Published in final edited form as:

Stat Med. 2017 April 30; 36(9): 1383–1394. doi:10.1002/sim.7226.

Two-sample binary phase 2 trials with low type I error and low sample size

Samuel Litwin^{1,*†‡}, Stanley Basickes², and Eric A. Ross¹

¹Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111, USA

²Greenfield Manufacturing, 920 Levick Street Philadelphia, Pennsylvania, 19111, USA

Summary

We address design of two-stage clinical trials comparing experimental and control patients. Our end-point is success or failure, however measured, with null hypothesis that the chance of success in both arms is p_0 and alternative that it is p_0 among controls and $p_1 > p_0$ among experimental patients. Standard rules will have the null hypothesis rejected when the number of successes in the (E)xperimental arm, E , sufficiently exceeds C , that among (C)ontrols. Here, we combine one-sample rejection decision rules, $E \geq m$, with two-sample rules of the form $E - C > r$ to achieve two-sample tests with low sample number and low type I error. We find designs with sample numbers not far from the minimum possible using standard two-sample rules, but with type I error of 5% rather than 15% or 20% associated with them, and of equal power. This level of type I error is achieved locally, near the stated null, and increases to 15% or 20% when the null is significantly higher than specified. We increase the attractiveness of these designs to patients by using 2:1 randomization. Examples of the application of this new design covering both high and low success rates under the null hypothesis are provided.

Keywords

binary comparisons; phase II clinical trials; low sample number; low type I error; two-stage

2 Introduction

The greatest drawback to early stage randomized comparative trials is the large number of patients required. One-sample designs, such as [2], using historical controls do not account for increased success due to better standards of care than were in place during assembly of historical data. One-sample trials may not be a random sample of the population of interest, so are hard to generalize. They may use patient eligibility rules or outcome measures that differ from historical ones. Moreover, such trials can be corrupted by investigator bias. Still, one-sample tests are popular because they can be small. One method of reducing the sample size of comparative trials is to increase the allowable type I error to 20%. This approach is

*Correspondence to: S. Litwin, Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center.

†samuel.litwin@fccc.edu

‡Research was supported in part by NIH contract/grant number P30 CA006927

taken in Jung [1] who achieves minimal sample numbers for binary outcome, two-stage, two-arm tests having 15% or 20% type I error and 80% or 85% power.

Our standard of comparison is Jung's, [1], two-stage, two-arm minimax design. We succeed in reducing type I error from, e.g. 20% to about 5%. This reduction comes at the cost of slightly increased sample number and only local, instead of global control of type I error. In Jung's designs, type I error is close to the stated value, 15% or 20%, for the entire range of control success rate. We control type I error near the stated null hypothesis and allow it to increase gradually as the control success rate increases beyond this value. We also obtain designs that reduce the overall sample number while reducing type I error at the stated null to about 5%. These smaller sample designs require the user to have precise knowledge of the control success rate.

We bring this about by removing some points from the critical region that would, in a standard two-sample context, lead to rejection of the null. These omitted points reflect lower than expected control (null) performance, and experimental performance below its alternative expectation.

For simplicity, we first describe our approach in a single stage setting. Let E and C stand for Experimental and Control successes. We use the usual two-sample decision rule: $E - C \geq r$ but also require that $E \geq m$. If the control success rate is lower than anticipated, but the experimental rate approximates the anticipated control rate, then the chance the $E - C \geq r$ leads to high type I error. However, for sufficiently large m it is unlikely that $E \geq m$, so this source of type I error can be made very small. The additional rule is just the usual one-sample test rule and amounts to simple sculpting of the critical region.

In section 4, below, we extend the approach to a two-stage trial with 2:1 randomization and early termination for lack of efficacy. Unlike Jung's approach, we have not attempted to minimize overall sample size using these modified rules. Rather we control the chances of early study termination under both null and alternative. We require that the chance of early termination under the null be at least 50% and, under the alternative at most 5%. These determine the sample size for our initial cohort. This simplifies the search for the overall sample number. Subject to these limits we provide extensive tables supporting information that correspond to those in [1]. supporting information also includes software that allows users to change the early termination parameters as well as all other design conditions.

Finally, we respond to the medical communities desire to make trials more attractive to patients by using 2:1 randomization. This does not seem to significantly increase the required sample numbers. Several new trial designs are discussed and compared to those tabulated in [1]. Two of the examples in [1] are compared as well. In addition to direct comparison of designs in our tables to those in [1] we discuss off-table improvements that can be obtained by varying some of the several design parameters.

3 Example of critical region sculpting in 1:1 one-stage setting

Figure 1 illustrates simple critical region sculpting to improve type I error. Consider 1:1 randomization in a single-stage test of the null hypothesis that $p_0 = p_1 = 0.25$ versus the

alternative that $p_0 = 0.25$ and $p_1 = 0.4$. The usual comparison of proportions will reject the null if $E - C \geq 5$ with $n = 58$ observations per arm. The test will have 80% power and 16.69% type I error. Small open circles at the cross hairs in the figure show the means of E and C under null and alternative. The blue line in the figure, $E = C + 5$, is the boundary of the critical region in this case. The broken red line, $E = \max(20, C + 3)$ is the boundary of a sculpted critical region. By eliminating the green points, $E < 20$, from the original critical region, type I error is reduced by 11.75% and power is reduced by 5.77%. Expanding this critical region to include the black points where, $E \geq \max(20, C + 3)$ augments power by 5.75%. The resulting power is still 80% with type I error reduced to under 6% with no change in sample size. Lost, however, is the global robustness of the unsculpted test. If the control success rate is increased by 7.5% from 25% to 32.5%, (to half-way between the hypothesized p_0 and p_1), and the experimental success rate increased the same amount, from 40% to 47.5%, type I error is increased to 25.4% with power increased to 87.9%. The UNSculpted test would garner only a 2% increase in type I error and 1% loss in power from these same shifts in proportions.

