Published in final edited form as:

Stat Med. 2017 December 30; 36(30): 4765–4776. doi:10.1002/sim.7454.

A Model-Based Conditional Power Assessment for Decision Making in Randomized Controlled Trial Studies

Baiming Zou¹, Jianwen Cai¹, Gary G. Koch², Haibo Zhou², and Fei Zou¹

¹Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA

²Department of Biostatistics, University of North Carolina - Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

Conditional power based on summary statistic by comparing outcomes (such as the sample mean) directly between two groups is a convenient tool for decision making in randomized controlled trial studies. In this paper, we extend the traditional summary statistic-based conditional power with a general model-based assessment strategy, where the test statistic is based on a regression model. Asymptotic relationships between parameter estimates based on the observed interim data and final unobserved data are established, from which we develop an analytic model-based conditional power assessment for both Gaussian and non-Gaussian data. The model-based strategy is not only flexible in handling baseline covariates and more powerful in detecting the treatment effects compared with the conventional method, but also more robust in controlling the overall type I error under certain missing data mechanisms. The performance of the proposed method is evaluated by extensive simulation studies and illustrated with an application to a clinical study.

Keywords

Conditional Power; Consistency; Maximum Likelihood Estimate; Multivariate Normal; Nonlinear Data; Randomized Controlled Trial

1. Introduction

Even though randomized controlled trial (RCT) designs are frequently used in various clinical studies for scientific research, in practice, these study designs are often based on limited prior study information. As a result, researchers could face a dilemma near the end of study for the final experimental results only being suggestive with a relatively small p -value (say $0.05 < p < 0.15$) rather than statistically significant (say $p < 0.05$). Under this situation, researchers may want to extend their study to collect additional information in order to increase power for the final study to be significant, based on a priori plan with a priori allocated type I error. Similarly if the final result is not highly significant (say $p < 0.025$) and the originally planned sample size is not large which is not uncommon in many pilot studies, researchers would be interested in extending their study to strengthen the scientific findings

*Correspondence to: Baiming Zou, Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA.

¹bzou@phhp.ufl.edu

as well. Furthermore, in practical executions of many clinical studies, the accrual rate is usually much slower than expected and far below the planned sample size is recruited at the end of study period. Under this scenario, researchers may consider to extend the study to collect additional information based on the data collected thus far.

In a recent type 1 diabetes (T1D) study (i.e. [1]), the efficacy of a combination therapy of anti-thymocyte globulin (thymoglobulin) and pegylated GCSF (neulasta) in maintaining C-peptide production in the established T1D patients was investigated with 25 participants randomized to two groups (17 in test drug and 8 in control). The study showed a marginally significant ($p \approx 0.05$) treatment difference near its planned completion. Since this pilot study is for the feasibility research, investigators are interested in extending the study to consolidate their scientific findings. Although extending the study can provide more convincing and robust scientific findings, a subsequent question is how much more information is needed given the current observed data. Conditional power ([2]; [3]) offers researchers a convenient tool to accomplish this goal. It is often used in clinical research for early stopping of a study due to futility or for extending the study due to its potential efficacy ([4]; [5]).

Conditional power is defined as the conditional probability for the final test statistic to be significant conditional on the observed data, with some speculations on the trend of future unobserved data. In the conventional assessment of conditional power, the test statistic is based on the summary statistic by comparing the outcomes directly ([3]; [6]), for example the sample mean, between two groups. The test statistic based on the summary statistic immediately leads to a straightforward decomposition of final test statistic into two independent parts: one due to the observed interim data and the other from the future, unobserved data. Therefore, the conditional distribution of final test statistic given the observed data can be obtained easily, and the conditional power can be conveniently assessed with an analytic form. Furthermore, under well designed randomization process, systematic differences for the baseline covariates are eliminated. Thus, the unbiased treatment effects estimate can be easily obtained by comparing the outcomes from the two groups directly. This guarantees the validity of the traditional conditional power assessment. A detailed review of conditional power methods can be found in ([6]). Under two-stage design, conditional power is often used for sample size adjustment in different clinical studies. To avoid the redundancy, we refer readers for the review by Shih ([7]) in this regard.

Although the summary statistic-based method is valid in general and easy to use, it is not necessarily efficient. It has the mathematical convenience but only makes use of partial information without taking into account other potentially valuable information embedded in the observed data. In randomized controlled trial studies, in addition to the treatment effects, there often exist other factors affecting the outcomes, i.e. the so-called prognostic factors. Ignoring prognostic factors may not cause bias in the treatment effects estimate for normally distributed outcomes, but it often results in reduced efficiency ([8]; [9]). Furthermore, in the scenario of missing data, which is not uncommon in many clinical studies, or imbalanced baseline prognostic factors, the treatment effects estimate based on the summary statistic could be severely biased ([10]). In addition, for some data types following nonlinear models (binary data), the treatment effects estimate by comparing the log-odds ratio between two

groups with prognostic factors omitted can be biased even if these factors are perfectly balanced by the randomization procedure ([11]; [12]). This biased treatment effects estimate could lead to inappropriate conditional power assessment and erroneous decision making in clinical research.

