TECHNICAL REPORT # 58
APRIL 2012

Design resampling for interim sample size recalculation

Sergey Tarima, Peng He, Tao Wang, Aniko Szabo
Division of Biostatistics, Medical College of Wisconsin

Abstract

Internal pilot designs allow re-estimation of the sample size at the interim analysis using available information on nuisance parameters. In general, this affects the Type I and II error rates. We propose a method based on resampling the whole design at the interim analysis, starting with sample size recalculation at the observed interim analysis values of nuisance parameters, and finishing with the decision to accept or reject the null hypothesis. This internal resampling is performed under both the null and under the alternative hypotheses allowing the estimation of the bias of the type I error and power. Finally, the bias corrected error rates are used in the original sample size calculation procedure to obtain an updated sample size. We explore the proposed resampling approach under a set of simulation scenarios and compare it with several others previously published internal pilot designs.

KEYWORDS: Internal Pilot; Sample Size; Power Calculation; Hypothesis Testing; Study Design.

1 Introduction

Ethical, financial, and recruitment constraints prevent researchers from enrolling arbitrarily many patients for a study to achieve statistically significant results. Pilot studies are used to provide information on parameters needed to determine an appropriate sample size for a larger confirmatory

study for which external funding is sought. The error variance and baseline rates are common examples of such parameters. In observational studies, there are often additional nuisance parameters involved in the sample size estimation, such as the proportion of patients belonging to each group, or the distribution and effect of other demographic variables which will be adjusted for in the analysis. These parameters can be estimated from the pilot data and incorporated into the sample size estimation. The effect size is also needed for sample size calculation, but it is usually defined as the smallest clinically relevant value.

Having spent substantial effort and resources on the pilot study, investigators often would wish to include the pilot data in the final data analysis. However the use of this pilot data in both the design of subsequent data collection (estimation of the sample size) as well as in the final analysis requires an adjustment in order to control the type I error rate and power appropriately. The methodology of internal pilot studies provides the tools for such adjustments. Multiple authors have discussed procedures for sample size re-estimation based on estimates of nuisance parameters obtained at an interim analysis. Proschan [7] and Friede and Kieser [2] review interim pilot designs. The majority of work in this area has been focused on two-arm prospective randomized studies with normally distributed or binomial outcomes. Sample size re-estimation for a general linear hypothesis tested through the general linear model was explored by Coffey and Muller [1]; an interesting extension toward linear mixed models was suggested by Glueck and Muller [4]. However neither of these papers accounts for randomness in the distribution of the covariate of interest. An interim pilot design for the analysis of covariance model (ANCOVA) was explored via simulations by Friede and Kieser [3], while Gurka et al [5] analyzed the effect of unknown group size proportions.

In this manuscript we propose a general methodology of constructing interim pilot designs for any situation for which a sample size calculation formula exists. Such a formula may include nuisance parameters that are not available at the beginning of the study, but could be estimated if data were available. The method is based on resampling the entire design, starting with sample size recalculation at the interim analysis using the values of nuisance parameters observed in the internal pilot data (see Section 2). This resampling is performed under both the null and alternative hypotheses allowing the estimation of the bias of the type I error and power. We correct for this bias at the interim sample size recalculation to obtain a final

sample size. In Section 3 we explore the proposed approach under a set of simulation scenarios and compare its performance with several others previously published internal pilot designs. An illustrative example is presented in Section 4. The article is concluded with a short summary in Section 5.

2 Methodology

Study design

We explore an internal pilot design when $n_1 (< n_{max})$ subjects are enrolled in the study as an internal pilot and the second stage sample size is recalculated at the interim analysis to secure the desired type I error, α, and power, $1-\beta$. We assume that there always exists an upper bound for the total sample size, n_{max}, chosen for example, from budgetary or time constraints.

Consider the problem of independent sampling, where each observation Y_i is generated from a known p.d.f. (or p.m.f.) $f_Y(y|\theta, \eta)$ with unknown parameters θ and η. We plan to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta \neq \theta_0$ with power $100(1-\beta)\%$ achieved at $\theta = \theta_1$. The parameters η are not in the research focus and will be treated as nuisance.

Instead of focusing on the behavior of test statistics we direct our attention to the properties of decision functions, $\delta(\mathbf{D})$, where $\delta(\cdot)$ is a binary function (1 to reject H_0, 0 otherwise) associated with a study design $\mathbf{D}$. In this manuscript, we define a study design as (1) a set of data collection rules including the sample size calculation/recalculation procedure, (2) a definition of a test statistic, and (3) a definition of the decision rule itself. These definitions should cover all possible situations including rules for "exceptions", such as having no observations in a certain category, or zero variance, etc. It is convenient to consider parameterized study designs, $\mathbf{D}(\alpha, \beta, \theta_0, \theta_1, \text{other parameters})$, where in addition to the previously described α, β, θ_0, and θ_1 other parameters may be present, for example n_1 and n_{max}.

Fixed sample size and internal pilot designs

To clarify the introduced concept of study design and notations, we show how several common designs fit into the described framework.

