Observational Studies 2 (2016) 65-85 Submitted 10/15; Published 9/16

An Interim Sample Size Recalculation for Observational Studies

Sergey Tarima **starima@mcw.edu**
Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Rd
Wauwatosa, WI 53226

Peng He **hep@medimmune.com**
Medimmune
1 MedImmune Way
Gaithersburg, MD 20878

Tao Wang **taowang@mcw.edu**
Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Rd
Wauwatosa, WI 53226

Aniko Szabo **aszabo@mcw.edu**
Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Rd
Wauwatosa, WI 53226

Abstract

Interim sample size re-estimation (SSR) often affects the Type I and II error rates. We propose and investigate a method based on resampling the whole study design at the interim analysis. This resampling starts with SSR at the observed interim analysis values of nuisance parameters and finishes with the decision to accept or reject the null hypothesis. The proposed approach finds a new critical value and an updated sample size. As shown in a Monte-Carlo simulation study this resampling method shows superior performance for logistic regression when compared with the naïve SSR. Another set of simulation studies shows comparable performance of the resampling and several previously published procedures for a 2-sample t-test with random allocation. An illustrative example highlights the benefits of our approach for logistic regression analysis.

Keywords: Interim sample size reestimation, logistic regression, observational studies

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.

Having spent substantial effort and resources on the pilot study, investigators often wish to include the pilot data in the final data analysis. This inclusion however has to be planned upfront at the study design stage. The use of the pilot data along with subsequently collected data requires an adjustment in order to control or better control the type I error rate and power. The methodology of interim sample size re-estimation (SSR) 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 (2005) and Friede and Kieser (2006) reviewed interim SSR. The majority of work in this area has been focused on two-arm prospective randomized studies with normally distributed or binomial outcomes. SSR for a general linear hypothesis tested through the general linear model was explored by Coffey and Muller (1999). An interesting extension toward linear mixed models was suggested by Coffey and Muller (2003). An interim pilot design for the analysis of covariance model (ANCOVA) was explored via simulations by Friede and Kieser (2011), while Gurka et al. (2010) analyzed the effect of unknown group size proportions.

In this manuscript we propose a general methodology for interim SSR where 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 first stage (see Section 2). This resampling is performed under both the null and alternative hypotheses allowing the estimation and correction for bias of the type I error and power. We show that deviation from the desired type I error and power is $o\left(\frac{1}{\sqrt{n}}\right)$, where n is the pilot sample size. The proposed resampling method targets testing procedures in the presence of multiple nuisance parameters where researchers rely on asymptotic behavior of test statistics.

Design resampling considered in this manuscript is a general approach and may be used in various setting. More specifically, potential applications exist in the area of non-randomized studies for testing hypotheses about regression coefficients, for example, in a logistic regression model.

Section 3 applies the developed approach to the logistic regression model and illustrates its benefits in a Monte-Carlo simulation study. Section 4 investigates asymptotic behavior of the power estimators accounting and not accounting for interim SSR. Power properties against local alternatives are considered. Additional simulation studies are reported in Section 5. An illustrative example in Section 6 applies the resampling to a logistic regression model with three predictors, one binary and two continuous. The article is concluded with a short summary in Section 7.

2. Study design and sample size recalculation

We explore study designs when $n(< n_{max})$ subjects are enrolled in the study as the first stage data and the second stage sample size is recalculated at the interim analysis to better control the desired power, $1-\beta$, while maintaining the type I error, α. 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 and n_{max}. Nuisance parameters are not controlled by the study design and will not be included as arguments of $\mathbf{D}$. At the same time the statistical properties of a statistical procedure used in tandem with a study design do depend on η.

For example, if $\mathbf{D}_f = \mathbf{D}_f\left(\alpha,\beta,\theta_0,\theta_1,\eta^{(0)}\right)$ denotes the fixed sample size study design with an assumed value $\eta^{(0)}$ for the unknown η, the power function for the tandem of this $\mathbf{D}$ and a test statistic T_v is

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

where $v(\alpha,\beta,\theta_0,\theta_1,\eta^{(0)})$ is a sample size formula and $k(v)$ is a critical value.

Fixed sample designs, $\mathbf{D}_f$, can be augmented to become designs with SSR by adding rules/parameters describing the interim SSR procedure. A naïve approach to interim SSR uses $\hat{\eta}$, an interim analysis estimate of η without further adjustments. In this manuscript we let $\hat{\eta}$ be the MLE of η at the interim analysis. The interim MLE of the nuisance parameters was used before in adaptive designs literature, see for example, Lane and Flournoy (2012) and Lane et al. (2014).

Moreover, the total sample size is bracketed between n and n_{max} since it cannot be less than the size of the first stage data, (n), and cannot be greater than n_{max}, which value is chosen from practical budgetary or study time constraints. Then, the naïve SSR design has a different set of arguments,

$$\mathbf{D}_{ip} = \mathbf{D}_{ip}(\alpha,\beta,\theta_0,\theta_1,n,n_{max}),$$

is an alternative to $\mathbf{D}_f$, which does not use $\eta^{(0)}$ but depends on n and n_{max}. Its power conditional on $\hat{\eta}_\theta$ is

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

where $\hat{\eta}_\theta$ depends on n, n_{max}, the true value η, and possibly θ.

A naïve SSR for a two sample t-test is known to have an increased type I error (Wittes and Brittain, 1990). In more complex situations such as SSR for a logistic regression model we can see not only an increase but also a decrease of the type I error.

2.1 Integrated type I error and power

Interim SSR treats the final sample size as 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 may be achieved by, for example, Stein's approach (Stein, 1945), which estimates the unknown η using pilot data only. This is associated with a loss of empirical information and a modification of the test statistic, which we avoid throughout this manuscript. Stein's approach is also difficult to generalize when a more complex situation than a linear model is considered. In general, unless specific steps are taken in a limited set of study designs, the type I error rate is rarely exactly controlled at a desired level. This is especially true when the test statistic depends on the nuisance parameters, like many MLEs in generalized linear models, for example, the multiple logistic regression model. Even the so called "exact logistic regression" is not really exact but only conditionally exact. These are the settings where we expect most applications of our methodology.

For a design $\mathbf{D}$, the type I error and power are

$$E_{\hat{\eta}}\left[E_{Y|\hat{\eta}}\left[\delta\left(\mathbf{D}\left(\alpha,\beta,\theta_0,\theta_1,n,n_{max}\right)\right)\middle|H_0,\hat{\eta}\right]\middle|\eta\right] = a\left(\alpha,\beta|\mathbf{D},\eta\right) \neq \alpha \tag{3}$$

and

$$E_{\hat{\eta}}\left[E_{Y|\hat{\eta}}\left[\delta\left(\mathbf{D}\left(\alpha,\beta,\theta_0,\theta_1,n,n_{max}\right)\right)\middle|H_1,\hat{\eta}\right]\middle|\eta\right] = 1 - b\left(\alpha,\beta|\mathbf{D},\eta\right) \neq 1-\beta. \tag{4}$$

The outer expectation is taken with respect to the distribution of $\hat{\eta}$. This distribution depends on the true value of the nuisance parameter η, which keeps the true type I error and power unknown in real-life applications.

2.2 Sample size recalculation via resampling

We propose a new approach to SSR after the first stage data that conditionally on $\hat{\eta}$ maintains both the (integrated) type I and II error rates.