Motivated by example above and desiring to allow the conditional power to be more broadly applicable in practical scientific studies, we extend the conventional assessment procedure of conditional power with a general model-based framework to utilize other valuable information contained in the observed data. Specifically, we construct the test statistic based on the maximum likelihood estimate (MLE) of the treatment effects from a regression model which includes such information from important prognostic factors. Further, if these factors are imbalanced because of missing data, including them in the model fit would potentially reduce bias in the conditional power assessment as we demonstrated in simulation study (see Section 3). Through an asymptotic relationship between the MLEs of treatment effects from the observed data and final observations (including the observed and future unobserved data), we establish a powerful and convenient way to assess the conditional power analytically for both Gaussian and non-Gaussian data. Extensive simulation results reveal that the proposed framework consistently outperforms the existing conditional power assessment.

The rest of the paper is arranged as follows. We present a detailed mathematical description of model-based conditional power assessment in Section 2. Asymptotic relationships of MLEs based on observed interim data and final unobserved data are outlined in Section 2 while detailed derivations are given in Appendix. Extensive simulation results under various settings are presented in Section 3. The practical usage of model-based conditional power is illustrated in Section 4 with an application to a diabetes clinical study. Discussions can be found in Section 5.

2. Model-Based Conditional Power Assessment Scheme

To fix the notation, let $(y_i, z_i, \mathbf{x}_i)$ be the observed response, binary treatment assignment status (1 for treated and 0 for control) and baseline covariates respectively of subject i . Further, we let the first n_1 and n_0 subjects from the treated and control groups be the observed samples while the remaining $N_1 - n_1$ and $N_0 - n_0$ subjects be future unobserved samples from the treated and control group, respectively. In the sequel, we assume observations are independently and identically distributed (IID). Here *obs* and *all* are used to differentiate variables from the observed and final (including both observed and unobserved) samples. To assess the treatment effects, we consider the following generalized linear model:

$$E(y_i | (z_i, \mathbf{x}_i)) = g^{-1}(\beta_0 + \beta_z z_i + \mathbf{x}_i \beta_{\mathbf{x}}). \quad (1)$$

where g is a given link function. We test the treatment effects with the following hypotheses

$$H_0: \beta_z = 0 \text{ vs } H_a: \beta_z \neq 0.$$

Summary Statistic-Based Conditional Power

In practice, for the randomized controlled trial data, such as continuous or binary outcomes, it is common to test the treatment effects with the following summary statistic based on the observed samples as $\xi_{obs} = \sqrt{\frac{n_0 n_1}{n_0 + n_1}} \frac{\bar{y}_1 - \bar{y}_0}{\sigma}$ with $\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1,i}$ and $\bar{y}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{0,i}$. The test statistic based on the final samples can be similarly constructed as

$\xi_{all} = \sqrt{\frac{N_0 N_1}{N_0 + N_1}} \frac{\tilde{y}_1 - \tilde{y}_0}{\sigma}$ with $\tilde{y}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} y_{1,i}$, $\tilde{y}_0 = \frac{1}{N_0} \sum_{i=1}^{N_0} y_{0,i}$ and $\sigma^2 = \text{Var}(y_{0,i}) = \text{Var}(y_{1,i})$, is assumed as known or available through a consistent estimate. Conventionally, for a given rejection threshold r , the conditional power is defined as $CP_{\theta}(\xi_{obs}) = \Pr(|\xi_{all}| > r \mid \xi_{obs}; \theta)$ where θ represents the parameters associated with the data, including β_z and σ^2 .

Model-Based Conditional Power

Denote the maximum likelihood estimate (MLE) of β_z from model (1) based on the observed data and the final data as $\hat{\beta}_{z,obs}$ and $\hat{\beta}_{z,all}$, respectively. Under the general regularity conditions, $\hat{\beta}_{z,obs}$ and $\hat{\beta}_{z,all}$ follow normal distributions $N(\beta_z, \sigma_{obs}^2)$, and $N(\beta_z, \sigma_{all}^2)$ respectively, with corresponding asymptotic variances σ_{obs}^2 and σ_{all}^2 . Furthermore, as shown in [13], under the assumptions that the observed and unobserved data are independently and identically distributed (IID), the joint distribution of $\hat{\beta}_{z,obs}$ and $\hat{\beta}_{z,all}$ is bivariate normal as follows:

$$\begin{pmatrix} \hat{\beta}_{z,obs} \\ \hat{\beta}_{z,all} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \beta_z \\ \beta_z \end{pmatrix}, \Sigma\right)$$

where $\Sigma = \begin{pmatrix} \sigma_{obs}^2 & \rho \sigma_{obs} \sigma_{all} \\ \rho \sigma_{obs} \sigma_{all} & \sigma_{all}^2 \end{pmatrix}$, $\rho \equiv \lim_{N,n \rightarrow \infty} \text{Corr}(\hat{\beta}_{z,all}, \hat{\beta}_{z,obs})$ and