4 Early Stopping

Early stopping is especially important in phase II trials where resources must not be wasted on ineffective therapy. Below, we require the chance of early stopping under the null to be at least 50%. Similarly, we require the chance of early stopping in error to be at most 5%. However, these limits are easily adjusted to suit the user.

Let C_1 and E_1 be the numbers of successes among n_1 control and $2n_1$ experimental patients, respectively, in the first stage. Using this 2:1 randomization the study will be terminated after the first stage if either $E_1 - 2C_1 \leq a_1$ or $E_1 < m_1$. We find the smallest n_1 for which there are numbers a_1 and m_1 such that:

$$p(\text{early termination} | H_0) \equiv es_0 = p(E_1 - 2C_1 \leq a_1 \cup E_1 < m_1 | H_0) \geq 0.5$$

and

$$p(\text{early termination} | H_1) \equiv es_1 = p(E_1 - 2C_1 \leq a_1 \cup E_1 < m_1 | H_1) \leq 0.05$$

These requirements simplify the search for optimal sample numbers by making computation of the first cohort size and decision boundaries, a_1 and m_1 , independent of the second cohort. It comes at a small cost to an experiments overall sample number but provides control of early stopping probabilities.

Let x be the success rate of the control arm and y be that of the experimental arm. Let

$$f(x, y) = \sum_{B_1 \cap L_1} b(c_1, n_1, x) b(e_1, 2n_1, y)$$

where e_1 and c_1 are experimental and control first cohort successes and

$b(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$ is the binomial distribution. L_1 identifies the ranges of e_1 and c_1 , and B_1 the conditions for early stopping:

$$L_1 = \{(e_1, c_1) : (0 \leq c_1 \leq n_1) \cap (0 \leq e_1 \leq 2n_1)\}$$

$$B_1 = \{(e_1, c_1) : (e_1 - 2c_1 \leq a_1) \cup (e_1 < m_1)\}$$

Then, if the specified null success rate is p_0 and the alternative rate is p_1 the above relations become: $es_0 = f(p_0, p_0) = 0.5$ and $es_1 = f(p_0, p_1) = 0.05$. For given values of p_0 and p_1 we find the smallest n_1 for which there is at least one pair (a_1, m_1) , satisfying both inequalities. n_1 is incremented by one until such a value is reached. If there is more than one pair (a_1, m_1) the pair that maximizes the chance of early stopping under the null is selected. This specifies the first stage of the design.

5 Second Cohort

Continuation beyond the initial patient cohort is defined by the event

$B'_1 = E_1 - 2C_1 > a_1 \cap E_1 \geq m_1$. Let E_2 and C_2 be the numbers of successes among experimental and control subjects respectively in the second cohort. Let n_2 and $2n_2$ be the total numbers of controls and experimentals in the combined cohorts so $n_2 - n_1$ and $2(n_2 - n_1)$ are the numbers of second cohort controls and experimentals respectively. Armed with n_1 , a_1 and m_1 , we computed:

$$p(B'_1 \cap E_1 + E_2 - 2(C_1 + C_2) \geq r_2 \cap E_1 + E_2 \geq m_2)$$

under both null and alternative hypotheses to obtain type I error and power respectively for $r_2 \in [-2n_2, 2n_2]$ and $m_2 \in [0, 2n_2]$. Let

$$B_2(e_1, c_1, e_2, c_2) = \{(e_1, c_1, e_2, c_2) : (e_1 + e_2 - 2(c_1 + c_2) \geq r_2) \cap (e_1 + e_2 \geq m_2)\}$$

where c_2 and e_2 are second cohort control and experimental successes and

$$L_2 = \{(e_2, c_2) : (0 \leq c_2 \leq n_2 - n_1) \cap (0 \leq e_2 \leq 2(n_2 - n_1))\}$$

Let $L_3 = L_1 \cap L_2$, and let

$$g(x, y) = \sum_{B'_1 \cap B_2 \cap L_3} b(c_1, n_1, x) b(e_1, 2n_1, y) b(c_2, n_2 - n_1, x) b(e_2, 2(n_2 - n_1), y)$$

Then, if β is the type II error, the conditions on power, $1 - \beta$, and type I error, α , are: $g(p_0, p_1) \geq 1 - \beta$ and $g(p_0, p_0) \leq \alpha$, respectively.

With n_1 , a_1 and m_1 fixed, we found the smallest n_2 for which some combination of r_2 and m_2 met both power and type I error restrictions. n_2 , as with n_1 , is incremented by one until the smallest possible value is found. The search for n_2 is much slower than that for n_1 , but fast enough to be practical. Most p_0, p_1 combinations yielded multiple solutions in terms of parameters r_2 and m_2 . We then identified a combination with the smallest type I error that met the power condition. We found that this procedure yielded designs, that had very small sample numbers, but whose type I error is very sensitive to small changes in p_0 . This prompted a revision in our method of selecting the parameters.

6 Range of robustness

We found that a small increase in control success probability led to a disproportionate increase in type I error when we selected parameters as above. For example, in an extreme case we found $g(p_0, p_0) = 0.03$ but $g(p_0 + 0.05, p_0 + 0.05) > 0.25$. This effect was traced to parameter m_2 . As $m_2 > 0$ is increased in $g(p_0 + \delta, p_0 + \delta)$, m_2 becomes less relevant for $E_1 + E_2 = m_2$. To remedy this we identified parameters that fixed type I error to one of two design levels: 15% or 20%, not at p_0 , but at either $p_0 + 0.05$ or the midpoint, $(p_0 + p_1)/2$, and created tables for these choices. We created tables such that: $g(p_0 + \delta, p_0 + \delta) = \alpha$ for $\alpha \in \{0.15, 0.2\}$ and $\delta = 0.05$ or $\delta = (p_1 - p_0)/2$. This, of course, increases sample numbers accordingly and while not specifying type I error at the stated null, this value is tabulated and is consistently in the vicinity of 5%.