Designs with fixed sample size calculation ($\mathcal{D}_1$): These designs are associated with a rule for calculating a fixed sample size (v) at the start of the study based on α, β, θ_0 and θ_1, and an assumed value of η. The functional form of a test statistic, $T_v = T_v(Y_1, ..., Y_v)$ with a corresponding critical value calculation, is also defined in this design.

Let $\mathbf{D}_{1t}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}\right)$ $(\in \mathcal{D}_1)$ denote the study design for the one sample t-test. The sample size is calculated for given α and β under an assumed value of the standard deviation, $\eta^{(0)}$. Thus, $\eta^{(0)}$ is the only nuisance parameter of $\mathbf{D}_{1t}$. The statistical properties of $\mathbf{D}_{1t}$ are described by the power function

$$P(\theta|\mathbf{D}_{1t}) = Pr\left(\left|T_{v(\alpha,\beta,\theta_0,\theta_1,\eta^{(0)})}\right| > k(v)|\theta, \mathbf{D}_{1t}\right), \tag{1}$$

where $v(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)})$ is the sample size formula, T_v is the t-statistic, $k(v) = t_{v-1,\alpha/2}$ is the critical value.

A two sample t-test with a fixed allocation ratio (r) always keeps $r100\%$ units allocated to group 1 and $(1-r)100\%$ to group 2. We denote this design as $\mathbf{D}_{2t}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}, r\right)$. The nuisance parameters $\eta^{(0)} = \left(\eta_1^{(0)}, \eta_2^{(0)}\right)$ denote guesses on the standard deviation (η_1) and the mean of group 1 (η_2). We do not include r among the nuisance parameters since we assumed that the allocation ratio is controlled by the investigator via, say, a randomized block design. If the data are normally distributed, the distribution of the two sample t-test does not depend on η_2 under either the null or the alternative hypotheses. Only $\eta_1^{(0)}$ affects the sample size calculations.

If the allocation ratio is not fixed, then r describes the long term allocation ratio (i.e., the probability that an observational unit belongs to group 1). Then, the distribution of the test statistic is a binomial mixture of two t-statistics with varying degrees of freedom. This design is denoted as $\mathbf{D}_{2tr}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}\right)$, where $\eta^{(0)} = \left(\eta_1^{(0)}, \eta_2^{(0)}, \eta_3^{(0)}\right)$. Here we added $\eta_3^{(0)}$, a guessed value of the true allocation ratio η_3. The use of η_3 instead of r differentiates parameters responsible for random and fixed allocation ratios.

Internal pilot designs ($\mathcal{D}_2$): Fixed sample designs can be augmented to become internal pilot designs by adding rules/parameters describing the interim sample size re-estimation procedure. A naive approach to interim sample size recalculation uses $\hat{\eta}$, the interim analysis estimate of the nuisance parameter without further adjustments. Then, the naive internal pilot design

for the one sample t-test,

$$\mathbf{D}_{1t,IPN}\left(\alpha,\beta,\theta_0,\theta_1,n_1,n_{max}\right)\left(\in\mathcal{D}_2\right),$$

is an alternative to $\mathbf{D}_{1t}$, which does not use $\eta^{(0)}$ but depends on n_1 and n_{max}. Its power function is

$$P\left(\theta|\mathbf{D}_{1t,IPN}\right)=Pr\left(\left|T_{v(\alpha,\beta,\theta_0,\theta_1,\hat{\eta}_\theta)}\right|>k(v)|\theta,\mathbf{D}_{1t,IPN}\right), \tag{2}$$

where $\hat{\eta}_\theta$ depends on η, n_1, n_{max} and possibly θ. In this manuscript we assume that $\hat{\eta}_\theta$ is independent of θ, that is $\hat{\eta}=\hat{\eta}_\theta$.

A naive internal pilot-based sample size recalculation for a two sample t-test will be denoted by $\mathbf{D}_{2t,IPN}$. This design was first analyzed by Wittes and Brittain [9]. We also consider the internal pilot design $\mathbf{D}_{2t,IPS}$ suggested by Stein [8], which slightly modifies the functional form of the two-sample t-statistic, whereas $\mathbf{D}_{2t,IPN}$ uses the classical two sample t-statistic for T_v.

Internal sample size recalculation makes the final sample size a random variable, which makes the distribution of the test statistic T_v and therefore the critical value of the test difficult to calculate. Exact control of the type I error is achieved by $\mathbf{D}_{2t,IPS}$, but this is rather an exception than a rule for internal pilot designs. In general, the true type I error rate is rarely controlled,

$$E\delta\left(\mathbf{D}_{2t,IPN}\left(\alpha,\beta,\theta_0,\theta_1,n_1,n_{max}\right)|H_0\right)=a\left(\alpha,\beta|\mathbf{D}_{2t,IPN}\right)\neq\alpha.$$

The desired power is not controlled in either Stein's or the naive internal pilot designs,

$$E\delta\left(\mathbf{D}\left(\alpha,\beta,\theta_0,\theta_1,n_1,n_{max}\right)|H_1\right)=1-b\left(\alpha,\beta|\mathbf{D}\right)\neq 1-\beta.$$

Sample size recalculation via resampling

We propose a new approach to sample size re-estimation after the internal pilot that maintains both the type I and type II error rates. This approach is applicable to any internal pilot design.