$\frac{\sigma_{all}^2}{\sigma_{obs}^2} \equiv \lim_{N,n \rightarrow \infty} \frac{\text{Var}(\hat{\beta}_{z,all})}{\text{Var}(\hat{\beta}_{z,obs})}$ with $N = N_0 + N_1$, $n = n_0 + n_1$. Therefore, the conditional

distribution of $\hat{\beta}_{z,all}$ given $\hat{\beta}_{z,obs}$ follows $N(\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}} (\hat{\beta}_{z,obs} - \beta_z), \sigma_{all}^2 (1 - \rho^2))$. On this basis,

we construct the model-based final test statistic as: $\eta_{all} \equiv \frac{\hat{\beta}_{z,all}}{\sigma_{all}}$, for the determination of conditional power, where the conditional distribution of η_{all} given $\hat{\beta}_{z,obs}$ follows $N(\delta, 1)$

with $\delta = \frac{\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}} (\hat{\beta}_{z,obs} - \beta_z)}{\sigma_{all} \sqrt{1 - \rho^2}}$.

Following some mathematical derivation procedures, we can write the model-based conditional power as:

$$\begin{aligned}
 CP_{\beta_z} &\equiv Pr(|\eta_{all}| > r | \hat{\beta}_{z,obs}; \beta_z) \\
 &= 1 - \Phi\left(r - \frac{\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}}(\hat{\beta}_{z,obs} - \beta_z)}{\sigma_{all}\sqrt{1-\rho^2}}\right) + \Phi\left(-r - \frac{\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}}(\hat{\beta}_{z,obs} - \beta_z)}{\sigma_{all}\sqrt{1-\rho^2}}\right) \\
 &= 1 - \Phi\left(r - \frac{\beta_z + \rho \sigma_{all} \xi_{obs}}{\sigma_{all}\sqrt{1-\rho^2}}\right) + \Phi\left(-r - \frac{\beta_z + \rho \sigma_{all} \xi_{obs}}{\sigma_{all}\sqrt{1-\rho^2}}\right) \\
 &= CP_{\beta_z}(\xi_{obs}) \text{ where } \xi_{obs} \equiv \frac{\hat{\beta}_{z,obs} - \beta_z}{\sigma_{obs}} \sim N(0, 1).
 \end{aligned}$$

To assess the model-based conditional power, we need the rejection threshold r . When the conditional power is used as a criterion for decision making in the interim analysis, r is pre-specified. When the conditional power is used for extending an RCT, the rejection region is determined by the overall type I error that investigators would like to control. This will be discussed in details as presented in the sequel.

Controlling Overall Type I Error—It is well known that the type I error rate will be inflated with the increase of sample size. Various solutions under different study settings have been proposed to control the overall type I error properly with varying sample size ([14]; [15]; [16]). Here, we adopt a two-stage sampling scheme [15] for our setting by regarding the extension of an RCT study as a two-stage sampling procedure with three scenarios. Specifically, if the observed test statistic is highly significant, i.e. $\frac{|\hat{\beta}_{z,obs}|}{\sigma_{obs}} > r_u$, then we would not need to extend the study but reject the null. On the contrary, if the observed test statistic is clearly non-significant, i.e. $\frac{|\hat{\beta}_{z,obs}|}{\sigma_{obs}} < r_l$ (the p-value is greater than a pre-specified value), then it is not worth extending the study and the null is not rejected. The study would be extended only if the observed test statistic falls in between r_l and r_u . That is the overall type I error for the potential extension of a study can be split into two parts and written as:

$$\alpha = Pr\left(\frac{|\hat{\beta}_{z,obs}|}{\sigma_{obs}} > r_u | H_0\right) + Pr\left(|\eta_{all}| > r; r_l < \frac{|\hat{\beta}_{z,obs}|}{\sigma_{obs}} < r_u | H_0\right) \quad (2)$$

The first part can be regarded as the type I error by rejecting the null based on the observed data, which is similar to the α -spending function in the group sequential clinical trial [17]. Practically, it is the maximum p-value, denoted as p_u , the researcher would use to reject the null based on the observed interim data without the need to continue or extend the study.

The second part is the type I error resulting from extending the study. Under H_0 ,

$\xi_{obs} \equiv \frac{\hat{\beta}_{z,obs}}{\sigma_{obs}} \sim N(0, 1)$. Furthermore, under the null, the conditional distribution of η_{all} given

ξ_{obs} follows $N(\delta_0, 1)$ where $\delta_0 = \xi_{obs} \frac{\rho}{\sqrt{1-\rho^2}}$. Hence, under the null and conditional on ξ_{obs} ,

$\eta_{all} - \delta_0$ follows a standard normal distribution. Therefore, we rewrite equation (2) as follows:

$$\alpha = p_u + \int_{r_l}^{r_u} \left[1 - \Phi\left(r - \frac{\rho \xi_{obs}}{\sqrt{1-\rho^2}}\right) + \Phi\left(-r + \frac{\rho \xi_{obs}}{\sqrt{1-\rho^2}}\right) \right] \phi(\xi_{obs}) d\xi_{obs} \quad (3)$$

where $r_u = z \frac{p_u}{1 - \frac{p_u}{2}}$, $r_l = z \frac{p_l}{1 - \frac{p_l}{2}}$, p_u and p_l are predefined.