7 The Tables

Jung, [1], contains 4 tables: 80% or 85% power and 15% or 20% type I error. Tables for 80% power and 20% type I error are presented here. Similar tables for the other three cases can be found in supporting information. In each table entry p_0 and p_1 differ by either 10%, 15% or 20%. There are 33 (p_0, p_1) combinations corresponding to those in [1]. Type I error is not fixed at the specified null, but restricted to be at most 20% at $p_0 + 0.05$ in table 3, or at $(p_0 + p_1)/2$ in table 4. The resulting (smaller) type I error at the specified null is also tabulated. Sample numbers and decision parameters are listed for power of at least 80%. For each (p_0, p_1) combination the tables list the chance of early stopping under the null, es_0 , (always at least 50%); this chance under the alternative, es_1 , (always at most 5%); type I error at the stated null, α ; power at the stated alternative, pwr ; type I error at the stated null plus 5%, $g(p_0 + 0.05, p_0 + 0.05) = \alpha_1$; power at the stated null and alternative (both) plus 5%, $g(p_0 + 0.05, p_1 + 0.05) = pwr_1$; type I error at the midpoint, α_2 ; and power at the stated null and alternative (both) plus half the difference $p_1 - p_0$, pwr_2 . When type I error is computed at $p_0 + 0.05$ power is also computed at $p_1 + 0.05$, and similarly for the midpoint computation. Table 5 juxtaposes sample numbers for early and final decision in tables 4 and 5 to those in [1].

Table notation

p_0 : Chance of success in controls and in experimentals under the null. p_1 : Chance of success in experimentals under the alternative.

N_1 : total sample number of first patient cohort consisting of $n_1 = N_1/3$ controls and $2n_1 = 2N_1/3$ experimental patients. In table 5 columns 8 and 11 refer to the 1:1 randomization in [1]. In these columns (only) $N_1/2$ is both the control and experimental first stage sample size. $N_2/2$ is both the control and experimental total. Note that lower case is used in the power computation formulas where control and experimental sample numbers are needed, but total cohort sizes, N_1 and N_2 (below) are used in the tables.

a_1 and m_1 : If $E_1 - 2C_1 \leq a_1$ or $E_1 < m_1$ the study is terminated for futility after the first cohort. Note that N_1 , a_1 and m_1 are the same in columns 3,4 and 5 of tables 3 and 4, and N_1 is the same in columns 3 and 13 of the *comparison* table 5. These parameters are determined by the specified chances of early stopping under the null and alternative. They are not affected by power or type I error requirements.

N_2 : final total sample number. $N_2 - N_1$ is the size of the second cohort. There are $(N_2 - N_1)/3$ second cohort controls and $2(N_2 - N_1)/3$ second cohort experimental patients. As before in relation to [1], columns 8 and 9 in table 5 $(N_2 - N_1)/2$ is both the control and experimental second cohort sample number.

r_2 and m_2 : the null is to be rejected if $E_1 + E_2 \geq m_2$ AND $E_1 + E_2 - 2(C_1 + C_2) \geq r_2$.

es_0 and es_1 : early stopping probabilities under the stated null or alternative respectively. $es_0 = f(p_0, p_0)$ and $es_1 = f(p_0, p_1)$.

Type I error is tabulated under three conditions:

1. when both null and alternative probabilities of success are given by the tabulated p_0 , (α in tables). $\alpha = g(p_0, p_0)$.
2. when both are given by $p_0 + 0.05$, (α_1 in tables) $\alpha_1 = g(p_0 + 0.05, p_0 + 0.05)$.
3. when both are given by $(p_0 + p_1)/2$, (α_2 in tables). $\alpha_2 = g((p_0 + p_1)/2, (p_0 + p_1)/2)$.

Power is also tabulated under three conditions:

1. when p_0 and p_1 are as tabulated, (pwr in tables) $pwr = g(p_0, p_1)$.
2. when both p_0 and p_1 are increased by 5%, (pwr_1 in tables) $pwr_1 = g(p_0 + 0.05, p_1 + 0.05)$.
3. when both p_0 and p_1 are increased by the value of $(p_1 - p_0)/2$, (pwr_2 in tables). $pwr_2 = g((p_0 + p_1)/2, (p_0 + p_1)/2 + (p_1 - p_0))$.

Power computations, pwr , pwr_1 and pwr_2 as well as es_0 are rounded to the floors of their values and tabulated as integer percentages. Type I errors, α_1 and α_2 are rounded to their ceilings to an integer percentage. α and es_1 are rounded up to the nearest 0.1% and also tabulated as percentages.

For some high values of p_0 and p_1 , $p_1 + (p_1 - p_0)/2 > 1$. In such cases the experimental arm chance of success is set to 1.0 for the indicated computation.

8 Comparison with Sin-Ho Jung's designs in [1]

Our designs have several advantages over those in [1]. They have: much smaller type I error at the stated null, usually a smaller first cohort size, controlled early stopping probabilities, sample numbers slightly larger or much smaller (depending on design parameters) and patients more positive view of 2:1 randomization.

The first advantage, substantial reduction of type I error in the immediate vicinity of the null hypothesis, deteriorates to near zero either at the null plus 5% or at the midpoint between the null and alternative. At the null, type I error is close to 5% (for most table entries) while the design is determined to meet the higher stated type I error, 20%, at $p_0 + 0.05$ or the midpoint. Type I error at the stated null, p_0 , is always smaller than that stated in each table header. Beyond the given offset (5% or midpoint of p_0 and p_1) the design in [1] will have fairly constant type I error while that in our design will increase. See examples in figures 2 and 3.