Key idea: For a design $\mathbf{D}\in\mathcal{D}_2$ we find α_{new} and β_{new} to control the desired type I error and power,

$$E\delta\left(\mathbf{D}\left(\alpha_{new},\beta_{new},\theta_0,\theta_1,n_1,n_{max}\right)|H_0\right)=\alpha$$

and

$$E\delta\left(\mathbf{D}\left(\alpha_{new},\beta_{new},\theta_0,\theta_1,n_1,n_{max}\right)|H_1\right)=1-\beta.$$

This definition leads to a fully defined internal pilot procedure $\mathbf{D}^a(\alpha,\beta,\theta_0,\theta_1,n_1,n_{max})$, since all the details about sample size re-estimation, final hypothesis testing, etc are already defined in $\mathbf{D}$.

Implementation: At the interim analysis we estimate $\hat{\eta}$ and perform the following resampling procedure with M iterations. For each $i=1,...,M$, we generate $\left(Y_1^{(i)},...,Y_{n_1}^{(i)}\right)$ from $f_Y\left(y|\theta_0,\hat{\eta}\right)$, estimate $v_i\in[n_1,n_{max}]$ based on these n_1 observations, generate additional (v_i-n_1) observations $\left(Y_{n_1+1}^{(i)},...,Y_{v_i}^{(i)}\right)$ from $f_Y\left(y|\theta_0,\hat{\eta}\right)$, and calculate $T_{v_i}^{(i)}$ on this i^{th} sample. We add the subscript i to highlight dependence on iteration. The estimated type I error rate is

$$\hat{a}\left(\alpha,\beta|\mathbf{D}\right)=\frac{1}{M}\sum_{i=1}^{M}I\left(T_{v_i}^{(i)}>k_i\right)\neq\alpha,$$

where k_i is the critical value for an originally assumed distribution of $T_{v_i}^{(i)}$. On the logit scale ($\text{logit}(x)=\ln\left(x/(1-x)\right)$) the bias-corrected α_{new} can be expressed as

$$\text{logit}(\alpha_{new})=\text{logit}(\alpha)-[\text{logit}(\hat{a})-\text{logit}(\alpha)]$$

or

$$\alpha_{new}=\frac{\alpha^2\left(1-\hat{a}\right)}{\left(1-\alpha\right)^2\hat{a}+\alpha^2\left(1-\hat{a}\right)}. \tag{3}$$

Then, we perform a similar resampling procedure to find β_{new}. For $i=1,...,M$, we generate $\left(Y_1^{(i)},...,Y_{n_1}^{(i)}\right)$ from $f_Y\left(y|\theta_1,\hat{\eta}\right)$, estimate $v_i\in[n_1,n_{max}]$ on these n_1 observations using α_{new} and β in the sample size formula, generate additional (v_i-n_1) observations $\left(Y_{n_1+1}^{(i)},...,Y_{v_i}^{(i)}\right)$ from $f_Y\left(y|\theta_1,\hat{\eta}\right)$, and calculate $T_{v_i}^{(i)}$ on this i^{th} sample. The estimated power

$$1-\hat{b}\left(\alpha_{new},\beta|\mathbf{D}\right)=\frac{1}{M}\sum_{i=1}^{M}I\left(T_{v_i}^{(i)}>k_i\right)\neq1-\beta$$

leads to the bias-corrected value

$$\beta_{new} = \frac{\beta^2 \left(1 - \hat{b}\right)}{(1-\beta)^2 \hat{b} + \beta^2 \left(1 - \hat{b}\right)}. \tag{4}$$

This internal resampling scheme assumes that the functional form of the data distribution, including the distribution of covariates, is known. For example, we can assume that the covariates follow a normal distribution. Alternatively, a nonparametric bootstrap resampling of the internal pilot data can be used. The illustrative example in Section 4 uses a combination of nonparametric (the distribution of covariates) and parametric (the distribution of outcome conditional on covariates) bootstraps for internal resampling.

3 Simulation Studies

Here we consider a few scenarios showing how the proposed general methodology works and compare it with other applicable approaches.

Internal pilot for a one sample t-test

Let $Y_1, ..., Y_{n_1}, ...$ be an i.i.d. sample from $N(\theta, \eta^2)$ and the one sample t-test is planned for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, with target power $100(1-\beta)\%$ achieved at $\theta = \theta_1$. The test statistic on v observations is $T_v = \sqrt{v}\bar{Y}_v/S_v$, where $\bar{Y}_v = v^{-1}\sum_{i=1}^{v} Y_i$ and $S_v^2 = (v-1)^{-1}\sum_{i=1}^{v}\left(Y_i - \bar{Y}_v\right)^2$. Under H_1, T_v has a noncentral t-distribution with $v-1$ degrees of freedom and noncentrality parameter $\omega_1 = \sqrt{v}(\theta_1 - \theta_0)/\eta$. The cumulative distribution function will be denoted as $P\left(T_v > x | v-1, \omega_1\right)$. Then, the sample size at a known η is a solution to

$$\begin{aligned} P\left(|T_v| > k | v-1, \omega_1\right) &= 1-\beta, \\ P\left(|T_v| > k | v-1, 0\right) &= \alpha, \end{aligned} \tag{5}$$

where the critical value k is also found from these equations.