The conditional power with the overall type I error controlled at α is therefore

$$1 - \Phi\left(r - \frac{\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}} (\hat{\beta}_{z,obs} - \beta_z)}{\sigma_{all} \sqrt{1-\rho^2}}\right) + \Phi\left(-r - \frac{\beta_z + \rho \frac{\sigma_{all}}{\sigma_{obs}} (\hat{\beta}_{z,obs} - \beta_z)}{\sigma_{all} \sqrt{1-\rho^2}}\right). \quad (4)$$

To assess the conditional power with (4), we need to know parameters σ_{obs} , σ_{all} and ρ given $\hat{\beta}_{z,obs}$ in addition to the treatment effects to be detected, β_z . In practice, β_z can be pre-determined from other existing studies or based on observed data thus far as in our real study example in Section 4. Empirically, parameter σ_{obs} can be estimated from model (1) using the observed data. The estimation of the parameters σ_{all} and ρ , however, is not straightforward since they involve the future unobserved data. Under our setting, i.e. the ratios of treated vs control samples in the observed data and future unobserved data are controlled to be asymptotically the same (i.e. $\frac{n_0}{n_1} \approx \frac{N_0 - n_0}{N_1 - n_1} \approx \frac{N_0}{N_1}$), we can further show (in Appendix) that

$$\rho \equiv \lim_{N, n \rightarrow \infty} \text{Corr}(\hat{\beta}_{z,all}, \hat{\beta}_{z,obs}) = \sqrt{\tau} \quad (5)$$

$$\frac{\sigma_{all}^2}{\sigma_{obs}^2} \equiv \lim_{N, n \rightarrow \infty} \frac{\text{Var}(\hat{\beta}_{z,all})}{\text{Var}(\hat{\beta}_{z,obs})} = \tau \quad (6)$$

where $N = N_0 + N_{1,n} = n_0 + n_1$ and $\tau \equiv \lim_{N, n \rightarrow \infty} \frac{n}{N}$ is the asymptotic proportion of the observed sample size relative to the final overall sample size. Based on it, we propose a finite sample estimate of σ_{all}^2 as $\hat{\sigma}_{all}^2 = \frac{n}{N} \hat{\sigma}_{obs}^2$ and estimate ρ , i.e. the asymptotic correlation between the treatment effects estimates based on all and observed data, as $\hat{\rho} = \sqrt{\frac{n}{N}}$ and τ by $\hat{\tau} = \frac{n}{N}$. As demonstrated in the numerical studies in Section 3, these parameter estimates have very good finite sample performances.

In summary, for final sample size requirement to extend an RCT based on the observed interim data, the following are the steps to use the proposed model-based conditional power formulas to calculate the final sample size N with a pre-specified conditional power $1 - \beta$ by controlling the overall type I error at a significance level α :

1. Fit model (1) with the observed data to obtain $\hat{\beta}_{L,obs}$ and $\hat{\sigma}_{obs}$.
2. By the observed number of samples n thus far, attempt a small final sample size $N(>n)$ and estimate τ as $\hat{\tau} = \frac{n}{N}$.
3. Substitute ρ and ρ^2 in equation (3) with $\sqrt{\hat{\tau}}$ and $\hat{\tau}$, respectively, to obtain the numerical solution of r as the rejection threshold such that the targeted overall type I error α is controlled.
4. Plug in parameter estimates from steps 1), 2) and 3) into equation (4) to calculate the conditional power denoted as p_c . If $p_c < 1 - \beta$, then increase N , otherwise, decrease N and repeat steps 2), 3) and 4) until the appropriate N is obtained.

It should be emphasized, in the proposed procedure, n is the actual observed sample size thus far and it is not pre-specified. But p_u (or r_u) and p_l (or r_l) are design parameters and they need to be pre-specified. In practice, a more likely application of interest would be to calculate the conditional power given the final sample size N , and we do not need to bother to adjust the final sample size N to reach the pre-specified conditional power in step 4 as noted above.

3. Simulation Studies

To compare the performance of the proposed model-based conditional power assessment with the conventional summary statistic method, we conduct simulation studies for two different data types: the continuous normal data and the binary outcomes.

3.1. Simulation Studies for Gaussian Data

Before comparing the performance of the model-based strategy versus the summary statistic-based method, we investigate the finite sample performance for the proposed parameter estimators (5) and (6) for the model-based conditional power assessment. Specifically, we perform the first set of simulations with the following data generating model to mimic the clinical trial data structure where there exists another prognostic factor x influencing the outcome in addition to the treatment effects:

$$y_i = 0.2 + \beta_z z_i + \beta_x x_i + \epsilon_i \quad (7)$$

where $\beta_x = 0.8$, $x_i \sim N(0, 1)$ and $\epsilon_i \sim N(0, 1)$. We fixed the final sample size of each arm at 100 and varied the proportion of data observed from 0.1 to 0.9 by randomly selecting the corresponding proportion of data and fitting model (1) with an identity link using all and the available observations to get the corresponding treatment effects estimate with data generated from both the alternative (i.e. $\beta_z = 0.35$) and null (i.e. $\beta_z = 0$) models. Both the variances for the treatment effects estimate based on all and interim observations and the correlation between these two treatment effect estimates are all based on 1000 Monte Carlo replications and thus they can be regarded as the true sample variances and correlation. Results for this simulation setting are depicted in Figure 1.