Power is given in the tables for the stated null versus the stated alternative. Power when both p_0 and p_1 are increased the same amount will usually be greater than that achieved by [1] at the same offset. This is a result of rather small values of r_2 .

The table designs all provide at least 50% chance of early stopping under the null and at most 5% under the alternative. These choices fix the values of n_1 , a_1 and m_1 . They minimize n_1 under these conditions and make the search for minimum n_2 much shorter.

When the user is confident that the null probability p_0 is no more than 5% greater than his estimate, the design, mostly, provides a smaller total sample number than that in [1]. The larger sample sizes required under the midpoint construction also provide very small type I error at the stated null. The minimax designs in [1] sometimes allow multiple solutions (n_1 , a_1) for the same minimum n_2 . When this occurred Jung chose the designs to minimize the expected sample number rather than the minimum type I error. We did not attempt to reconfigure the tables in [1] to make the other choice when it existed.

9 Examples

Lung Cancer

In short duration trials of progression-free-survival, PFS, it can be difficult or impossible to determine when a disease progressed. Data of this sort are, at best, interval censored. Patients are usually seen every month or six weeks when their progress is ascertained. In these cases, the endpoint is often PFS at a particular time such as 6 months. Standard survival analysis methods may not be appropriate here [3]. A phase II study of HDC (heat shock protein Hsp90 conjugated to SN-38) versus Topotecan in small cell lung cancer is being considered. It should test the null hypothesis that progression free survival (PFS) will be 50%, $p_0 = 0.5$, at six months follow-up in both arms. The hoped for alternative is that PFS

in the HDC arm will be at least 70%, $p_1 = 0.7$, at six months. Four designs are compared in figure 2.

Design from [1]—Table 5 indicates that $N_1 = 34$, $N_2 = 70$ (from table III in [1]) for 80% power and 20% type I error at the null. This design has $p(ES|H_0) = 0.3$ and $p(ES|H_1) = 0.041$, (our computation). This 1:1 randomized design in [1] will be compared to two 2:1 designs in our tables and an off-table design.

Design with type I error determined at $p_0 + 0.05$ —Table 3, has type I error of 20% at $(p_0 + 0.05) = 0.55$. $N_1 = 30$ with early stopping if $E_1 - 2C_1 \leq -7$ or $E_1 < 11$ yielding $p(ES|H_0) = 0.59$ and $p(ES|H_1) = 0.049$. With fewer first cohort patients this design has almost double the chance early stopping under the null over that of the above design in [1]. With $N_2 = 51$ the null is rejected if $E_1 + E_2 - 2(C_1 + C_2) \leq -1$ and $E_1 + E_2 \leq 21$. The overall sample number is significantly smaller than that in [1]. Study power is 81% at the stated p_0 , p_1 combination and type I error 9.4% at the null is significantly lower than that in [1]. Type I error increases to 20% if $p_1 = p_0 = 0.55$ and to 32% if $p_1 = p_0 = 0.60$. This and other table 3 designs, have only a narrow range of robustness. However, if the user is confident that $p_0 = 0.5$ then this design has great advantage over [1]. Power increases to 88% if $p_0 = 0.55$ and $p_1 = 0.75$ and to 90% if $p_0 = 0.6$ and $p_1 = 0.8$. See figure 2A.

Design with type I error determined at midpoint—Table 4 has type I error of 20% at $(p_0 + p_1)/2 = 0.6$. N_1 , a_1 , m_1 , $p(ES|H_0) = es_0$, and $p(ES|H_1) = es_1$ are the same as table 3 and thus confer the same early stopping advantage as the table 3 design. With $N_2 = 78$ the null is rejected if $E_1 + E_2 - 2(C_1 + C_2) \leq -5$ and $E_1 + E_2 \leq 30$. This 8 patient increase over that in [1] is all in the second cohort, so is not realized in case of early stopping. Study power is 80% at the stated p_0 and p_1 combination and type I error 7.7% at the null. Type I error increases to 15% if $p_1 = p_0 = 0.55$ and to 20% if $p_1 = p_0 = 0.60$. The extension of robustness range along with low type I error at the null makes this design competitive with that of [1]. Power increases to 83% if $p_0 = 0.55$ and $p_1 = 0.75$ and to 84% if $p_0 = 0.6$ and $p_1 = 0.8$.

Off-table midpoint design—Individual results may be adjusted away from the tables to improve performance. If we tighten the requirement that $p(ES|H_0) > 0.5$ to $p(ES|H_0) > 0.6$ the design, corresponding to table 4, gives $N_1 = 36$, $N_2 = 75$, $p(ES|H_1) = 0.047$, and importantly, type I error is reduced to 4.2% at the null. It increases to 12% at $p_0 = p_1 = 0.55$ and to 20% at $p_0 = p_1 = 0.6$. Power increases to 85% if $p_0 = 0.55$ and $p_1 = 0.75$ and to 87% if $p_0 = 0.6$ and $p_1 = 0.8$. This design narrows the increase in N_2 over that in [1] to 5 patients and reduces the type I error at the null to under 5%.

This design commits at most 25 patients to the control treatment and 50 to the experimental one. The design in [1] commits 35 patients to each treatment. The modified design above has an initial cohort of 12 controls and 24 experimental patients has over 60% chance of early stopping under the null, over twice that in [1]. Within their ranges of robustness the table 3 and 4 designs and the modified one above have type I error at most equal to that of [1] and much less than that of [1] over most of the range of interest. See figure 2.