For a design $\mathbf{D}$ we find α_{new} and β_{new} to control the desired type I error and power conditional on the first stage data based estimates of η, so that

$$E_{\hat{\eta}}\left[E_{Y|\hat{\eta}}\left[\delta\left(\mathbf{D}\left(\alpha_{new},\beta_{new},\theta_0,\theta_1,n,n_{max}\right)\right)\middle|H_0,\hat{\eta}\right]\middle|\hat{\eta}\right] = \alpha \tag{5}$$

and

$$E_{\hat{\eta}}\left[E_{Y|\hat{\eta}}\left[\delta\left(\mathbf{D}\left(\alpha_{new},\beta_{new},\theta_0,\theta_1,n,n_{max}\right)\right)\middle|H_1,\hat{\eta}\right]\middle|\hat{\eta}\right] = 1-\beta. \tag{6}$$

This definition leads to a design with SSR, $\mathbf{D}^a(\alpha,\beta,\theta_0,\theta_1,n,n_{max})$, based on the SSR formula and the hypothesis testing procedure defined in $\mathbf{D}$. The superscript a in $\mathbf{D}^a(\cdot)$ declares that the design has been adjusted to secure the desired type I error and power conditional on $\hat{\eta}$.

The solution to (5) and (6) will exactly control both the type I and II errors if the distribution of $\hat{\eta}$ and the value η (as a possible parameter of the distribution of $\hat{\eta}$) are known exactly. In this case, the formulas (5) and (6) will coincide with the type I error and power (3) and (4). The likelihood principle states that all sample information is incorporated in the likelihood function. We use the MLE estimation on first stage data to obtain the profile likelihood of the vector of nuisance parameters. This likelihood conditional on $\hat{\theta}$ (the observed value θ at the interim analysis) is integrated out in (5) and (6). Thus, we argue that conditionally on $\hat{\theta}$ the profile likelihood of η absorbs all information from the first stage that is relevant to the nuisance parameters.

Below we consider an algorithm for finding α_{new} and β_{new} in Equations (5) and (6). This algorithm uses two loops. The inner loop estimates type I and type II errors at some fixed values of α_{new} and β_{new}. The outer loop changes α_{new} and β_{new} until the solution of (5) and (6) is reached.

Before the loops start their work, at the interim analysis, we estimate $\hat{\eta}$ and set initial values of α_{new} and β_{new} to α and β respectively.

2.2.1 Inner loop

The inner loop performs the following resampling procedure M times. For each iteration i $(i = 1, ..., M)$, we

1. using a pseudo-random number generator, draw a random sample $\left(Y_1^{(i)}, ..., Y_n^{(i)}\right)$ from $f_Y\left(y|\theta_0, \hat{\eta}\right)$,
2. obtain the total sample size $v_i \in [n, n_{max}]$ based on these n observations; v_i is a function of α_{new} and β_{new} as well as the nuisance parameters estimated on $\left(Y_1^{(i)}, ..., Y_n^{(i)}\right)$,
3. using pseudo random numbers, generate additional $(v_i - n)$ observations $\left(Y_{n+1}^{(i)}, ..., Y_{v_i}^{(i)}\right)$ from $f_Y\left(y|\theta_0, \hat{\eta}\right)$, and
4. calculate $T_{v_i}^{(i)}$ on this i^{th} sample.

This resampling procedure is similar to a bootstrap procedure under the assumption on θ, but the key difference is that we simulate the performance of the whole study design. We reproduce everything: the first stage data, interim SSR, the second stage data, the final test statistic, and the binary decision the test statistic leads to.

Another important distinct aspect of this resampling procedure is that the underlying random variable Y $(\sim f_Y)$ can have two parts: the outcome, which is usually one dimensional and covariates, which are potentially multivariate. Then, the generation of $Y^{(i)}$ is divided into two sequential steps. First, covariates are generated by a non-parametric bootstrap procedure. Second, the outcome is generated using a model-based functional form of the conditional dependence with conditioning on θ_0, $\hat{\eta}$ and the generated values of covariates. The second step is analogous to a parametric bootstrap.

The covariates can also be generated from a parametric distribution if the functional form of the distribution of covariates is known. Our choice of non-parametric resampling of covariates works especially well with categorical covariates since non-parametric resampling

becomes sampling from a multinomial distribution with parameters substituted by their first stage MLEs.

The estimated type I error rate is

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

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

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

or

$$\alpha_{new}=\frac{\alpha^2\left(1-\hat{a}\right)}{\left(1-\alpha\right)^2\hat{a}+\alpha^2\left(1-\hat{a}\right)}. \tag{7}$$

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

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

leads to the bias-corrected value

$$\beta_{new}=\frac{\beta^2\left(1-\hat{b}\right)}{\left(1-\beta\right)^2\hat{b}+\beta^2\left(1-\hat{b}\right)}. \tag{8}$$

The use of logit assures us that the chosen α_{new} and β_{new} are secured within the unit interval.

2.2.2 Outer loop

Formulas (7) and (8) correct for bias but do not necessarily produce a solution to (5) and (6). To solve (5) and (6) with respect to α_{new} and β_{new} we incorporate formulas (7) and (8) inside an outer loop. Let o denote the iterator of this outer loop, then (7) and (8) in their logit form become

$$l\alpha_{new}^{(o+1)} = l\alpha_{new}^{(o)}-[l\hat{a}^{(o)}-l\alpha], \tag{9}$$
$$l\beta_{new}^{(o+1)} = l\beta_{new}^{(o)}-[l\hat{b}^{(o)}-l\beta]. \tag{10}$$

Equations (9) and (10) represent a simplified version of Newton's root-finding procedure for solving the bivariate equation

$$f(l\alpha_{new},l\beta_{new})=\left[l\hat{a}\left(\alpha_{new},\beta_{new}\right)-l\alpha,l\hat{b}\left(\alpha_{new},\beta_{new}\right)-l\beta\right]=[0,0]$$

Newton's algorithm applied separately for each component of $f(\cdot)$ leads to the iterative procedure

$$l\alpha_{new}^{(o+1)} = l\alpha_{new}^{(o)} - \frac{l\hat{a}\left(\alpha_{new}^{(o)}, \beta_{new}^{(o)}\right) - l\alpha}{f'_{l\alpha_{new}}\left(l\alpha_{new}^{(o)}, l\beta_{new}^{(o)}\right)}, \tag{11}$$

$$l\beta_{new}^{(o+1)} = l\beta_{new}^{(o)} - \frac{l\hat{b}\left(\alpha_{new}^{(o)}, \beta_{new}^{(o)}\right) - l\beta}{f'_{l\beta_{new}}\left(l\alpha_{new}^{(o)}, l\beta_{new}^{(o)}\right)}. \tag{12}$$

Our simulation studies show that the slope of the $f(\cdot)$ with respect to $l\alpha_{new}$ is approximately equal to 1 for a given $l\beta_{new}$. The slope of $f(\cdot)$ with respect to $l\beta_{new}$ may vary around 1 conditional on $l\alpha_{new}$, which does not affect the convergence in our simulation experiments. Convergence is usually reached in 2 to 6 outer iterations.

More complex situations with a multi-dimensional nuisance parameter may require a more sophisticated optimization procedure. One of the easiest to implement is the secant method because it does not require the derivation or estimation of the first or second derivatives of $f(\cdot)$.

We highlight that v_i is a random variable, a function of α_{new}, β_{new}, and the estimated nuisance parameter, which changes at every iteration of the inner loop. This makes our convergence criteria depend only on α_{new} and β_{new}. The values of α_{new} and β_{new} change only at the outer loop iterations, the inner loop is needed only to estimate the type I and II errors associated with specific choices of α_{new} and β_{new}. To improve precision of type I and II error estimation we increase the number of inner loop iterations at every iteration of the outer loop. The actual value of the total recalculated sample size is obtained as $v_i = v_i(\alpha_{new}, \beta_{new}, \theta_0, \theta_1, \hat{\eta})$ from the existing sample size formula.

In our simulation experiments the outer loop uses 30 iterations, the number of iterations in the internal loop M_o is increasing by 5000 for each new outer iteration $M_{o+1} = M_o + 5000$, $M_1 = 5000$.