Investigating Figure 1, it appears that the ratio between the variance of the treatment effects estimate based on the interim data and that based on all samples, i.e. $\frac{\hat{\sigma}_{all}^2}{\hat{\sigma}_{obs}^2}$, changes with the

proportion of observed samples, i.e. $\hat{\tau}$, almost fits on a 45 degree straight line. The same observation applies to the relationship for the square of correlation between the treatment effects estimate based on the interim data and that based on all samples, i.e.

$\hat{\rho} = \widehat{Corr}(\hat{\beta}_{z,obs}, \hat{\beta}_{z,all})$, versus the observed sample proportion. That is $\frac{\hat{\sigma}_{all}^2}{\hat{\sigma}_{obs}^2} \propto \hat{\tau}$ and $\hat{\rho}^2 \propto \hat{\tau}$.

Simulation results clearly demonstrate the very good finite sample performance of the parameter estimators based on (5) and (6) regardless data generated from the null or alternative models.

To evaluate the performance of the proposed model-based strategy relative to the conventional summary statistic method, we conduct simulation studies with two scenarios: with and without missing data. Similar observations for the performance were obtained under different design parameter settings. For ease of presentations, we present results for the study design parameters with $p_u = 0.025$, $p_l = 0.15$ and the theoretical overall type I error $\alpha = 0.05$, to mimic some practical pilot clinical trials where the originally planned sample size is usually small. The data generating model is the same as model (7) used in the first set of simulations. In addition to comparing the performance in the metric of conditional power, we also evaluate the empirical performance for controlling the overall type I error under the two-stage sampling scheme (2) for the proposed strategy relative to the conventional method with the data generated from model (7) except that we set $\beta_z = 0.0$, i.e. data are generated from the null model. The overall empirical type I error is assessed by the two-stage sampling scheme (2). That is, if the observed test statistic $\xi_{obs} \equiv \frac{\hat{\beta}_{z,obs}}{\sigma_{obs}}$ satisfies $|\xi_{obs}| > z_{1-\frac{p_u}{2}}$ or

$z_{1-\frac{p_l}{2}} < |\xi_{obs}| < z_{1-\frac{p_u}{2}}$ with the final test statistic $\eta_{all} \equiv \frac{\hat{\beta}_{z,all}}{\sigma_{all}}$ satisfying $|\eta_{all}| > r$, then a type

I error has occurred. For the conventional summary statistic-based method, the overall

empirical type I error is evaluated in the same manner except that the test statistic is based on the summary statistic.

No Missing Data Scenario—In this set of simulation studies, the outcomes are generated via model (7) in an ideal scenario without any missing data. We conduct extensive simulation studies under no missing data setting, for example, different observed sample sizes and different effect size, similar observations on the performance of the model-based strategy versus the summary statistic method are obtained. For ease of presentation, we present the simulation results for the scenario with the observed sample size for each arm fixed at $n = 20$ under various targeted final sample sizes to mimic practical pilot clinical studies where the initially planned sample sizes are usually small. For each setting, the total number of simulations is run for 5,000 Monte Carlo replications. Results under this setting are summarized in the first part of Table 1.

Results from the first part of Table 1 reveal the remarkable conditional power gains for the model-based strategy over the summary statistic method while controlling the overall type I error at the same level. For instance, when the observed sample size in each arm is 20 and the final sample size in each arm to be collected is 50, the average conditional powers are 0.635 versus 0.498 for the model-based strategy and the summary statistic method, respectively. Simulation results clearly demonstrate the benefit of model-based strategy in practical applications.

To investigate the performance of model-based conditional power when the model used to fit the data is not a true model, we generate the data with a quadratic term of x added to model (7) as follows but it is omitted in data fit for model-based conditional power:

$$y_i = 0.2 + \beta_z z_i + \beta_x x_i + 0.2x_i^2 + \epsilon_i$$

Results for this setting is summarized in the second part of Table 1. Based on it, we have similar observations as the first part, i.e. the overall empirical type I error rates of both methods are controlled to the targeted level but the proposed method is still more powerful than the conventional method. Compared with the first part, the conditional powers from both methods are lowered since the model used to fit data is not an optimal model and the treatment effect estimator is not efficient though it is still an unbiased estimator.