Prostate Cancer

A phase II study is proposed to compare Bicalutimide to Bicalutimide plus Metformin in prostate cancer patients. The binary end point of the study is reduction of PSA to undetectable level or not after 32 weeks of treatment. The null hypothesis is that Bicalutimide alone or in combination with Metformin will produce undetectable PSA in 20% of patients. The alternative of interest is that the combination will reduce PSA to undetectable levels in at least 40% of patients. Three designs are described and summarized below.

Design from [1]—Table 5 indicates that $N_1 = 34$, $N_2 = 54$ (from table III in [1]) for 80% power and 20% type I error at the null. This design has $p(ES|H_0) = 0.41$ and $p(ES|H_1) = 0.076$, (early stopping probability is our computation, not tabulated in [1]). This 1:1 randomized design in [1] will be compared to a 2:1 design in our tables and an off-table design. Off-table design was obtained by slightly relaxing the design parameters.

Design with type I error determined at midpoint—Table 4 has maximum type I error of 20% at $(p_0 + p_1)/2 = 0.3$. N_1 , a_1 , m_1 , $p(ES|H_0) = es_0$, and $p(ES|H_1) = es_1$ are the same as table 3 and thus confer the same early stopping advantage as the table 3 design. With $N_2 = 69$ the null is rejected if $E_1 + E_2 - 2(C_1 + C_2) \leq 5$ and $E_1 + E_2 \leq 12$. This 15 patient increase is all in the second cohort, so is not realized in case of early stopping. Study power is 80% at the stated p_0 and p_1 combination and type I error 7.7% at the null. Type I error increases to 15% if $p_1 = p_0 = 0.25$ and to 19% if $p_1 = p_0 = 0.30$. The extension of robustness range along with low type I error at the null makes this design competitive with that of [1]. Power is maintained at 80% if $p_0 = 0.25$ and $p_1 = 0.45$ and reduced to 79% if $p_0 = 0.3$ and $p_1 = 0.5$.

Off-table design with same N_2 as [1]— If the requirement for type I error at $p_0 + 0.05$ is set at 14% and the power requirement lowered to 79% a design with $N_1 = 27$ and $N_2 = 54$ is produced, thus matching the N_2 with that of [1] and N_1 7 smaller. Figure 3 shows this off-table design in comparison with [1] and the above design with $N_2 = 69$. While the $N_2 = 69$ midpoint design is uniformly better than [1], the much smaller off-table design is uniformly more powerful, and has the smallest type I error close to the null. The off-table design has smaller type I error than [1] for $0.2 \leq p_C \leq 0.2858$, a considerable fraction of the range of interest. Its type I error is 3.7% at the null. Its power rises to nearly 90% by the point where its type I error exceeds that of [1]. Its power is 0.7968 at the null, but rises quickly, dominating the other designs.

CALGB 50502—(taken from [1]) In this (example 2 in [1]) Jung extends his method to allow 2 : 1 randomization. It tests a null hypothesis that $p_C = p_E = p_0 = 0.7$ versus an alternative that $p_C = p_0 = 0.7$ and $p_E = p_1 = 0.85$ with 80% power and 15% type I error. The design requires $N_1 = 108$ patients in the initial cohort and $N_2 = 144$ total. Our design corresponding to these parameters requires $N_1 = 33$ patients in the initial cohort and $N_2 = 120$ total. It has type I error of 2.2% at the null, 10% at $p_0 + 0.05$ and 15% at $p_0 + 0.075$.

The example of CALGB 50502 using 1:1 randomization (example 1 in [1]) requires $N_1 = 62$ initial cohort patients and $N_2 = 126$ total with the parameters the same as the previous paragraph. Our design is a serious contender to both the 2:1 and 1:1 designs in [1].

Hypothetical comparison of two surgical meshes—A new mesh used for tissue containment in breast reconstruction is proposed to reduce the incidence of serious complications (e.g. removal of mesh and replacement due to infection). The current mesh results in 80% success. A study of 16 new mesh patients resulted in zero complications. Table 4 lists $N_1 = 18$ and $N_2 = 57$ to compare a control success rate of 80% to a experimental rate of 95%. Power is 81% and type I error is 3.4% when $p_E = p_C = 0.8$. Type I error is 13% when $p_E = p_C = 0.85$, and 20% when $p_E = p_C = 0.875$. The design in [1] lists $N_1 = 24$ and $N_2 = 66$ with type I error of 20% and 80% power.

10 Discussion

A two-sample, early stopping design using 2:1 randomization with simple critical region sculpting is shown to have low type I error and require a much smaller sample than the usual two-sample test. It provides local control of type I error near the null hypothesis. This error increases gradually as the true control rate increases beyond its specified value. The design benefits early stage clinical trials with access to modest patient numbers, limited budgets or require rapid evaluation.

In the CALGB 50502 example 2, [1], Jung obtains strict control of power and type I error over $0 < p_0 < 1$ using a worst case procedure, namely, assuming the null success rate is one-half. This conservative rule may not be needed if the success probability range of interest is bounded away from one-half. Our design takes this range directly into account.

The last few entries in table 5 shows smaller values of N_2 for our design than the corresponding ones in [1]. The CALGB 50502 hypothesis, $((p_0, p_1) = (0.7, 0.85))$ just borders this range. For high values of p_0 and p_1 , especially of p_1 , our design benefits from the smaller variances acquired by using the 5% or midpoint offsets.

The one-sample rejection rule, $E_1 + E_2 \geq m_2$, added to the two-sample test, is a simple form of critical region sculpting. Increasing the size of m_2 rapidly decreases type I error and allows r_2 in the difference of proportions rule, $E_1 + E_2 - 2(C_1 + C_2) \geq r_2$, to be smaller. This increases the power at the expense of type I error for p_0 much greater than its stated null value. We trade-off the uniform robustness in [1] for greatly reduced local type I error and comparable sample numbers to those in [1]. Our total sample numbers are usually greater than those in [1], but much smaller than a design of that type with type I error in the 5% range.