2.3 Other approaches

To the best of our knowledge, manipulation of both input α and β to find a solution to (5) and (6) has not been considered before. Moreover there were no suggested approaches for SSRs for generalized linear models and multiple nuisance parameters.

Previously, the design with SSR was suggested for linear contrast testing in univariate linear regression models with a single nuisance parameter, σ^2 (Coffey and Muller, 1999). This approach is implemented in SAS IML (Kairalla et al., 2008), and is based on numeric integration to solve (5) with respect to α_{new}. The target power was secured by sequential sample size increases. This SAS IML software also suggests a sensitivity analysis for inaccurate estimates of σ^2 and deals with several different modifications of the test statistic. The absence of a compiled binary program requires substantial computing resources, which limits the quality of Monte-Carlo simulation experiments. This software also implements the bounding method, see Coffey et al. (2007), which adjusts α to the most conservative scenario for controlling the type I error.

In contrast, our approach is based on Monte-Carlo type integration and finds a simultaneous solution to both (5) and (6). Moreover, we are not limited to a single nuisance parameter as illustrated in our simulation studies. The suggested resampling approach allows us to solve (5) and (6) for various study designs, whereas a numeric grid-based or other non-Monte-Carlo integration approaches require substantial analytic work to derive proper formulas.

The use of bootstrap resampling was used before (Betensky and Tierney, 1997) to generate potential realizations of future data. In contrast, we resample the whole study design and apply our methods in more general settings of multiple nuisance parameters.

Following section compares in a simulation study the control of type I error and power between the naïve and adjusted approaches.

3. Interim sample size recalculations for multiple logistic regression

Multiple logistic regression is a much more complex model for interim SSR than the two sample t-tests. Below we report the difficulties we faced and the solutions we applied.

1. *Difficulty:* The MLEs of nuisance parameters, $\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_p$ depend on the parameter of interest β_1.
   Solution: The MLEs of the nuisance parameters were obtained under the MLE of β_1, $\hat{\beta}_1$. Then, we use the same estimates of $\beta_0, \beta_2, \beta_3, ..., \beta_p$ under H_0 and H_1, but the distributions of $\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_p$ continue being different under H_0 and H_1 which leads to different on average sample sizes. The convenient part of this estimation is that we are keeping consistent and asymptotically efficient estimates of the nuisance parameters under any β_1.

2. *Difficulty:* The distribution of covariates is not known.
   Solution: We used non-parametric resampling to generate samples of the covariates.

3. *Difficulty:* Absence of the closed form solution for the MLEs requires Newton-Raphson approximation and makes the resampling scheme computationally intensive,
   Solution: The computational intensity was alleviated by implementing Newton-Raphson algorithm for logistic regression along with the resampling scheme described in Section 2.2.2 in C++. Then, a *DLL*-library was compiled and called from *R* during the simulation experiment. A Windows computer with a four core CPU (INTEL I5) and 4GB of memory was used in this simulation experiment. Single *DLL* calls took mainly 1 to 10 minutes depending on when the convergence criteria is met.

4. *Difficulty:* Resampling may end up with datasets with quasi-separation (no unique MLEs) or singular design matrices.
   Solution: These situations were resolved from a simple premise that such interim samples stop the study and declare inconclusiveness. Thus, these scenarios were eliminated from consideration during resampling.

5. *Difficulty:* There is a possibility that resampling may generate first stage data with all outcomes equal to 0 or all equal to 1.
   Solution: Similarly, for such situations we declared inconclusiveness and eliminated such samples from our resampling.

In Monte-Carlo simulation studies reported in this section we consider two continuous random predictors X_{i1} and X_{i2} generated from two standard independent normal distributions. The outcome Y_i was a Bernoulli random variable with

$$P\left(Y_i = 1 | X_{i1} = x_1, X_{i2} = x_2\right) = \left(1 + \exp\left(-\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)\right)^{-1}, \tag{13}$$

where $\beta_0 = \beta_2 = 0$. The objective is to test $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 = 1.127$ at a 5% significance and 80% power.

The naïve interim SSR will be performed using

$$N = \mathcal{I}^{-1}\left(\beta_1\right) \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{1.127^2}, \tag{14}$$

where $\mathcal{I}^{-1}\left(\beta_1\right)$ is the second diagonal element of $\mathcal{I}^{-1}\left(\mathbf{b}\right)$, the asymptotic variance of $\hat{\beta}_1$. We estimate $\mathcal{I}^{-1}\left(\beta_1\right)$ on the first stage data as $n\left(SE(\hat{\beta}_1)\right)^2$, where $SE(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$ readily available from a logistic regression fit on the first stage data.

3.1 A monte-Carlo study with naïve sample size recalculation

A Monte-Carlo simulation study with $50,000$ simulations under H_0 and $50,000$ simulations under H_A with $n = 20$, $n_{max} = 100$, $\delta = 1.127$ and naïve SSR showed a type I error of 0.0303 and statistical power of 0.8737. Thus, the naïve SSR shows a substantial deflation of the type I error and a serious overpowering, which immediately leads to larger than needed sample sizes. Wald's P-value < 0.05 was used to claim significance, and its asymptotic nature led to bias in type I and II errors at small to moderate sample sizes.

3.2 A Monte-Carlo study with a naïve SSR adjusted for study design.

We performed a Monte-Carlo study for the logistic regression model (13). Under H_0 the study was based on $11,000$ repetitions and for each repetition we performed *both* the naïve and adjusted sample recalculations. The type I error at the naïve SSR was 0.031, which repeats a similar finding in Section 3. The adjusted values $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$ led to type I error of 0.042. When we averaged $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$ across all $11,000$ repetitions, the averaged $\hat{\alpha}_{new}$ was 0.073 and the averaged $1 - \hat{\beta}_{new}$ was 0.707. The sample size on average decreased from 43.3 to 32.8. In a similar manner we performed $2,100$ repetitions under H_A. The power on average went down from 0.882 to 0.810, the averaged sample size went down as well, from 71.2 to 58.2. The averaged $\hat{\alpha}_{new}$ and $1 - \hat{\beta}_{new}$ were 0.073 and 0.719, which are very similar to our Monte-Carlo experiment under H_0.

Table 1 shows an example of internal steps of an iterative procedure for finding $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$. The values of $\hat{a}$ $(= 0.031400)$ and $1 - \hat{b}$ $(= 0.886600)$ in the first line of Table 1 estimate the type I error and power at $\alpha (= 0.05)$ and $1 - \beta (= 0.8)$. Using the formulas from Section 2.2.1 we obtain a new α, $\hat{\alpha}_{new} (= 0.078722)$, and a new $1 - \beta$, $1 - \hat{\beta}_{new} (= 0.671751)$, which lead to the estimated values of the type I error of 0.051000 and power of 0.791900. In this example, even a single iteration gives a good approximation to the solution of (5) and (6). Similar findings were observed for t-tests when the major change to $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$ was done at the first iteration. This benefit of the first iteration, however, is not always present. The illustrative example in Section 6 clearly shows that in some situations more iterations may be needed.

Table 1: Steps of iterative procedure searching for $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$ securing the desired $\hat{a}$ and $\hat{b}$, $n = 20$, $n_{max} = 100$, M defines a sample size for an outer loop iteration, the convergence is met when $\Delta = (\hat{a} - \alpha)^2 + \left(\hat{b} - \beta\right)^2 < 0.00001$.