Given an assumed value $\eta^{(0)}$, one can proceed with fixed sample size calculation for the design $\mathbf{D}_{1t}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}\right)$.

Design $\mathbf{D}_{1t,IPN}(\alpha, \beta, \theta_0, \theta_1, n_1, n_{max})$ does not formally depend on η and uses the internally estimated

$$\hat{\eta} = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_i - \bar{Y}\right)^2}$$

when solving Equations (5).

We define $\mathbf{D}^a_{1t,IPN}(\alpha, \beta, \theta_0, \theta_1, n_1, n_{max})$ as $\mathbf{D}_{1t,IPN}(\alpha, \beta, \theta_0, \theta_1, n_1, n_{max})$ augmented with internal recalculation of α and β using Equations (3) and (4). Then, v is found by resolving

$$\begin{aligned} P\left(|T_v| > k | v - 1, \hat{\omega}_1\right) &= 1 - \beta_{new}, \\ P\left(|T_v| > k | v - 1, 0\right) &= \alpha_{new}, \end{aligned} \tag{6}$$

where $\hat{\omega}_1 = \sqrt{v}(\theta_1 - \theta_0)/\hat{\eta}$.

Tables 1 and 2 summarize the results of this simulation study with internal pilot size of $n_1 = 5$ and 10. For the fixed sample design $\mathbf{D}_{1t}$ we consider the unrealistic case of $\eta = \eta^{(0)}$, that is perfect knowledge of the nuisance parameter. Thus for this example the goal is to be as close as possible to this ideal case. Both tables show the expected inflation of the type I error rate by the naive design $\mathbf{D}_{1t,IPN}$ accompanied by loss of power with increasing values of the standard deviation. The resampling-adjusted design $\mathbf{D}^a_{1t,IPN}$ eliminates most of the type I error inflation and power deflation, especially for $n_1 = 10$. The price for this correction is a higher expected sample size with more variability.

Internal pilot for a two sample t-test with a fixed group allocation

In this section we consider the randomized block design. The randomization is performed within blocks to ensure the desired allocation ratio r. The size of the block (b) determines the magnitude of the sample size increments. Then, the internal pilot of n_1 experimental (or observational) units consists of $m_1 = n_1/b$ blocks separating n_1 units to two treatment groups with n_{10} and n_{11} observations respectively, $n_1 = n_{10} + n_{11}$. Mathematically, we consider two data generating processes

$$Y_{10}, ..., Y_{n_{10}0}, ..., Y_{v_0 0}, ... \sim N(\eta_2, \eta_1^2)$$

Table 1: Monte-Carlo Type I error, Power, and Sample Sizes; $100,000$ simulations; one sample t-test designs, $n_1 = 10$, $n_{max} = 300$.

	$\mathbf{D}_{1t}$	$\mathbf{D}_{1t,IPN}$	$\mathbf{D}^{a}_{1t,IPN}$
η	Type I error		
1.6	0.0492	0.0643	0.0573
2	0.0500	0.0612	0.0513
3	0.0495	0.0553	0.0473
3.5	0.0494	0.0526	0.0473
	Power		
1.6	0.8177	0.8091	0.8367
2	0.8086	0.7841	0.8216
3	0.8040	0.7601	0.8001
3.5	0.8043	0.7517	0.7943
	EN(SD)		
1.6	23	22.73(9.33)	26.86(12.22)
2	34	33.89(14.80)	40.93(18.01)
3	73	73.17(33.29)	86.68(38.06)
3.5	99	98.53(45.05)	115.89(51.21)

Table 2: Monte-Carlo Type I error, Power, and Sample Sizes; $100,000$ simulations; one sample t-test designs, $n_1 = 5$, $n_{max} = 300$.

	$\mathbf{D}_{1t}$	$\mathbf{D}_{1t,IPN}$	$\mathbf{D}^a_{1t,IPN}$
η		Type I error	
0.6	0.0501	0.0523	0.0515
1	0.0515	0.0727	0.0682
2	0.0487	0.0685	0.0519
3	0.0503	0.0589	0.0448
3.5	0.0504	0.0574	0.0458
		Power	
0.6	0.8985	0.9387	0.9335
1	0.8030	0.8327	0.8596
2	0.8076	0.7319	0.7897
3	0.8033	0.7057	0.7663
3.5	0.8034	0.6953	0.7560
		EN(SD)	
0.6	6	6.00(1.59)	6.18(2.29)
1	10	10.56(5.39)	13.22(8.51)
2	34	33.88(22.24)	46.87(30.06)
3	73	73.30(49.34)	97.85(61.65)
3.5	99	97.79(64.78)	127.84(76.90)

and

$$Y_{11}, ..., Y_{n_{11}1}, ..., Y_{v_1 1}, ... \sim N(\eta_2 + \theta, \eta_1^2),$$

where n_{10}, n_{11}, v_0 and v_1 satisfy

$$\frac{n_{10}}{n_{10}+n_{11}} = \frac{n_{10}}{n_1} = \frac{v_1}{v_1+v_2} = \frac{v_1}{v} = r.$$