The chances of early stopping under null and alternative in our design, control N_1 , the size of the initial cohort. This fact can be used to balance the initial and final cohorts if desired. Off-table designs were found that reduce N_2 by manually searching early stopping requirements from 20% to 80% under the null and from 2% to 5% under the alternative. Some tabulated designs yielded smaller values of N_2 and some did not. We have not systematically checked, but small reductions in N_2 are possible in several of the tabulated designs.

Table 5 may be used directly or used as starting points for more refined searches. There are too many adjustable parameters to allow simple systematic search, but small adjustments

may yield worth-while reductions in sample numbers, type I error or other properties of interest. Programs to facilitate adjustments are included in supporting information (see Software, below).

A log-rank test was compared to the off-table lung cancer design in section 9 above. Using the same total sample number, 75, 2:1 randomization, 13 months accrual and 6.5 months followup, assuming exponential PFS the log-rank test attains only 70% power with 5% type I error. However, type I error with the log-rank test, is very nearly constant over a very wide range.

As our hypothetical surgical mesh example illustrates, when the control rate is very high, (e.g. at least 65%) our tabulated sample numbers are comparable or less than those in [1]. This stems from the fact that our computations use $p_0 + 0.05$, $p_0 + 0.075$ or $p_0 + 0.1$ with similar adjustments to p_1 . At the upper end of the scale these changes significantly reduce variances under both null and alternative. At the low end of the scale the opposite is true.

Reviewers drew attention to a design limitation illustrated by the following example: Suppose the hypothesized null and alternative success rates are 20% and 40%. Further, suppose that the control rate is in fact only 10%, but the experimental rate is 30%, still 20% higher than the controls but well below the stated alternative. The difference may be important, but it is much less likely that the experimental group will achieve both $E_1 - m_1$ and $E_1 + E_2 - m_2$. An important difference may be lost.

To be worthy of further consideration for subsequent phase III trials, clinicians often insist that the observed experimental success rate must achieve a minimal level in addition to exceeding the observed success rate among controls. A lower experimental success rate is considered a failure. In this case, over-specification of the control success rate will not be relevant. Here, $E_1 - m_1$ and $E_1 + E_2 - m_2$ do most of the work and type I error at the true (low) control success rate is very small. Such users realize the benefit of small sample size and very small type I error. Power loss resulting from overestimation of control success rate can be obtained via our supplied software supporting information.

The standard of care in clinical trials tends to improve over time. Success estimates based on historical controls lag behind this effect. Thus, estimates of control success rates tend to underestimate that of the current standard of care. This was a strong motivator of our design.

The design exchanges the extreme generality of the omnibus test in favor of acceptably small type I error with restricted range of application and small sample numbers. The method is being used in a wide range of clinical trials at Fox Chase Cancer Center in Philadelphia.

A future paper will also expand the current design to include early stopping for significant efficacy as well as 1:1 randomization.

11 Software

Programs are available to explore perturbations of designs found in the tables or get new designs based on off-table parameters. In particular, limits on type I error, power, early stopping limits and robustness limits (type I error at 5% or midpoint) are adjustable parameters.

Off-table designs are often of interest. Such designs can be guided by table entries close to the desired properties. Our main search engines, Fort21h and RpgmNum, can be used to modify design parameters: power, offset (5% or midpoint) for type I error, limit of type I error at this offset, maximum chance of early stopping under the alternative, minimum chance of early stopping under the null, and, of course, p_0 and p_1 . Further refinements can be studied using RpgmPtb, which allows design parameters: N_1 , N_2 , a_1 , m_1 , r_2 , and m_2 as well as p_0 and p_1 to be set at run time. RpgmRob provides additional computations to establish robustness properties. Programs are supplied with ‘readme’ files containing instructions and examples of their use in supporting information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Jung, Sin-Ho. Randomized phase II trials with a prospective control. *Statistics in Medicine*. 2008; 27:568–583. DOI: 10.1002/sim.2961 [PubMed: 17573688]
2. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
3. The EORTC Lung Cancer Group. Progression-free survival rate as primary end-point for phase II cancer clinical trials: Application to mesothelioma. *J Clin Oncol*. 2006; 24:3007–3012. DOI: 10.1200/JCO.2005.05.1359 [PubMed: 16809726]

Figure 1. Example of critical region sculpting

Figure 2. Four designs for lung cancer example

A: Experimental and control success rates are equal with $0.5 \leq p_C = p_E \leq 0.7$ for type I error computation. For power computation under constant difference p_C is as above and $p_E = p_C + 0.2$. Power is $g(p_C, p_C + 0.2)$ (constant difference), and type I error, $g(p_C, p_C)$ (equal probabilities), for designs from: table 3 with $N_2 = 51$ (blue curves), table 4 with $N_2 = 78$ (green curves) and off-table design with $N_2 = 75$ (red curves). Computations for design in [1] are black lines. **B:** Contours of constant power = $g(p_C, p_E)$, power $\in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.80, 0.85, 0.9\}$ for off-table design with $N_2 = 75$ Red line is $p_E = p_C$, blue line is $p_E = p_C + 0.2$. **C:** Contours of 80% power and 20% power (type I error). Black [1], blue table 3 with $N_2 = 51$ and red is the off-table design with $N_2 = 75$. This show that tables 3 and off-table 4 designs have better power and that the off-table design has better type I error up to the range mid-point.