M	Δ	$\hat{a}$	1- $\hat{b}$	$\hat{\alpha}_{new}$	1- $\hat{\beta}_{new}$
5000	0.007846	0.031400	0.886600	0.078722	0.671751
10000	0.000067	0.051000	0.791900	0.077223	0.682654
15000	0.000107	0.056067	0.808387	0.069035	0.671003
20000	0.000063	0.042700	0.796950	0.080458	0.675174
25000	0.000055	0.055520	0.795080	0.072650	0.681820
30000	0.000015	0.047033	0.802400	0.077101	0.678542
35000	0.000002	0.049857	0.798429	0.077316	0.680675

4. Asymptotic properties

First, we emphasize that the interim SSR proposed in this manuscript does not exactly control the type I error and power for finite samples. On the other hand, there are many hypothesis testing procedures, such as multiple logistic regression, which even without interim sample size recalculation secure type I error control only asymptotically. This section shows that our design resampling approach controls the type I error and power asymptotically against a local alternative in the presence of interim SSR.

4.1 Power curve

The type I error and the power are two points on the power curve

$$\begin{aligned} PWR(\alpha, \beta|\mathbf{D}, \theta, \eta) = & \quad E_{\hat{\eta}}\left[E_{Y|\hat{\eta}}\left[\delta\mathbf{D}(\alpha, \beta, \theta_0, \theta_1, n, n_{max})\left|\hat{\eta}\right.\right]\right|\theta, \eta\Big] \\ = & \int_G E_{Y|\tau}\left[\delta\mathbf{D}(\alpha, \beta, \theta_0, \theta_1, n, n_{max})\left|\tau\right.\right] g_n(\tau|\theta, \eta)\, d\tau \\ = & \int_G Pr\left[\delta\mathbf{D}(\alpha, \beta, \theta_0, \theta_1, n, n_{max}) = 1\left|\tau\right.\right] g_n(\tau|\theta, \eta)\, d\tau, \end{aligned} \tag{15}$$

where $g_n(\tau|\theta, \eta)$ is the distribution of $\hat{\eta}$ defined for all possible values of the first stage data $Y_1, ..., Y_n$.

It is critical to define $\hat{\eta}$ for all possible values of the first stage data since the decision δ should be made at any $Y_1, ..., Y_n$. This means that when something unusual happens, such as lack of convergence of a Newton-Raphson algorithm in its search for the MLE $\hat{\eta}$ (singular design matrix, quasi-separation, etc.) we still should be able to make a statement about δ. This is why we introduced the concept of design into the picture and we call these unusual situations "exceptions" by analogy with computer science terminology. For example, what should be done if quasi-separation happened? One possibility is to stop the study and declare it inconclusive. Other options include acceptance of H_0 or H_1, or we can proceed with further sampling under an assumed or administratively assigned values of $\hat{\eta}$. Then, all

possible outcomes of $\hat{\eta}$ can be separated into two mutually exclusive categories G_{reg}, where certain regularity conditions hold and the MLE can be found and G_{ex} where "exceptions" should be resolved in an administrative or common sense manner, $G = G_{reg} \cup G_{ex}$.

To eliminate the undue influence of "exceptions" on asymptotic properties we require $\sqrt{n}Pr(\hat{\eta} \in G_{ex}) = o(1)$. For example, if the support of $\hat{\eta}$ is the whole real line, $\mathcal{R}$, then $G_{reg} = \mathcal{R}$ and G_{ex} may refer to exceptional events such as missing value (NA), divergence (DV), quasi-separability (QS), perfect colinearity (CL). Then $G = \{\mathcal{R} \cup NA \cup DV \cup QS \cup CL\}$. For many exceptional situations, this is a realistic assumption. For example, one exceptional situation for a binary outcome is associated with the possibility when all outcome values are the same in the first stage data, but the probability of this event converges to zero with the power law, which is faster than $\frac{1}{\sqrt{n}}$.

The decision function δ at $\hat{\eta} \in G_{reg}$ fully depends on a chosen test statistic, $T_N(\eta)$, its sample size formula $N = N(\alpha, \beta|\eta)$, and the planned decision rule on rejecting H_0, $T_N(\eta) > k_N$. The critical value $k_N = k_N(\eta)$ is selected under the assumption that $T_N \sim F$, where F is an (assumed) distribution of T_N. This distributional assumption may be correct for a fixed sample size but in interim SSR settings it is misspecified. The decision rule based on T_N, $T_N(\eta) > k_N$, can be substituted by the decision rule based on its corresponding P-value, $Pr(T_N(\eta) > k_N) < \alpha$.

We rewrite the true power function as

$$PWR(\alpha, \beta|\mathbf{D}, \theta, \eta) = Pr(\hat{\eta} \in G_{ex}) \int_{G_{ex}} Pr\left[Pr(\delta = 1)\left|\tau\right.\right] g_n(\tau|\theta, \eta)\, d\tau + Pr(\hat{\eta} \in G_{reg}) \int_{G_{reg}} Pr\left[Pr(T_N(\tau) > k_N) < \alpha\left|\tau\right.\right] g_n(\tau|\theta, \eta)\, d\tau, \quad (16)$$

The first term of the right hand side of (16) integrates over the situations when exceptions happen. If the decision δ for a finite number of exceptions is defined on a case by case basis then $g_n(\tau|\theta, \eta)$ is a discreet probability measure defined on G_{ex} and the integral becomes a weighted sum. We are less interested in the first term of (16) since we require it to quickly converge to zero and disappear when asymptotic properties are investigated.

The $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$ describes a family of power curves indexed by α and β. In Section 4.2 we assume that α and β are two fixed quantities extracting a single power curve from the family. Obviously, this curve is not known due to the unknown θ and η.

4.2 Local power curve estimation

Let $\left(\hat{\theta}, \hat{\eta}\right)$ be the first stage MLE of (θ, η). Then, by the invariance property of the MLE $PWR\left(\alpha, \beta|\mathbf{D}, \hat{\theta}, \hat{\eta}\right)$ is the MLE of $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$ defined at $\hat{\eta} \in G_{reg}$. Since we are mainly interested in power at specific values of θ (θ_0 or θ_1) we use $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$ to estimate $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$. We estimate η at the observed alternative, the MLE $\hat{\theta}$. From the asymptotic properties of the MLE

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta \\ \hat{\eta} - \eta \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_{\theta\theta} & I_{\theta\eta} \\ I_{\theta\eta} & I_{\eta\eta} \end{pmatrix}^{-1}\right), \quad (17)$$

where the $I_{\cdot,\cdot}$ are the elements of the Fisher information matrix for a single observation, we find the asymptotic distribution of $\hat{\eta}$,

$$\sqrt{n}\left(\hat{\eta}-\eta\right) \xrightarrow{d} N\left(0, v_{\eta}^{2}\right), \tag{18}$$

where $v_{\eta}^{2}=\left(I_{\eta\eta}-I_{\theta\eta}I_{\theta\theta}^{-1}I_{\theta\eta}\right)^{-1}$.

Without loss of generality we are investigating a simple null $\theta=0$ versus a local alternative at $\theta=\frac{h}{\sqrt{n}}$. In the local asymptotics framework the sample size formula accounting for the local alternative is

$$N=nv_{\theta}^{2}\frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^{2}}{h^{2}}, \tag{19}$$

where $v_{\theta}^{2}=\left(I_{\theta\theta}-I_{\theta\eta}I_{\eta\eta}^{-1}I_{\theta\eta}\right)^{-1}$. Equation (19) states that the total sample size should depend linearly on the sample size for the first stage data and v_{θ}^{2}.