For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ on v observations (without internal sample size recalculation and known r and η_1) the two-sample t-test statistic

$$\begin{aligned} T_v &= \sqrt{\frac{v_2 v_1 (v-2)}{v}} \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{(v_2-1)S_1 + (v_1-1)S_0}} \\ &= \sqrt{v-2} \frac{\left(\bar{Y}_1 - \bar{Y}_0\right) \eta_1^{-1} \left(v_2^{-1} + v_1^{-1}\right)^{-1/2}}{\sqrt{(v_2-1)S_1\eta_1^{-2} + (v_1-1)S_0\eta_1^{-2}}} \\ &= \frac{Z + \omega_2}{\sqrt{(v-2)^{-1}\chi^2_{v-2}}} \end{aligned} \tag{7}$$

has a noncentral t-distribution with $v-2$ degrees of freedom and noncentrality parameter

$$\omega_2 = \frac{\theta_1 - \theta_0}{\eta_1 \sqrt{v_2^{-1} + v_1^{-1}}},$$

where $\bar{Y}_j = v_j^{-1} \sum_{i=1}^{v_j} Y_{ij}$, $S_j^2 = (v_j - 1)^{-1} \sum_{i=1}^{v_j} \left(Y_{ij} - \bar{Y}_j\right)^2$, $j = 0, 1$, Z is a standard normal random variable, and χ^2_{v-2} is a χ^2 random variable with $v-2$ degrees of freedom.

Then, v and k are found from

$$\begin{aligned} P\left(|T_v| > k | v-2, \omega_2\right) &= 1 - \beta, \\ P\left(|T_v| > k | v-2, 0\right) &= \alpha. \end{aligned} \tag{8}$$

Design $\mathbf{D}_{2t}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}\right)$ uses a two dimensional nuisance guess $\eta^{(0)} = \left(\eta_1^{(0)}, \eta_2^{(0)}\right)$. Its internal pilot counterpart $\mathbf{D}_{2t,IPN}\left(\alpha, \beta, \theta_0, \theta_1, n_1, n_{max}\right)$ uses $\hat{\eta}^{(0)} = (\hat{\eta}_1, \hat{\eta}_2)$ based on the interim pilot for sample size recalculation. We also consider a "restricted" naive internal pilot, $\mathbf{D}_{2t,IPNR}$, which is a variation of $\mathbf{D}_{2t,IPN}$, which sets $2n_1$ as the lower bound for the total sample size. This design approximates the Wittes-Brittain[9] idea of performing the interim analysis at half the originally planned sample size, and allowing only

Table 3: Monte-Carlo Type I error, Power, and Sample Sizes; 100, 000 simulations; two sample t-test designs; $n_1 = 20$ (10 per group); fixed allocation, $r = 0.5$

η_1	$\mathbf{D}_{2t}$	$\mathbf{D}_{2t,IPS}$	$\mathbf{D}_{2t,IPN}$	$\mathbf{D}_{2t,IPNR}$	$\mathbf{D}^a_{2t,IPN}$
	Type I error				
1	0.0497	0.0499	0.0584	0.0508	0.0526
1.5	0.0515	0.0506	0.0547	0.0542	0.0499
2	0.0492	0.0501	0.0523	0.0522	0.0493
2.5	0.0508	0.0500	0.0514	0.0514	0.0499
	Power				
1	0.8059	0.8147	0.8333	0.8952	0.8286
1.5	0.8044	0.8038	0.8159	0.8174	0.8108
2	0.8023	0.8043	0.8134	0.8134	0.8061
2.5	0.8020	0.8016	0.8091	0.8091	0.8002
	EN(SD)				
1	34	36.28(11.41)	36.28(11.41)	43.08(6.40)	37.34(12.35)
1.5	74	80.08(26.31)	80.08(26.31)	80.30(25.94)	81.24(25.67)
2	128	141.48(46.94)	141.48(46.94)	141.50(46.93)	140.38(45.23)
2.5	200	220.62(72.99)	220.62(72.99)	220.62(72.99)	215.88(70.60)

increase in the targeted study size. The use of internal resampling to find α_{new} and β_{new} defines $\mathbf{D}^a_{2t,IPN}$.

Simulations in Tables 3 and 4 consider only $r = 0.5$ for different sizes of internal pilot. This simplification allows us to compare the performance of several designs $\mathbf{D}_{1t,IPN}$ (fixed sample size, under the unrealistic assumption of a correctly guessed $\eta = \eta^{(0)}$), $\mathbf{D}_{2t,IPN}$ (naive internal pilot), $\mathbf{D}_{2t,IPNR}$ (restricted naive internal pilot), $\mathbf{D}_{2t,IPS}$ (naive internal pilot with Stein's approach to modify the test statistic, see [8]) and $\mathbf{D}^a_{2t,IPN}$ (internal resampling adjusted naive approach).