Figure 3. Off-table design for the prostate example

A: Experimental and control success rates $p_0 = 0.2$ and $p_1 = 0.4$. Off-table design with $N_2 = 54$ (red), table 4 with $N_2 = 69$ (green). Computations for design in [1] are black curves. **B:** Contours of constant power for $N_2 = 54$ off-table design. **C:** Contours of 80% power and 20% type I error. Black: [1], green: table 4 with $N_2 = 69$, red: the off-table design. Compared to [1] the off-table design has smaller type I error for $p_C < 0.2858$ and higher power throughout the plotted range. The table 4 design has lower type I error and higher power throughout this range.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Summary of 4 designs for lung cancer study

design	N_1	N_2	es_0	es_1	α	α_1	α_2	pwr	pwr_1	pwr_2
[1]	34	70	30	4.1	20	20	20	80	80	81
table 3	30	51	59	4.9	9.4	19	32	81	90	92
table 4	30	78	59	4.5	7.7	15	20	80	83	84
off-table	36	75	60	4.7	4.2	12	20	80	85	87

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Summary of 3 designs for prostate cancer study

design	N_1	N_2	es_0	es_1	α	α_1	α_2	pwr	pwr_1	pwr_2
[1]	34	54	41	7.6	20	21	23	80	78	78
table 4	27	69	52	4.5	7.7	15	19	80	80	79
off-table	27	54	52	4.5	3.7	13	26	79	87	89

Table 3
Parameters, sampling numbers and decision boundaries for early stopping studies

p_0	p_1	N_1	a_1	m_1	N_2	r_2	m_2	es_0	es_1	α	pwr	α_1	pwr_1	α_2	pwr_2
5	15	45	-5	2	114	4	7	55	5.0	4.3	80	20	78	20	78
5	20	33	-5	2	51	1	5	69	4.9	2.4	81	20	87	30	87
5	25	18	-2	1	30	2	3	55	4.5	4.9	80	17	80	26	78
10	25	36	-4	3	69	1	9	57	4.7	3.3	81	19	88	29	88
10	30	21	-5	2	39	2	5	58	4.9	7.3	82	19	83	29	82
15	30	36	-5	4	81	2	13	51	4.9	4.3	80	19	86	26	86
15	35	24	-5	3	39	1	7	56	4.9	7.2	80	20	85	32	86
20	35	42	-5	6	87	1	17	51	5.0	5.1	80	20	88	29	89
20	40	27	-4	4	45	2	9	52	4.5	9.3	80	20	83	30	83
25	40	45	-7	8	96	1	22	51	4.8	5.1	80	20	89	29	90
25	45	27	-5	5	51	2	12	53	4.8	9.0	81	20	84	30	85
30	45	48	-6	10	108	2	28	51	4.9	5.5	81	20	88	28	89
30	50	27	-7	6	51	1	14	53	5.0	8.8	81	20	87	32	89
35	50	45	-10	11	105	1	31	50	5.0	5.8	80	20	89	29	91
35	55	33	-4	8	51	1	16	51	4.4	8.4	80	19	86	31	88
40	55	51	-6	14	114	3	37	51	4.9	6.1	80	19	86	26	87
40	60	36	-4	10	54	1	19	52	4.2	7.3	80	18	87	30	89
45	60	45	-9	14	111	2	40	50	5.0	6.0	80	20	88	28	90
45	65	33	-7	11	54	2	20	60	4.9	9.9	80	20	85	30	87
50	65	54	-8	19	108	1	43	57	4.9	5.4	80	19	90	28	92
50	70	30	-7	11	51	1	21	59	4.9	9.4	81	20	88	32	90
55	70	42	-9	16	102	1	44	51	5.0	5.7	80	19	90	28	92
55	75	30	-5	12	48	2	21	59	4.6	10.6	80	20	85	29	87
60	75	48	-6	20	99	2	46	55	4.5	5.6	80	18	89	26	91
60	80	21	-5	9	48	1	23	51	4.7	9.2	83	20	90	31	93
65	80	42	-5	19	84	1	42	55	4.9	6.3	80	19	91	27	92
65	85	21	-5	10	39	1	20	57	4.7	10.6	82	20	89	31	92

p_0	p_1	N_1	a_1	m_1	N_2	r_2	m_2	es_0	es_1	a	pwr	α_1	pwr_1	α_2	pwr_2
70	85	33	-5	16	72	2	38	51	4.1	8.2	80	20	88	27	90
70	90	18	-3	9	36	1	20	53	3.6	8.8	82	18	91	28	93
75	90	27	-3	14	60	1	34	51	4.7	7.7	82	20	92	28	95
75	95	15	-2	8	27	1	16	54	3.0	10.2	83	18	86	28	76
80	95	18	-2	10	48	1	29	51	4.6	7.6	83	19	92	27	88
85	95	36	-3	21	108	1	67	52	4.3	2.9	80	20	97	20	97

Table 4
Parameters, sampling numbers and decision boundaries for early stopping studies