To further investigate the performance of model-based conditional power for model misspecification scenario, we generate the data with a quadratic term of x only as follows but the data is fit with the linear term of x for model-based conditional power assessment:

$$y_i = 0.2 + \beta_z z_i + \beta_x x_i^2 + \epsilon_i$$

Results based on this setting is presented in the third part of Table 1. Inspecting this part, we obtain similar observations as the first two parts, i.e. the overall empirical type I error rates

of both methods are still controlled to the targeted level quite well. Compared with the first part, the conditional powers from both methods are remarkably lowered but they are comparable indicating the robustness of model-based method.

To investigate the performance of model-based conditional power under skewed error distribution, we generate the data via the following model similar to (7) except that the error term $\epsilon_i \sim \exp(1)$ instead of $\sim \mathcal{N}(0, 1)$:

$$y_i = 0.2 + \beta_z z_i + \beta_x x_i + \epsilon_i$$

We summarize the results for this skewed error distribution setting in the fourth part of Table 1. Based on it, we have similar observations as the first part, i.e. the overall empirical type I error rates of both methods are controlled closely to the targeted level but the proposed method is still more powerful than the conventional method indicating that skewed error distribution has ignorable impact on the performance of both methods.

Missing Data Scenario—In many clinical studies, missing data often occur for various reasons. For example, older patients in the treated group may miss endpoint measurements more often due to adverse events while younger patients in the control group could miss the measurements more frequently due to unsatisfactory benefits. To mimic this missing scenario (i.e. missing completely at random for our proposed method based on the terminology of Rubin [18]), we generate data based on the following missing mechanism settings to investigate the performance of the proposed framework. In this set of simulations, data are generated from the same model as in Table 1 with 60 subjects randomized to each arm and the targeted sample sizes are set at 80, 100, and 120 in each arm, respectively. However, in order to reflect the dynamics of missing data, we set missing status with unequal missing rates for the two arms. Specifically, the missing mechanism is set as 15% of those with $x_T > 0$ for the treated group and 10% of those with $x_C < 0$ in the control group (Missing Data Scenario I). The observations with missing data are excluded from the assessment of conditional power for both summary statistics and model-based methods. Simulation results for this missing data specification are presented in the first part of Table 2. As a comparison, the missing mechanism is reversed as the following: in the treated arm 10% of those with $x_T < 0$ are randomly specified as missing while 15% of those with $x_C > 0$ in the control arm are randomly marked as missing (Missing Data Scenario II). Results for this simulation setting are given in the second part of Table 2.

Checking the results from the first part of Table 2, the summary statistic conditional power is evidently lower than the model-based conditional power as we observed in the previous simulation studies. More importantly, the conventional scheme fails to control the overall type I error when there are missing data. In contrast, the proposed model-based strategy consistently controls the overall type I error at the targeted level. For instance, when the targeted final sample size is (85, 90) with actual observed interim sample size (51, 54) for the treated and control arm respectively, the empirical overall type I error via the conventional scheme is 0.144 while it is 0.047 for the model-based strategy which indicates that the conventional scheme here falsely rejects null more than three times more often than it

should. Furthermore, examining the first part of Table 2 closely, it demonstrates that the treatment effects estimate based on the available observations via the summary statistic method is severely biased downward while the treatment effects estimate from the model-based strategy is approximately unbiased. The biased treatment effects estimate leads to the invalid conditional power evaluation and failure for controlling the overall type I error under the summary statistic scheme. Inspecting the second part of Table 2, we notice that the conditional power of the conventional scheme is very close to that from the model-based method. However, this conclusion is misleading if we investigate the treatment effects estimate from the summary statistic, which is now severely biased upward resulting in an erroneous conditional power calculation in addition to the severe inflation of the type I error rate. In contrast, the treatment effects estimate from the model-based method is approximately identical to the true effect size and the empirical overall type I error is consistently controlled to the targeted theoretical level, and these results further demonstrate the robustness of the model-based conditional power strategy and its usefulness in clinical studies.

3.2. Simulation Studies for Binary Outcomes

Similar to the simulation studies for Gaussian data, we investigate the finite sample performance of the model-based conditional power for binary outcomes with data generated from the following logistic model:

$$\log \text{it}[Pr(y_i = 1 | z_i, x_i)] = 0.2 + \beta_z z_i + \beta_x x_i \quad (8)$$

where $\beta_x = 2.0$ and $x_i \sim N(0, 1)$. We simulate data from the alternative model with $\beta_z = 0.8$ and the null model with $\beta_z = 0$. We fix the final sample size in each arm as 300 and varied the proportion of interim data observed from 0.1 to 0.9 by randomly selecting the corresponding proportion of data as observed. The variance estimations of the treatment effects estimates based on all and available observations are calculated from total 1000 Monte Carlo replications, and results are presented in Figure 2. Again, the variance ratio and the square of correlation change with the observed sample proportion all fit on a 45 degree

straight line, i.e. $\frac{\hat{\sigma}_{all}^2}{\hat{\sigma}_{obs}^2} \propto \hat{\tau}$ and $\hat{\rho}^2 \propto \hat{\tau}$, demonstrating the good finite sample performances of

the proposed parameter estimators under the binary outcomes scenario.

Table 3 summarizes the conditional power and empirical overall type I error with total 5000 simulations under each setup based on the aforementioned data generating model (8).