Internal pilot for a two sample t-test with a random group allocation

We saw that for a fixed sample size calculation with a fixed allocation ratio the test statistic (7) has a central (under H_0) or noncentral (under H_1)

Table 4: Monte-Carlo Type I error, Power, and Sample Sizes; $100,000$ simulations; two sample t-test designs; $n_1 = 10$ (5 per group); fixed allocation, $r = 0.5$

η_1	$\mathbf{D}_{2t}$	$\mathbf{D}_{2t,IPS}$	$\mathbf{D}_{2t,IPN}$	$\mathbf{D}_{2t,IPNR}$	$\mathbf{D}^{a}_{2t,IPN}$
	Type I error				
1	0.0507	0.0508	0.0636	0.0579	0.0526
1.5	0.0496	0.0503	0.0546	0.0537	0.0467
2	0.0499	0.0499	0.0510	0.0509	0.0469
2.5	0.0504	0.0496	0.0515	0.0515	0.0491
	Power				
1	0.8081	0.8140	0.8401	0.8446	0.8213
1.5	0.8093	0.8077	0.8261	0.8259	0.7995
2	0.8010	0.8030	0.8184	0.8183	0.7883
2.5	0.8046	0.8021	0.8167	0.8167	0.7852
	EN(SD)				
1	34	41.89(20.46)	41.90(20.46)	42.44(19.75)	41.77(19.36)
1.5	74	92.95(45.96)	92.95(45.96)	92.99(45.90)	88.01(40.94)
2	128	163.86(81.28)	163.86(81.28)	163.87(81.27)	150.85(72.31)
2.5	200	254.98(122.50)	254.98(122.50)	254.98(122.50)	232.94(112.03)

t-distribution. However random allocation of subjects to groups leads to a different distribution. Since only the noncentrality parameter dependens on v_1 and v_2, the distribution under H_0 does not change, but under H_1 it becomes a mixture with

$$P\left(|T_v| > k|v-2, \theta_1, \theta_0, \eta_3\right) = \sum_{v_1=0}^{v} \frac{v!}{v_1! v_2!} \eta_3^{v_1} \left(1-\eta_3\right)^{v_2} P\left(|T_v| > k|v-2, \omega_2(v_1, v_2)\right). \tag{9}$$

Moreover, the test statistic is not defined if $\min(v_1, v_2) \leq 1$ and has to be extended to these possible situations. For example, at $v_1 = 1$ or $v_2 = 1$ one can estimate the pooled standard deviation on one sample only; for the case $v_1 = v_2 = 0$ one can set $T_v = 0$. Thus, even a fixed sample size calculation faces substantial complications in deriving the distribution of the two sample t-test statistic under H_1.

In practice, the random aspect of the allocation is usually ignored in the sample size estimation formulas and the formula for a fixed allocation is used instead. Fixed allocation sample size calculation leads to two numbers v_1 and v_2, but real life recruitment rarely allows to enroll exactly v_1 and v_2 subjects in groups 1 and 2 respectively. Then, the investigator faces a dilemma: stop after recruiting $v_1 + v_2$ subjects even if there are not enough subjects in one of the groups, or continue recruiting until at least v_1 and v_2 subjects are enrolled in the groups 1 and 2 respectively. Our simulation study in Table 5 considers a fixed sample size calculation design ($\mathbf{D}_{2tr}$) with a recruitment until both numbers are met. By analogy with previous simulation examples we consider $\mathbf{D}_{2tr,IPN}$ (naive sample size recalculation), $\mathbf{D}_{2tr,IPNR}$ (restricted naive sample size recalculation, the total sample size is at least twice as large as the size of the internal pilot) and the adjusted by interim resampling design $\mathbf{D}^a_{2tr,IPN}$.

The results show that while type I error rate is only slightly inflated for the internal pilot designs, all the designs except our proposed method have above targeted power.

4 Example

The examples described in the simulation studies were relatively simple to allow comparisons with existing internal pilot methodology. In this section we show a more realistic example of a hypothetical observational study.

Table 5: Monte-Carlo Type I error, Power, and Sample Sizes; $100,000$ simulations; two sample t-test designs; $n_1 = 20$; random allocation

η_1	η_3	$\mathbf{D}_{2tr}$	$\mathbf{D}_{2tr,IPN}$	$\mathbf{D}_{2tr,IPNR}$	$\mathbf{D}^a_{2tr,IPN}$
		Type I error			
0.5	1	0.0480	0.0560	0.0502	0.0562
0.5	1.5	0.0500	0.0540	0.0535	0.0506
0.5	2	0.0499	0.0520	0.0520	0.0509
0.25	1	0.0508	0.0555	0.0529	0.0553
0.25	1.5	0.0497	0.0517	0.0516	0.0497
0.25	2	0.0502	0.0519	0.0519	0.0508
		Power			
0.5	1	0.8455	0.8543	0.9028	0.8070
0.5	1.5	0.8369	0.8444	0.8454	0.8181
0.5	2	0.8247	0.8384	0.8385	0.8116
0.25	1	0.8419	0.8669	0.8834	0.8264
0.25	1.5	0.8431	0.8515	0.8516	0.8235
0.25	2	0.8296	0.8429	0.8429	0.8145
		EN(SD)			
0.5	1	38.64(7.50)	41.12(16.36)	46.69(12.96)	36.78(16.54)
0.5	1.5	80.85(10.73)	90.51(36.48)	90.68(36.23)	84.99(33.60)
0.5	2	136.99(13.84)	158.68(62.09)	158.69(62.08)	147.03(56.55)
0.25	1	50.08(9.89)	58.99(28.36)	61.11(26.50)	54.06(29.26)
0.25	1.5	109.18(14.39)	128.22(59.05)	128.26(58.97)	120.27(58.77)
0.25	2	184.04(18.60)	222.00(95.56)	222.00(95.56)	206.66(95.59)