Since the maximum likelihood ratio (MLR) tests are asymptotically optimal by the Neyman-Pearson lemma we also assume that the chosen test statistic T_N retains the same optimal properties asymptotically, otherwise we could use $\hat{\theta}$ instead. We also assume that regularity conditions ensure that T_N converges to a chi-squared random variable with one degree of freedom under the null. Under the local alternative the convergence is to a noncentral chi-squared random variable with the noncentrality parameter $h^{2}v_{\theta}^{-2}$ and one degree of freedom. Then, as $n\to\infty$

$$Pr\left[T_{N}(\eta)>k_{N}\right]\to Pr\left[\chi_{1}^{2}\left(\frac{h^{2}}{v_{\theta}^{2}}\right)>Q(\chi_{1}^{2},1-\alpha)\right], \tag{20}$$

where $Q(\chi_1^2, 1-\alpha)$ is the $(1-\alpha)$ level quantile of the central chi-squared distribution and $\chi_1^2(\cdot)$ is a noncentral chi-squared random variable with the non-centrality parameter defined by its argument in parentheses. Asymptotically, direct dependence on η goes away whereas dependence on $I_{\eta\eta}$ and $I_{\eta\theta}$ stays and is incorporated in v_{θ}^{2}. Hence, v_{θ}^{2} is the only asymptotic nuisance parameter.

The asymptotic approximation of the local power function (20) will be needed for practical application of Lemma 1 and Theorem 2.

Lemma 1 *Under certain regularity conditions, for bounded and almost everywhere differentiable functions $F(\cdot)$, at large n,*

$$\sqrt{n}\left(\int F(\tau)g_{n}\left(\tau|\hat{\eta}\right)d\tau-\int F(\tau)g_{n}\left(\tau|\eta\right)d\tau\right)\overset{d}{\approx}\zeta \tag{21}$$

with ζ is a zero mean random variable with variance $v_{\zeta}^{2}=v_{\eta}^{2}\left(\frac{\partial}{\partial\eta}F(\eta)\right)^{2}$.

Proof Let $\tau=\eta+\tau'\frac{v_{\eta}}{\sqrt{n}}$. Then $d\tau=\frac{v_{\eta}}{\sqrt{n}}d\tau'$ and $g_{n}\left(\tau|\eta\right)d\tau=g_{n}^{st}\left(\tau'|\eta\right)d\tau'$, where $g_{n}^{st}\left(\tau'|\eta\right)=\frac{v_{\eta}}{\sqrt{n}}g_{n}\left(\eta+\tau'\frac{v_{\eta}}{\sqrt{n}}\right)$. Brenner et al. (1982) showed (Theorem 4.3) that under suitable regularity conditions the standardized densities, $g_{n}^{st}(\cdot)$, converge almost surely to the standard normal density, $\phi(\cdot)$,

$$g_{n}^{st}\left(\tau'|\eta\right)=\frac{v_{\eta}}{\sqrt{n}}g_{n}\left(\eta+\tau'\frac{v_{\eta}}{\sqrt{n}}\right)\to\phi(\tau'). \tag{22}$$

The delta method ensures that at a sufficiently large n

$$\sqrt{n}\left(\int F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)g_n^{st}\left(\tau'|\eta\right)d\tau' - \int F\left(\hat{\eta}+\tau'\frac{v_\eta}{\sqrt{n}}\right)g_n^{st}\left(\tau'|\hat{\eta}\right)d\tau'\right) \overset{d}{\approx} \zeta, \tag{23}$$

where ζ is a random variable with zero mean and variance

$$v_\zeta = v_\eta\left(\int \frac{\partial}{\partial\eta}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)g_n^{st}\left(\tau'|\eta\right)d\tau' + \int F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)\frac{\partial}{\partial\eta}g_n^{st}\left(\tau'|\eta\right)d\tau'\right) \tag{24}$$

From (22), $g_n^{st}\left(\tau'|\eta\right)\approx\phi(\tau')$ and due to its asymptotic independence of η, $\frac{\partial}{\partial\eta}g_n^{st}\left(\tau'|\eta\right)\approx 0$. Then, $v_\zeta \approx v_\eta\int\frac{\partial}{\partial\eta}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)\phi\left(\tau'\right)d\tau'$. Further, for regular F except possibly for a set of discontinuity points of zero measure there exists a sufficiently large sample size so that $\frac{\partial}{\partial\eta}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)$ is approximately constant forall τ' except for a subset of neglectable probability mass. The asymptotics

$$\int_{-\infty}^{-n^{2/3}}\frac{\partial}{\partial v_\theta^2}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)\phi\left(\tau'\right)d\tau'\to 0,$$

$$\int_{n^{2/3}}^{\infty}\frac{\partial}{\partial\eta}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)\phi\left(\tau'\right)d\tau'\to 0,$$

and

$$\int_{-n^{2/3}}^{n^{2/3}}\frac{\partial}{\partial\eta}F\left(\eta+\tau'\frac{v_\eta}{\sqrt{n}}\right)\phi\left(\tau'\right)d\tau'\to\frac{\partial}{\partial\eta}F\left(\eta\right),$$

ensure that $v_\zeta^2\approx\left(v_\eta\frac{\partial}{\partial\eta}F\left(\eta\right)\right)^2$. ■

If we apply Lemma 1 and use a power function for F, $F(\eta)=Pr\left(T_N(\eta)>k_N\right)$, then we observe asymptotic $\sqrt{n}$-normality with use of the MLE $\hat{\eta}$ to estimate η. This large sample result can be used in tandem with (20) for asymptotic variance estimation.

Lemma 1 shows that the local asymptotic framework and potential presence of discontinuity points in the integrand of $PWR\left(\alpha,\beta|\mathbf{D},\theta,\eta\right)$,

$$Pr\left(T_N(\eta)>k_N\right)=\sum_{i=n}^{n_{max}}Pr(N=i)Pr\left(T_i(\eta)>k_i\right),$$

do not invalidate applicability of the delta method and lead to the same limiting distribution.

Let $Pr\left[T_N(\eta)>k_N\middle|\theta,\hat{\eta}\right]$ be the naïve estimate of the power curve with a perfect formula for the total sample size, which means $Pr\left[T_N(\eta)>k_N\middle|\theta_0,\eta\right]=\alpha$ and $Pr\left[T_N(\eta)>k_N\middle|\theta_1,\eta\right]=1-\beta$.

Theorem 2 *Under certain regularity conditions the naïve estimate, $Pr\left[T_N(\eta)>k_N\middle|\theta,\hat{\eta}\right]$, and the adjusted estimate, $PWR\left(\alpha,\beta|\mathbf{D},\theta,\hat{\eta}\right)$, of the local power function $PWR\left(\alpha,\beta|\mathbf{D},\theta,\eta\right)$ are asymptotically equivalent.*

Proof Under the assumption that the power function, $Pr\left[T_N(\tau) > k_N\right] (= Pr[\tau])$, is differentiable almost everywhere, Lemma 1 applies and leads to

$$\xi_n \stackrel{d}{=} \sqrt{n}\left(PWR\left(\alpha,\beta|\mathbf{D},\theta,\hat{\eta}\right) - PWR\left(\alpha,\beta|\mathbf{D},\theta,\eta\right)\right) \stackrel{d}{\to} N(0, C^2 v_\eta^2), \tag{25}$$

where $C = \frac{\partial}{\partial\eta} Pr\left[\eta\right]$.

