;8(3):273–278.

- DeGroot, MH. Probability and statistics. Vol. 2. Addison-Wesley Publishing Company; Reading, Massachusetts: 1975.
- Kieser M. Discussion. *Biometrical J* 2005;47:271–273.
- Lehmann, EM. Testing statistical hypotheses. John Wiley & Sons, Inc.; New York: 1959.
- Sampson AR, Sill MW. Drop-the-losers design: normal case. *Biometrical J* 2005;47:257–268.
- Sill MW, Sampson AR. Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. *Communications in statistics – theory and methods* 2007;36:801–813.
- Sill, M. PhD Dissertation. Department of Statistics, University of Pittsburgh; PA., USA: 2000. Estimation and Inference for Drop-the-Losers Designs.
- Tappin L. Unbiased estimation of the parameter of a selected binomial population. *Commun statist theor meth* 1992;21(4):1067–1083.
- Wittes J. Discussion. *Biometrical J* 2005;47:276–277.

Figure 1. Normalized likelihood functions for the probability of response in the case of no selection (solid line) and in the case of treatment selection based on the maximum response rate from 5 equally performing treatments (dashed line) when 20 responses are observed in 50 people.

Conditional $\alpha=0.05$ Power Curve For
Non-selected Treatment τ_2 Equal to Control; Selection after 25% of the Patients

$$H_0: \Delta_1 \leq 0 \text{ versus } H_1: \Delta_1 > 0$$

$$k=2, n_0=100, n_A=25, n_B=75, \pi_2=0.50, \pi_0=0.50$$

$$\text{Under The Condition } Q = \{\mathbf{X}: X_1 > X_2\}$$

Figure 2.

Conditional $\alpha=0.05$ Power Curve For
Non-selected Treatment τ_2 Worse than Control; Selection after 25% of the Patients

$$H_0: \Delta_1 \leq 0 \text{ versus } H_1: \Delta_1 > 0$$

$$k=2, n_0=100, n_A=25, n_B=75, \pi_2=0.10, \pi_0=0.50$$

$$\text{Under The Condition } Q = \{\mathbf{X}: X_1 > X_2\}$$

Figure 3.

Conditional $\alpha=0.05$ Power Curve For
Non-selected Treatment τ_2 Better than Control; Selection after 25% of the Patients

$$H_0: \Delta_1 \leq 0 \text{ versus } H_1: \Delta_1 > 0$$

$$k=2, n_0=100, n_A=25, n_B=75, \pi_2=0.70, \pi_0=0.50$$

$$\text{Under The Condition } Q = \{\mathbf{X}: X_1 > X_2\}$$

Figure 4.

Conditional $\alpha=0.05$ Power Curve For
Non-selected Treatment τ_2 Equal to Control; Selection after 50% of the Patients

$$H_0: \Delta_1 \leq 0 \text{ versus } H_1: \Delta_1 > 0$$

$$k=2, n_0=100, n_A=50, n_B=50, \pi_2=0.50, \pi_0=0.50$$

$$\text{Under The Condition } Q = \{\mathbf{X}: X_1 > X_2\}$$

Figure 5.

**Conditional $\alpha=0.05$ Power Curve For
Non-selected Treatment τ_2 Equal to Control; Selection after 75% of the Patients**

$$H_0: \Delta_1 \leq 0 \text{ versus } H_1: \Delta_1 > 0$$

$$k=2, n_0=100, n_A=75, n_B=25, \pi_2=0.50, \pi_0=0.50$$

$$\text{Under The Condition } Q = \{\mathbf{X}: X_1 > X_2\}$$

Figure 6.

Table 1

Simulated Sizes of UMPCU, Naïve, and Bonferroni Procedures under $Q^{(I)}$

($n_0 = 100$, $\pi_0 = 0.50$, $\Delta_{10} = 0$)

k=2	0.10			$\pi_2^{(2)}$			0.70		
	$C^{(3)}$	$N^{(4)}$	$B^{(5)}$	C	N	B	C	N	B
$n_A=25$, $n_B=75$	0.035	0.035	0.019	0.041	0.060	0.035	0.037	0.096	0.056
$n_A=50$, $n_B=50$	0.037	0.037	0.019	0.037	0.064	0.034	0.037	0.176	0.110
$n_A=75$, $n_B=25$	0.039	0.039	0.019	0.035	0.068	0.038	0.037	0.301	0.204
k=3	C	N	B	C	N	B	C	N	B
$n_A=50$, $n_B=50$	0.040	0.040	0.013	0.041	0.079	0.028	0.036	0.272	0.136

(1) For $k = 2$, $Q = \{\mathbf{X}: X_1 > X_2\}$. For $k = 3$, $Q = \{\mathbf{X}: X_1 > \max(X_2, X_3)\}$.

(2) When $k = 3$, $\pi_2 = \pi_3 = 0.10, 0.50, 0.70$.

(3) UMPCU test.

(4) Naïve test.

(5) Bonferroni test.