Measurements of prostate-specific antigen (PSA) levels are widely used for screening and diagnosing prostate cancer. PSA levels are known to be associated with measures of disease aggressiveness such as tumor stage as well as demographic characteristics predictive of screening behavior such as race/ethnicity, marital status, etc. A (hypothetical) investigator in Atlanta, GA wishes to conduct a study to evaluate whether the effect of Black versus White race on PSA levels is the same for localized versus regionally or distantly extended tumors. In practice he or she would turn to the SEER cancer registry, as we will for the source of data, but for the sake of the example let's assume that the information of interest is not available in the registry. In fact, PSA levels were not available in SEER until recently.

The specific goal of the study is to test the interaction effect of race (White vs Black) and tumor stage (localized vs others) on $\ln(PSA)$ values controlling for the effect of marital status (married vs others) and ethnicity (Hispanic vs others).

We use the linear regression model

$$\ln(PSA_i) = \beta_0 + \beta_1 \cdot W_i + \beta_2 \cdot L_i + \beta_3 \cdot W_i \cdot L_i + \beta_4 \cdot M_i + \beta_5 \cdot H_i + \epsilon_i, \quad (10)$$

where W_i, L_i, M_i, and H_i are, respectively, indicators of White race, localized tumor, married status, and Hispanic ethnicity of the i^{th} subject. The random noise ϵ_i is assumed to follow a normal model with the zero mean and a finite unknown variance σ^2. We formulate the research question about the interaction via $H_0 : \beta_3 = 0$ and wish to design a study that would have 80% power to detect a 1.5-fold difference in the race effect among the localized versus non-localized tumors, corresponding to $\beta_3 = \ln(1.5)$.

To calculate the study sample size we use the formula proposed by Hsieh et al [6]. If X represents the predictor of interest and Z stands the other predictors, then the sample size required to detect an effect with a partial regression coefficient of δ with power $100(1-\beta)\%$ at a two-sided significance level α is

$$N = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \phi^{-2}(r) + 3}{1 - R^2_{X.Z}}, \quad (11)$$

where $r = \delta\sigma_X/\sigma$, $R^2_{X.Z}$ is the multiple correlation coefficient of X and Z, and $\phi(r) = \frac{1}{2}\ln((1+r)/(1-r))$ is Fisher's z-transform. As typical, this expression includes multiple nuisance parameters that are difficult to obtain a priori.

Table 6: Linear regression on internal pilot data, $n_1 = 100$.

	Estimate	Std.Error	t value	p value
Intercept ($\hat{\beta}_0$)	5.4036	0.7208	7.497	<0.0001
White ($\hat{\beta}_1$)	-1.2507	0.6519	-1.918	0.0581
Localized ($\hat{\beta}_2$)	-1.4274	0.4916	-2.904	0.0046
Hispanic ($\hat{\beta}_4$)	0.1849	0.5394	0.343	0.7326
Married ($\hat{\beta}_5$)	-0.0928	0.2122	-0.437	0.6629
White × Localized ($\hat{\beta}_3$)	1.4046	0.6822	2.059	0.0423

To simulate the conduct of the study we extracted a sample of 8142 prostate cancer cases from the Atlanta Metropolitan area SEER registry. Our inclusion criteria limited our scope to records with black or white races, year of diagnosis 2004-2008, and observed PSA values. The data were sorted by year and month of diagnosis to mimic a prospective study. The results of internal pilot based on first $n_1 = 100$ observations are reported in Table 6.

To update α we performed $30,000$ internal resamplings. At the k^{th} iteration of this resampling, we performed the following steps:

1. used nonparametric bootstrap to resample the joint distribution of the predictors $\left(B_i^{(k)}, R_i^{(k)}, M_i^{(k)}, H_i^{(k)}\right)$, $i = 1, ..., n_1$;

2. generated simulated $\ln(PSA_i^{(k)})$ values under the null hypothesis ($\beta_3 = 0$) from the conditional normal distribution

$$N\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot B_i^{(k)} + \hat{\beta}_2 \cdot R_i^{(k)} + \hat{\beta}_4 \cdot M_i^{(k)} + \hat{\beta}_5 \cdot H_i^{(k)}, \hat{\sigma}^2\right), \qquad (12)$$

   where $\hat{\sigma}^2$ is the internal pilot estimate of σ^2;

3. fitted the regression model (10) on the k^{th} internal pilot resample;

4. estimated the total sample size $N^{(k)}$ using (11);

5. generated additional $N^{(k)} - n_1$ observations from the conditional distribution (12);

6. refitted the model (10) to the k^{th} resample with $N^{(k)}$ elements and tested H_0 with the P-value of the estimate of β_3.

Table 7: Regression model for the total sample, $N = 1837$.