Results from Table 3 reveal that the model-based conditional power strategy consistently outperforms the summary statistic conditional power method while the overall type I error is controlled at the same level. For instance, when the observed interim sample size is 80 and the final targeted sample size is set to 160 the conditional power based on the conventional method and the proposed approach are 0.530 versus 0.915. This set of numerical studies

further demonstrates the advantages of the proposed method in practical clinical research for binary outcomes.

To further investigate the performance of model-based conditional power for binary data with missing, we use the same missing data generating mechanism as Gaussian data based on model (8). Results for this setting are summarized as follows in Table 4.

Inspecting Table 4, it evidently reveals that the empirical overall type I error is consistently controlled to the targeted theoretical level closely by the model-based method regardless of which missing mechanism is used, while the type I error is remarkably inflated by the conventional scheme.

4. An Application Example

To illustrate a practical application of the proposed model-based conditional power assessment method, we apply it to the design of extending a type I diabetes clinical trial study ([1]). In this study, researchers investigate the efficacy of anti-thymocyte globulin (thymoglobulin) and pegylated GCSF (neulasta) versus placebo in retaining C-peptide production for the established type 1 diabetes (T1D) patients. During the study duration, 25 T1D subjects were recruited and randomized to the drug and placebo groups at 2:1 ratio (with 17 in drug and 8 in placebo, respectively) in double-blinded fashion. The primary outcome is the change from the baseline of two-hour C-peptide area under the curve (AUC) at 12 months.

Among the 25 participants, 1 subject from the drug group missed the 12-month measurement. Even though a marginally significant treatment effects ($p \approx 0.05$) exists, due to the small sample size, researchers are interested in extending the study to consolidate the scientific findings. We used the two methods discussed in this paper to determine how much more information is needed. Based on the available observations thus far, the treatment effects estimate (standard error) via the summary statistic and model-based methods are 0.83(0.40) and 0.86(0.37), respectively. In the model-based analysis, the baseline two-hour C-peptide AUC was included as a covariate in the model fit. Even though both methods suggest the treatment effects is a little bit above 0.8, to be somewhat conservative, in our design, we set the treatment effects to be detected as 0.80 and control the overall type I error at level 0.05 in the two-sided test with design parameters $p_u = 0.025$ and $p_l = 0.15$ as previously defined. Details of the final sample size requirement to reach the targeted conditional power based on the two methods discussed in this paper are summarized in Table 5.

According to the model-based approach, we would need 92 additional subjects to reach 0.99 conditional power with 61 assigned to the drug and 31 assigned to the placebo group, respectively. However, if the summary statistic conditional power method is used, we would need 158 more subjects to reach this goal with 105 in drug and 53 in placebo groups correspondingly. This example further demonstrates the advantage of the proposed scheme over the existing method for clinical studies.

5. Discussions

It is well known that the ANCOVA estimator is in general more efficient than the ANOVA estimator. Correspondingly, the model-based conditional power is thereby expected to be better than the summary statistic conditional power. Nevertheless, it is still of interest to have simulations which shed light on the extent to which it is better, particularly since the model-based conditional power has the complexity of involving the maximum likelihood treatment effects estimate that accounts for the future unobserved data. In this paper, we extend the conventional conditional power with a general model-based analytic assessment to make use of other useful information embedded in the observed data. Hence, the proposed approach is more efficient for treatment effects estimation. The proposed analytic method offers researchers a useful tool to assess the conditional power without the need to deal with the future unobserved data directly. Our numerical results demonstrate the advantages of the proposed framework over the traditional method under various settings. Our proposed approach depends on the asymptotic relationship for parameters in the test statistics based on the observed data and final samples. Intuitively, one may adopt an alternative method to estimate these parameters by using bootstrapping [19] with replacement on the available observations to obtain future unobserved data. However, this empirical version is not ideal since it overestimates the correlation parameter. Furthermore, it should be noted that the ideas for extending pilot studies in order to increase power for finding an effect of an intervention need to be interpreted with some caution, mainly because their intent is not in general but rather for situations where the possibility for such extension is pre-specified.

Compared with the conventional method, our approach is more robust to missing data by reducing the treatment effects estimation bias. Our simulation study shows that the model based strategy allows us to obtain a valid test that controls type I error well while the conventional method fails under various simulation set ups. However, how the proposed method behaves for data under more complicated missing mechanisms deserves further attention which is beyond the scope of this paper. In addition, it is worth pointing out that model-based treatment effect estimate is a conditional treatment effect while the treatment effect estimate based on summary statistics is a marginal treatment effect. For continuous data, the conditional and marginal treatment effects are equivalent (e.g. [20]). For binary outcomes, the marginal and conditional treatment effects are different if the treatment effect is evaluated through the odds ratio or risk ratio (e.g. [21]) since they are nonlinear treatment effects in these scenarios. However, as shown by [11] and [12], the treatment effect estimate could be biased if summary statistics is used and it may lead to inappropriate conditional power assessment.