If we use $\hat{\eta}$ naïvely by plugging it into the test statistic (T_N), the sample size formula (N), and the critical value (k), then we estimate the power curve (15) with $Pr\left[\hat{\eta}\right]$. The $Pr\left[\hat{\eta}\right]$ estimates $Pr\left[\eta\right]$, which is different from $PWR\left(\alpha,\beta|\mathbf{D},\theta,\eta\right)$ for finite samples. Denote the systematic error by

$$\Delta_n^\theta(\eta) = PWR\left(\alpha,\beta|\mathbf{D},\theta,\eta\right) - Pr\left[T_N(\eta) > k_N\middle|\theta,\eta\right], \tag{26}$$

where the subscript n reflects the dependence on the first stage sample size. Then $\Delta_n(\eta)$ is a functional non-random sequence converging to zero as $n \to \infty$. Under the assumption that exceptions are treated identically for the naïve and adjusted approaches,

$$\Delta_n^\theta(\eta) = Pr\left(\hat{\eta} \in G_{reg}\right)\left[\int Pr\left[T_N(\tau) > k_N\right] g_n\left(\tau|\theta,\eta\right) d\tau - Pr\left[T_N(\eta) > k_N\right]\right].$$

Then, for a sufficiently large n, $Pr\left(\hat{\eta} \in G_{reg}\right) = 1 + o\left(\frac{1}{\sqrt{n}}\right)$ and

$$\begin{aligned}
\sqrt{n}\Delta_n^\theta(\eta) &= \sqrt{n}\int Pr\left[\tau\right] g_n\left(\tau|\eta\right) d\tau - \sqrt{n} Pr\left[\eta\right] && (27)\\
&= \int \sqrt{n}\left(Pr\left[\eta + \tau'\frac{v_\eta}{\sqrt{n}}\right] - Pr\left[\eta\right]\right) g_n^{st}\left(\tau'|\eta\right) d\tau' && (28)\\
&= \int \sqrt{n}\left(\frac{\partial Pr\left[\eta\right]}{\partial\eta}\tau'\frac{v_\eta}{\sqrt{n}} + \frac{1}{2}\frac{\partial^2 Pr\left[\eta\right]}{\partial\eta^2}\left(\tau'\frac{v_\eta}{\sqrt{n}}\right)^2\right) g_n^{st}\left(\tau'|\eta\right) d\tau' + o\left(\frac{1}{\sqrt{n}}\right) && (29)\\
&= \frac{\partial Pr\left[\eta\right]}{\partial\eta} v_\eta^2 \int \tau' g_n^{st}\left(\tau'|\eta\right) d\tau' + \frac{\partial^2 Pr\left[\eta\right]}{\partial\eta^2}\frac{v_\eta^4}{\sqrt{n}}\int \tau'^2 g_n^{st}\left(\tau'|\eta\right) d\tau' + o\left(\frac{1}{\sqrt{n}}\right) && (30)\\
&= O\left(\frac{1}{\sqrt{n}}\right). && (31)
\end{aligned}$$

The first term in (30) depends on the standardized distribution of the nuisance parameter through $\int \tau' g_n^{st}\left(\tau'|\eta\right) d\tau'$, which converges to $\int \tau'\phi(\tau')d\tau'$ (=0) and the rate of the mean convergence is $\frac{1}{\sqrt{n}}$.

At the values of η where the sample size $N(\eta)$ (discreetly) changes, the presence of discontinuity points in $Pr\left[T_N(\eta) > k_N\middle|\theta,\eta\right]$ is possible. This, however, does not change the applicability of the delta method since discontinuity points are defined on a set of zero measure and the function $Pr\left[T_N(\tau) > k_N\middle|\theta,\tau\right]$ is integrated with respect to the distribution of the nuisance parameter $g_n(\cdot)$. Then,

$$\psi_n \stackrel{d}{=} \sqrt{n}\left(Pr\left[T_N(\hat{\eta}) > k_N\middle|\theta,\hat{\eta}\right] - Pr\left[T_N(\eta) > k_N\middle|\theta,\eta\right]\right) \stackrel{d}{\to} N(0, C^2 v_\eta^2). \tag{32}$$

Table 2: Simulation example to compare different estimates of the true power function; 2000 simulations under H_A and 10000 under H_0

Quantity		At H_0	At H_A
$PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$	(true)	0.029	0.865
$PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$	(proposed estimator)	0.033	0.872
$Pr\left[T_N(\eta) > k_N \middle| \theta, \hat{\eta}\right]$	(naive estimator)	0.038	0.919

Equations (32) and (27-31) ensure

$$\sqrt{n}\left(Pr\left[T_N(\hat{\eta}) > k_N \middle| \theta, \hat{\eta}\right] - PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)\right) \stackrel{d}{=} \psi_n + \sqrt{n}\Delta_n(\eta). \tag{33}$$

Thus, the asymptotic distributions of $Pr\left[T_N(\hat{\eta}) > k_N \middle| \theta, \hat{\eta}\right]$ and $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$ for estimating the power function $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$ are the same, and the systematic bias $\Delta_n = O\left(\frac{1}{n}\right)$. ■

4.3 Power Curve Estimation (simulation example)

Theorem 2 investigates only the importance of integration with respect to the distribution of $\hat{\eta}$ in estimating the true power function.

Here we revisit the simulation example from Section 3 but for estimating the power function without manipulating α and β in the sample size formula. For this example, we consider two estimators with SSR defined by Equation (14).

Table 2 reports type I error ($\theta = \beta_1 = 0$) and power under the alternative ($\theta = \beta_1 = 1.127$). In his table, $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$, shows the type I error and power when we perform interim sample size recalculation targeting 5% type I error and 80% power but ignore the interim SSR. The $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$ and $Pr\left[T_N(\eta) > k_N \middle| \theta, \hat{\eta}\right]$ are two plug-in estimates of $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$.

Then the bias associated with interim sample size recalculation and the use of Wald test is 0.021 (=0.050-0.029) for type I error and -0.065 (=0.800-0.865) for power. This bias is unknown for each particular study, but can be estimated for further correcting the bias like we did in the simulation example of Section 3. If we use $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$, then we estimate the bias as 0.017 (=0.05-0.033) and 0.072 (=0.800-0.872), which is a more accurate assessment of the bias than 0.012 (=0.050-0.038) and 0.119 (0.800-0.919) from $Pr\left[T_N(\eta) > k_N \middle| \theta, \hat{\eta}\right]$.

Even though the asymptotic properties of $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$ and $Pr\left[T_N(\eta) > k_N \middle| \theta, \hat{\eta}\right]$ are the same, Table 2 shows that it is still important to integrate out the distribution of $\hat{\eta}$ when sample size is not large enough. The $PWR(\alpha, \beta|\mathbf{D}, \theta, \hat{\eta})$ is a more accurate estimate of $PWR(\alpha, \beta|\mathbf{D}, \theta, \eta)$ than $Pr\left[T_N(\eta) > k_N \middle| \theta, \hat{\eta}\right]$ under both H_0 and H_A.

5. Simulation studies in t-test settings

Here we consider a few simple simulation scenarios to show how the proposed general methodology works for a two sample t-test with a random group allocation, and compare it with a few other applicable approaches.

Let

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

and

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

where n_{10} and n_{11} are the first stage sample sizes for groups 0 and 1, respectively. With random group allocation we cannot control n_{10} and n_{11}, but we can define the total first stage sample size $n(= n_{10} + n_{11})$. Then $\hat{\eta}_2 = n_{10}/n$ is an estimate of the allocation ratio η_2. Thus, the nuisance parameter $\eta = (\eta_1, \eta_2)$ is two dimensional.

5.1 Designs in simulation studies

Design $\mathbf{D}_{f,2t}\left(\alpha, \beta, \theta_0, \theta_1, \eta^{(0)}\right)$ uses a two dimensional nuisance guess $\eta^{(0)} = \left(\eta_1^{(0)}, \eta_2^{(0)}\right)$ for sample size calculation. Its counterpart $\mathbf{D}_{ip,2t}\left(\alpha, \beta, \theta_0, \theta_1, n, n_{max}\right)$ uses $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$ based on the interim pilot for SSR. We also consider a "restricted" naïve SSR design, $\mathbf{D}_{ipr,2t}$, a variation of $\mathbf{D}_{ip,2t}$, which sets $2n$ as the lower bound for the total sample size. This design employs the Wittes and Brittain (1990) idea of performing the interim analysis at half the originally planned sample size, and allowing only increase in the targeted study size. The use of α_{new} and β_{new} instead of α and β converts $\mathbf{D}_{ip,2t}$ to $\mathbf{D}^a_{ip,2t}$.