	Estimate	Std.Error	t value	p value
Intercept ($\hat{\beta}_0$)	4.8725	0.1923	25.333	<0.0001
White ($\hat{\beta}_1$)	-0.1577	0.1267	-1.245	0.2134
Localized ($\hat{\beta}_2$)	-0.4670	0.0989	-4.721	<0.0001
Hispanic ($\hat{\beta}_4$)	-0.0148	0.1706	-0.086	0.9311
Married ($\hat{\beta}_5$)	-0.1396	0.0445	-3.136	0.0017
White × Localized ($\hat{\beta}_3$)	0.0766	0.1333	0.575	0.5653

Note that in Step 3 it is possible that some of the regression parameters are inestimable due to singularity of the design matrix. In our example, non-localized disease is rare, with only a few cases present in the pilot sample, the resampling Step 1 occasionally leads to a zero vector $W \cdot L$ in the resampled design matrix. The treatment of such "exceptions" should be specified in the design of the study, which should include provisions for the situation that the internal pilot sample results in a non-full rank design matrix. The interim resampling procedure should follow the same rules. In this example, we assume that the investigator pre-planned to resolve this exception by setting the total sample size to the prespecified maximal value of $n_{max} = 3,000$.

The above $30,000$ iterations allow us to estimate $\hat{\alpha}$ with a standard error of 0.0013 and calculate α_{new}. In this example, $\alpha_{new} = 0.0492$.

A similar resampling scheme emulating the design behavior under H_1 leads to a new power $1-\beta_{new} = 0.7815$ and the final total sample size of 1837. The results of model fit on 1837 observations are reported in Table 7. We fail to reject the null hypothesis that there is no interaction ($0.5653 > 0.0491$). While it is unknown whether it is a correct decision, an analysis of all 8142 cases leads to a similar conclusion.

5 Discussion

In this manuscript we suggested considering the interim sample size recalculation from the prospective of internal design resampling. The proposed approach is very flexible, and can be applied for a wide variety of situations, including prospective clinical trials, observational, and retrospective studies. We defined a study design as a set of rules for sample size cal-

culation/recalculation, the chosen test statistic, and a course of actions for dealing with rare but possible situations such as singularity of a design matrix, etc. Ultimately the design is the ability to make a decision for all possible realizations of a random variable.

The use of internal design-resampling allows elimination of the formal dependence of a design on nuisance parameters and lowers their notorious effect on type I and II errors. We adjusted the type I error α and the power $1 - \beta$ using information from the internal pilot data. The adjusted type I error, α_{new}, and the adjusted power, $1 - \beta_{new}$, are used for sample size re-estimation at the interim analysis.

We emphasize that it is critical to have a clearly defined per-protocol design for all possible samples. Otherwise, internal design resampling may generate "impossible" sampling situations bringing additional uncertainty into the adjustment.

The internal resampling can be computationally challenging especially when a Monte-Carlo study is used. We implemented and performed three designs for internal sample size recalculation: the paired data t-test, the independent samples t-test with a pre-defined allocation ratio, and the independent sample t-test with random allocation. To speed up the Monte-Carlo simulations we used C code pre-compiled into a shared object for internal resampling part, and R code for the rest of the program.

Our simulation studies clearly show that our resampling methodology leads to a generally better control of type I and II error than the naive internal pilot design across all considered scenarios. For the simulation scenarios for the two-sample t-tests, the comparisons with Stein's, the naive, and the restricted Wittes and Brittain approaches, the internal design resampling also often keeps a better control for type I and II errors and has a smaller mean sample size.

We deliberately considered small internal pilot sample sizes (5 or 10 per group) since moderate to large sample sizes generate relatively accurate estimates of nuisance parameters and all considered methods would show similar performance. However our regression example shows that in situations with a larger number of nuisance parameters even 100 internal pilot samples might not be sufficient to assume no inflation of error rates.

The internal design resampling requires the defined in the study protocol internal pilot sample size (n_1) and the largest possible (n_{max}) sample sizes. If the tails of the distribution of the sample size obtained from the internal pilot heavily spill over these two lower and upper bounds for the sample size,

then the control for type I and II error becomes more problematic.

Overall, we strongly recommend researchers to clearly define their designs for all possible situation and incorporate internal sample size recalculation in their study designs. This may substantially improve the use of their resources, better control type I and II errors, and protect against nuisance parameters misspecification.

References

[1] Christopher S Coffey and Keith E Muller. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine*, 18(10):1199–214, May 1999.

[2] Tim Friede and Meinhard Kieser. Sample Size Recalculation in Internal Pilot Study Designs: A Review. *Biometrical Journal*, 48(4):537–555, August 2006.

[3] Tim Friede and Meinhard Kieser. Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical statistics*, 10(1):8–13, 2011.

[4] Deborah H Glueck and Keith E Muller. Adjusting power for a baseline covariate in linear models. *Statistics in Medicine*, 22(16):2535–51, August 2003.

[5] Matthew J Gurka, Christopher S Coffey, and Kelly K Gurka. Internal pilots for observational studies. *Biometrical journal. Biometrische Zeitschrift*, 52(5):590–603, October 2010.

[6] F. Y. Hsieh, Daniel A. Bloch, and Michael D. Larsen. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17(14):1623–1634, 1998.

[7] Michael A Proschan. Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics*, 15(4):559–74, January 2005.

[8] Charles Stein. A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258, 1945.

[9] Janet T Wittes and Erica Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–71; discussion 71–2, 1990.