α_2	0.20, power		0.80 $p(ES H_0)$		0.5, $p(ES H_1)$		0.05								
p_0	p_1	N_1	a_1	m_1	N_2	r_2	m_2	es_0	es_1	α	pwr	α_1	pwr_1	α_2	pwr_2
5	15	45	-5	2	114	4	7	55	5.0	4.3	80	20	78	20	78
5	20	33	-5	2	66	3	6	69	4.9	1.6	82	14	83	20	82
5	25	18	-2	1	36	3	0	55	4.5	5.9	80	15	77	20	74
10	25	36	-4	3	84	4	10	57	4.7	3.3	81	15	82	20	81
10	30	21	-5	2	54	3	8	58	4.9	1.8	80	10	84	20	84
15	30	36	-5	4	99	5	14	51	4.9	5.6	80	16	80	20	80
15	35	24	-5	3	60	4	10	56	4.9	4.2	80	13	81	20	81
20	35	42	-5	6	117	5	22	51	5.0	3.5	80	14	83	20	83
20	40	27	-4	4	69	5	12	52	4.5	7.7	80	15	80	19	79
25	40	45	-7	8	126	6	26	51	4.8	6.7	80	17	82	20	82
25	45	27	-5	5	75	5	17	53	4.8	5.5	80	13	81	19	81
30	45	48	-6	10	135	5	35	51	4.9	3.3	80	14	85	20	85
30	50	27	-7	6	78	5	20	53	5.0	6.2	80	14	82	20	82
35	50	45	-10	11	141	6	40	50	5.0	4.9	80	16	84	20	84
35	55	33	-4	8	78	5	23	51	4.4	6.3	80	14	82	20	82
40	55	51	-6	14	144	6	46	51	4.9	4.6	80	15	84	20	84
40	60	36	-4	10	81	5	27	52	4.2	5.5	81	14	83	20	83
45	60	45	-9	14	147	5	53	50	5.0	3.3	80	14	87	20	88
45	65	33	-7	11	81	4	31	60	4.9	3.3	80	11	86	20	88
50	65	54	-8	19	141	6	54	57	4.9	5.5	80	16	84	20	85
50	70	30	-7	11	78	5	30	59	4.9	7.7	80	15	83	20	84
55	70	42	-9	16	138	5	59	51	5.0	3.5	80	14	87	20	88
55	75	30	-5	12	72	4	32	59	4.6	4.7	80	12	85	20	87
60	75	48	-6	20	123	5	56	55	4.5	5.0	80	15	85	20	86
60	80	21	-5	9	72	1	36	51	4.7	2.0	81	8	93	20	96
65	80	42	-5	19	111	4	55	55	4.9	4.1	80	14	87	20	89
65	85	21	-5	10	60	2	32	57	4.7	2.6	80	9	91	19	94

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

ϵ_{AMD}	σ_2	ϵ_{AMD}	ϵ_2	ϵ_{AMD}	σ	ϵ_{SD}	ϵ_{SD}	m_2	ϵ_2	ϵ_N	ϵ_m	ϵ_1	ϵ_N	ϵ_d	ϵ_0
92	20	90	13	80	3.3	4.1	51	53	3	99	91	5	33	58	70
97	20	90	11	81	4.0	3.6	53	27	2	48	9	3	18	90	70
16	20	88	14	80	4.9	4.7	51	44	3	78	14	3	27	90	75
85	19	93	9	82	3.4	3.0	54	22	1	36	8	2	15	95	75
92	20	95	13	81	3.4	4.6	51	35	1	57	10	2	18	95	80
97	20	97	20	80	2.9	4.3	52	67	1	108	21	3	36	95	85

Table 5
Comparison of 2:1 to 1:1 in [1] randomization for 80% power and 20% maximum type I error

		2:1 table 3				1:1 table III in [1]				2:1 table 4			
		α_1		α_2		α		α		α_2		α	
p_0	p_1	N_1	N_2	α	α_1	α_2	N_1	N_2	α	α_1	α_2	N_1	N_2
5	15	45	114	4.3	20	20	58 ^a	90 ^a	20	26	26	45	114
5	20	33	51	2.4	20	30	24	44	14	22	25	33	66
5	25	18	30	4.9	17	26	16	32	10	18	22	18	36
10	25	36	69	3.3	19	29	52	68	16	20	21	36	84
10	30	21	39	7.3	19	29	20	34	19	24	26	21	54
15	30	36	81	4.3	19	26	54	82	20	23	23	36	99
15	35	24	39	7.2	20	32	36	50	16	19	21	24	60
20	35	42	87	5.1	20	29	58	96	19	21	21	42	117
20	40	27	45	9.3	20	30	34	54	20	21	23	27	69
25	40	45	96	5.1	20	29	84	106	20	22	22	45	126
25	45	27	51	9.0	20	30	48	64	20	21	22	27	75
30	45	48	108	5.5	20	28	74	120	19	20	20	48	135
30	50	27	51	8.8	20	32	46	70	18	19	20	27	78
35	50	45	105	5.8	20	29	96	120	20	21	21	45	141
35	55	33	51	8.4	19	31	40	70	19	20	20	33	78
40	55	51	114	6.1	19	26	70	124	20	21	21	51	144
40	60	36	54	7.3	18	30	42	70	20	20	20	36	81
45	60	45	111	6.0	20	28	100	128	20	21	21	45	147
45	65	33	54	9.9	20	30	40	70	20	20	20	33	81
50	65	54	108	5.4	19	28	70	128	20	20	20	54	141
50	70	30	51	9.4	20	32	34	70	20	20	20	30	78
55	70	42	102	5.7	19	28	64	122	20	20	20	42	138
55	75	30	48	10.6	20	29	36	68	20	19	19	30	72
60	75	48	99	5.6	18	26	76	116	20	19	19	48	123
60	80	21	48	9.2	20	31	36	66	19	19	18	21	72
													8

2:1 table 3				1:1 table III in [1]				2:1 table 4								
α_1		0.20		α		0.20		α_2		0.20						
p_0	p_1	N_1	N_2	α	α_1	α_2	N_1	N_2	α	α_1	α_2					
65	80	42	84	6.3	19	27	62	114	19	18	17	42	111	4.1	14	20
65	85	21	39	10.6	20	31	42	64	20	20	19	21	60	2.6	9	19
70	85	33	72	8.2	20	27	60	98	20	19	19	33	99	3.3	13	20
70	90	18	36	8.8	18	28	32	52	20	19	18	18	48	4.0	11	20
75	90	27	60	7.7	20	28	56	86	19	17	16	27	78	4.9	14	20
75	95	15	27	10.2	18	28	18	46	19	17	15	15	36	3.4	9	19
80	95	18	48	7.6	19	27	24	66	20	18	17	18	57	3.4	13	20
85	95	36	108	2.9	20	20	78	126	19	15	15	36	108	2.9	20	20