In practice, the random aspect of the allocation is usually ignored in the sample size estimation formula and the formula for a fixed allocation is used instead.

Our simulation study in Table 3 considers five study designs:

- $\mathbf{D}_{f,2t}$, a design with a fixed sample size calculation under the assumption that the values of the nuisance parameters are known exactly before the study starts,
- $\mathbf{D}_{ip,2t}$, internal pilot design with naïve SSR,
- $\mathbf{D}_{ipr,2t}$, internal pilot design with restricted naïve sample size recalculation, where the total sample size is at least twice as large as the size of the internal pilot,
- $\mathbf{D}^a_{ip,2t}$, the naïve internal pilot design adjusted by interim resampling.
- $\mathbf{D}^a_{bound,2t}$, the bounding method adjusting α to the most conservative situation. This approach was applied under the assumption of equally sized groups.

Table 3 shows that for the chosen simulation settings the adjusted for internal pilot design outperforms or shows similar properties to the $\mathbf{D}_{ip,2t}$, $\mathbf{D}^a_{bound,2t}$ and $\mathbf{D}_{ipr,2t}$ designs. Compared to $\mathbf{D}_{ip,2t}$, the $\mathbf{D}^a_{ip,2t}$ decreases the type I error inflation and makes its power closer to the targeted value. The adjusted design even corrects for the imperfection of the sample size formula, which results in under-powering for fixed sample size calculations. The bounding method produced the adjusted $\alpha = 0.0428$, which controlled the type 1 error, but was slightly underpowered. For $\eta_2 = 0.25$ there are approximately 5 pilot observations in a

Table 3: Monte-Carlo Type I error, Power, and Sample Sizes; 100,000 simulations; two sample t-test designs; $\theta_0 = 0$ (H_0); $\theta_1 = 1$ (H_1); $n = 20$; $n_{max} = 600$; random allocation; * denotes deviations from the targeted values significant at 5%

η_2	η_1	$\mathbf{D}_{f,2t}$	$\mathbf{D}_{ip,2t}$	$\mathbf{D}_{ipr,2t}$	$\mathbf{D}^a_{bound,2t}$	$\mathbf{D}^a_{ip,2t}$
		Type I error				
0.5	1	0.0496	0.0559*	0.0497	0.0500	0.0525*
0.5	1.5	0.0503	0.0536*	0.0531*	0.0466*	0.0486
0.25	1	0.0501	0.0560*	0.0523*		0.0517
0.25	1.5	0.0500	0.0544*	0.0540*		0.0515
		Power				
0.5	1	0.7673*	0.8042	0.8833*	0.7767*	0.8038
0.5	1.5	0.7919*	0.7940*	0.7958*	0.7579*	0.8012
0.25	1	0.7858*	0.8107*	0.8430*		0.8215*
0.25	1.5	0.7907*	0.7940*	0.7945*		0.8185*
		Average Sample Size (Standard Deviation)				
0.5	1	32(0)	34.7(11.7)	42.6(6.2)	33.7 (10.1)	36.1(12.7)
0.5	1.5	72(0)	76.3(26.1)	76.6(25.6)	72.6 (23.6)	80.2(28.0)
0.25	1	44(0)	48.9(22.2)	52.6(19.0)		56.3(34.0)
0.25	1.5	96(0)	107.0(49.2)	107.2(49.0)		128.3(79.6)

smaller group. This leads to poor estimation of η_2 and increases the power to 82%, however, even in this case the benefits of $\mathbf{D}^a_{ip,2t}$ are clear: the type I error is controlled better than in $\mathbf{D}_{ip,2t}$, whereas a similar quality of type I error control with $\mathbf{D}_{ipr,2t}$ is compensated by a better control of power.

The major benefit of the proposed method is two-fold: (1) it has much wider scope of possible applications since it does not target any specific estimating procedure and (2) no changes to the estimator and inference methods are needed beyond adjusting the cutoff on P-value for determining significance.

6. Example

In this section we show a more complex example of a hypothetical observational study on localized prostate cancer stage.

Localized prostate cancer is associated with substantially better prognosis compared to regionally or distantly extended tumors. We investigate a research question of whether race is associated with cancer stage at detection. Since prostate-specific antigen (PSA) levels and age (AGE) are known predictors of prostate cancer stage we want to adjust for their effect in the analysis of racial differences. The logistic regression analysis is a traditional choice for the analysis of binary outcomes adjusted for potential confounders.

We extracted Black and White subjects with newly diagnosed prostate cancers in 2012 from the Atlanta GA SEER cancer registry. We reduced the dataset to Medicare eligible

Table 4: Logistic regression for predicting localized prostate cancer.

Parameter	Estimate	Std. Error	z value	P-Val	Sample Size
Intercept ($\hat{\beta}_0$)	0.4982	4.1729	0.119	0.9050	100
Black ($\hat{\beta}_1$)	0.7619	0.8393	0.908	0.3640	(Internal
Age ($\hat{\beta}_2$)	0.0518	0.0612	0.846	0.3973	Pilot)
lPSA ($\hat{\beta}_3$)	-0.5814	0.2938	-1.979	0.0478	
Intercept ($\hat{\beta}_0$)	2.7156	1.9149	1.418	0.156	459
Black ($\hat{\beta}_1$)	0.5962	0.4265	1.398	0.162	(naïve
Age ($\hat{\beta}_2$)	0.0412	0.0275	1.498	0.134	Sample
lPSA ($\hat{\beta}_3$)	-0.8832	0.1494	-5.911	<0.0001	Size)
Intercept ($\hat{\beta}_0$)	2.6322	1.6719	1.574	0.1154	662
Black ($\hat{\beta}_1$)	0.8775	0.3736	2.349	0.0188	(Adjusted
Age ($\hat{\beta}_2$)	0.0445	0.0238	1.869	0.0616	Sample
lPSA ($\hat{\beta}_3$)	-0.9155	0.1213	-7.545	<0.0001	Size)
(Intercept) ($\hat{\beta}_0$)	2.0894	0.2968	7.041	<0.0001	20,697
Black ($\hat{\beta}_1$)	0.2915	0.0655	4.452	<0.0001	(Full
Age ($\hat{\beta}_2$)	0.0507	0.0042	12.074	<0.0001	Dataset)
lPSA ($\hat{\beta}_3$)	-0.8671	0.0225	-38.489	<0.0001	

patients (65 years of age and older) with observed PSA and localization status. We excluded patients with unknown PSA value. Finally, our dataset consisted of $20,697$ records.

The model equation for the logistic regression is

$$E\left(L_i|B_i, A_i, PSA_i\right) = \left[1 + \exp\left(-\beta_0 - \beta_1 \cdot B_i - \beta_2 \cdot A_i - \beta_3 \cdot log\left(PSA_i\right)\right)\right]^{-1}, \qquad (34)$$

where for each i^{th} individual L_i and B_i are the indicators of Localized cancer and Black race, A_i and $log\left(PSA_i\right)$ are person's age and natural logarithm of his PSA, respectively.

We formalized the research question into a hypothesis testing problem with $H_0 : \beta_1 = 0$ and wish to design a study that would have 80% power to detect a 3-fold change in adjusted odds of localized cancer between the racial groups. This corresponds to $H_A : \beta_3 = \log(3) \approx 1.1$. To calculate the sample size for this study the sample size formula (14) was used.

The internal pilot is the first 100 records from the dataset described above. The top quarter of Table 4 shows the model fit of the logistic regression model (34) on this internal pilot data (the sample size of 100).

When we use $\alpha = 0.05$, $1 - \beta = 0.8$, and $SE(\beta) = 0.83928$ from Table 4 in the sample size formula (14), the total sample size for the naïve approach is 459. The second quarter of Table 4 shows that the naïve approach does not detect significance of the race variable. The adjustment of α and $1 - \beta$ took 33 hours of personal computer time to reach convergence (see Table 5). In this table column M defines the number of iterations used to estimate the type I error and power at current values of $\hat{\alpha}_{new}$ and $1 - \hat{\beta}_{new}$. Column Δ calculates $(\hat{a} - \alpha)^2 + (\hat{b} - \beta)^2$ to assess when the convergence is achieved. Columns $\hat{\alpha}_{new}$ and $1 - \hat{\beta}_{new}$ calculate new values for α and $1 - \beta$ to be used instead of the original α and $1 - \beta$.

Table 5: Steps of iterative procedure searching for $\hat{\alpha}_{new}$ and $\hat{\beta}_{new}$ (prostate cancer example).

M	Δ	$\hat{a}$	$1-\hat{b}$	$\hat{\alpha}_{new}$	$1-\hat{\beta}_{new}$
10000	0.004928	0.049500	0.729800	0.050505	0.855571
20000	0.001325	0.050600	0.763600	0.049906	0.880034
30000	0.000568	0.050667	0.776167	0.049249	0.894314
40000	0.000217	0.048800	0.785325	0.050459	0.902464
50000	0.000128	0.051860	0.788840	0.048647	0.908317
60000	0.000068	0.048567	0.791883	0.050082	0.912395
70000	0.000024	0.050114	0.795143	0.049968	0.914770
80000	0.000031	0.049263	0.794500	0.050716	0.917386
90000	0.000037	0.050200	0.793922	0.050514	0.920188
100000	0.000001	0.050320	0.799260	0.050193	0.920527

When we use the adjusted values $\alpha_{new} = 0.050193$ and $1-\beta_{new} = 0.920527$, and $SE(\hat{\beta}_1) = 0.83928$ (internal pilot section from Table 4) in the sample size formula (14), the total sample size for the adjusted approach is 662. The third quarter of Table 4 shows that the adjusted approach concludes statistical significance of the race variable ($P-value < 0.050193$). Additional 203 observations did make a difference in our conclusion. The artificial nature of this illustrative example allows us to fit the logistic regression to the whole dataset shown in the last quarter of Table 4.

This example mimics a chart review study when an investigator does not know how many charts he or she has to review and has no pilot data to start with. In contrast to Section 3 there is almost no bias in the type I error. In that example a smaller sample size contributed to a serious deflation of the type I error, which is reflection of asymptotic nature of Wald tests for regression coefficients in multiple logistic regressions. We should also emphasize that truly exact tests for multiple logistic regression do not exist, and what is currently called the "exact test" is only *conditionally* exact. This example also demonstrates that our approach not only can correct for overpowering observed in Section 3, but also for possible underpowering.

7. Discussion

In this manuscript we considered interim sample size recalculation from the prospective of design resampling based on information available at the interim analysis. The proposed approach is very flexible, and can be applied to a wide variety of situations, including observational and retrospective studies. We defined a study design as a set of rules for sample size calculation/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, study designs should enable researchers to make unambiguous decisions for all possible realizations of random variables.

The use of internal design resampling reduces the dependence of the study design on nuisance parameters and lowers the bias of type I and II errors. We adjust 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 and for final decision making since the final P-value is to be compared versus α_{new} not α.

This final comparison against α_{new} may be challenging for clinicians who often only want to know how much more data they need. Thus, we urge statisticians who apply the design resampling to be careful in emphasizing this aspect of final decision making. Since some practitioners, e.g. clinicians, may still express skepticism in changing the cutoff for significance, we suggest using a multiplicative factor for P-value adjustment (the ratio of α to α_{new}) which serves the same purpose but will allow clinicians to put adjusted P-values in their manuscripts and to make traditional decisions by comparing P-values with the original α.

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 and bring additional uncertainty into the adjustment.

The internal resampling can be computationally challenging especially when a Monte-Carlo study is used. To speed up the Monte-Carlo simulations we used *C* and *C++* code pre-compiled into a dynamic link library for internal resampling part, and *R* code for the rest of the program.

Our Monte-Carlo simulation studies show that the proposed resampling methodology leads to a generally better control of type I and II error than the naïve internal pilot design across all considered scenarios. For the two sample t-test with random allocation we decreased inflation of the type I error while approximately controlling for statistical power. Stronger benefits are observed for interim SSR for logistic regression where no competing methods are available. Both the type I error and power are better controlled. This led to the averaged sample size decrease under both the null and alternative hypotheses.

Our approach is different from previously suggested ones, except for the bounding method Coffey et al. (2007), since we are not changing the original statistical procedure and its formula for sample size calculation. We only change α and β to α_{new} and β_{new} at the interim analysis to better control the desired type I and II errors when the final decision is made. This approach allows to write wrapper functions for previously developed statistical methods without changing the internals of the statistical methods.

The internal design resampling requires that the internal pilot sample size and the largest possible sample sizes are defined in the study protocol. 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 problematic.

In practice, the suggested sample size re-estimation is applicable to nonlinear models with multiple nuisance parameters often seen in observational studies. Even though the illustrative part of the manuscript is mainly focused on logistic regression, this approach can be applied to many parametric models where the estimators of the parameter of interest and nuisance parameters are asymptotically normal. Generalized linear models are the direct application area for the suggested approach. Our approach can also be applied to more complex regression models with asymptotically normal regression coefficients. We expect that observational studies will benefit the most from the suggested design resam-

pling methodology, as they have a large member of uncontrolled nuisance parameters and complicated analyses.

Overall, we recommend researchers to clearly define their designs for all possible situations and incorporate interim sample size recalculation with the suggested adjustments to α and β in their study designs. This optimizes the use of their resources, provides better control type I and II errors, and protects against misspecification of nuisance parameters.

References

Betensky, R. and Tierney, C. (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine*, 16:2587–2598.

Brenner, D., Fraser, D. A., and McDunnough, P. (1982). On asymptotic normality of likelihood and conditional analysis. *Canadian Journal of Statistics*, 10(3):163–172.

Coffey, C. S., Kairalla, J. A., and Muller, K. E. (2007). Practical Methods for Bounding Type I Error Rate with an Internal Pilot Design. *Communications in Statistics - Theory and Methods*, 36(11):2143–2157.

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

Coffey, C. S. and Muller, K. E. (2003). Properties of internal pilots with the univariate approach to repeated measures. *Statistics in Medicine*, 22(15):2469–85.

Friede, T. and Kieser, M. (2006). Sample Size Recalculation in Internal Pilot Study Designs: A Review. *Biometrical Journal*, 48(4):537–555.

Friede, T. and Kieser, M. (2011). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics*, 10(1):8–13.

Gurka, M. J., Coffey, C. S., and Gurka, K. K. (2010). Internal pilots for observational studies. *Biometrical journal. Biometrische Zeitschrift*, 52(5):590–603.

Kairalla, J. A., Coffey, C. S., and Muller, K. E. (2008). Glumip 2.0: Sas/iml software for planning internal pilots. *Journal of Statistical Software*, 28:1–32.

Lane, A. and Flournoy, N. (2012). Two-stage adaptive optimal design with fixed first-stage sample size. *Journal of Probability and Statistics*, 2012.

Lane, A., Yao, P., and Flournoy, N. (2014). Information in a two-stage adaptive optimal design. *Journal of Statistical Planning and Inference*, 144(1):173–187.

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

Stein, C. (1945). A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

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