It is also worth mentioning that the two-stage sampling scheme to determine the rejection cutoff for extending RCT studies may not work well enough to control the overall empirical type I error in the scenarios of extreme observed sample proportions, i.e. too much observed or unobserved. This is not limited to our proposed method but also applies to the conventional conditional power approach. Under this scenario, in practice, we recommend to use Monte Carlo simulation to generate the future data from the null distribution and obtain the empirical rejection cutoff instead of using the theoretical cutoff by solving the integration equation (3). Although the proposed analysis framework is general, extending the

proposed method to other response data types, such as longitudinal data, and extending to more than one look interim analysis, needs to be investigated.

Acknowledgments

B. Zou was partially supported by NIH grant 1UL1TR000064 from the National Center for Advancing Translational Sciences. Authors would like to thank the editor, associate editor, and two anonymous reviewers for their constructive comments.

Appendix: Derivation of (5) and (6)

Without loss of generality, let $\hat{\theta}$ denote the maximum likelihood estimate (MLE) of parameter θ in model $f(y|\theta)$ with true value θ_0 . Data $y_1, \dots, y_n, y_{n+1}, \dots, y_N$ are i.i.d. with y from model $f(y|\theta)$. Define $\ell(y|\theta) = \log f(y|\theta)$ and

$I(\theta_0) = \mathbb{E}_{\theta_0}(\ell'(y|\theta_0))^2 \equiv \mathbb{E}_{\theta_0}(\frac{\partial}{\partial \theta} \log f(y|\theta)|_{\theta=\theta_0})^2$. Therefore, we obtain:

$$\mathbb{E}_{\theta_0}(\ell''(y|\theta_0)) \equiv \mathbb{E}_{\theta_0}(\frac{\partial^2}{\partial \theta^2} \log f(y|\theta)|_{\theta=\theta_0}) = -I(\theta_0).$$

For MLE of θ based on n observed interim samples, i.e. $\hat{\theta}_{obs}$, we have: $\ell'_n(\hat{\theta}_{obs}) = 0$ with

$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(y_i|\theta)$. By the mean value theorem, we write:

$$0 = \ell'_n(\hat{\theta}_{obs}) \approx \ell'_n(\theta_0) + \ell''_n(\hat{\theta}_1)(\hat{\theta}_{obs} - \theta_0)$$

for some $\hat{\theta}_1 \in [\hat{\theta}_{obs}, \theta_0]$. Thus, we get $\sqrt{n}(\hat{\theta}_{obs} - \theta_0) \approx -\frac{\sqrt{n}\ell'_n(\theta_0)}{\ell''_n(\hat{\theta}_1)}$. Therefore, by the

asymptotic property of MLE, we have

$$\hat{\theta}_{obs} \approx \theta_0 + \frac{1}{nI(\theta_0)} \sum_{i=1}^n \ell'(y_i|\theta_0) \quad (9)$$

By the same argument, we have the following for the MLE of θ based on all N samples, i.e. $\hat{\theta}_{all}$

$$\hat{\theta}_{all} \approx \theta_0 + \frac{1}{NI(\theta_0)} \sum_{i=1}^N \ell'(y_i|\theta_0) \quad (10)$$

By the asymptotic normality of MLE, we obtain: $\sqrt{n}(\hat{\theta}_{obs} - \theta_0) \sim N(0, \frac{1}{I(\theta_0)})$ and

$\sqrt{N}(\hat{\theta}_{all} - \theta_0) \sim N(0, \frac{1}{I(\theta_0)})$. Therefore, we have:

$$\frac{\sigma_{all}^2}{\sigma_{obs}^2} \equiv \lim_{N,n \rightarrow \infty} \frac{Var(\hat{\beta}_{z,all})}{Var(\hat{\beta}_{z,obs})} = \lim_{N,n \rightarrow \infty} \frac{n}{N} \equiv \tau \text{ where } N = N_0 + N_1, n = n_0 + n_1.$$

Furthermore, by (9) and (10), we obtain:

$$Cov(\hat{\theta}_{obs}, \hat{\theta}_{all}) \approx \frac{1}{I^2(\theta_0)} \mathbb{E}_{\theta_0} \left\{ \left[\frac{1}{n} \sum_{i=1}^n l'(y_i | \theta_0) \right] \left[\frac{1}{N} \sum_{j=1}^N l'(y_j | \theta_0) \right] \right\}$$

Under the i.i.d. assumption of observations, by using the result $I'(\theta_0) = \mathbb{E}_{\theta_0}(l'(y | \theta_0)) = 0$, we rewrite the above covariance as:

$$Cov(\hat{\theta}_{obs}, \hat{\theta}_{all}) \approx \frac{n}{NI^2(\theta_0)} \mathbb{E}_{\theta_0} \left\{ \left[\frac{1}{n} \sum_{i=1}^n l'(y_i | \theta_0) \right] \left[\frac{1}{n} \sum_{j=1}^n l'(y_j | \theta_0) \right] \right\} = \frac{1}{NI^2(\theta_0)} \mathbb{E}_{\theta_0} \{ (l'(y | \theta_0))^2 \} = \frac{1}{NI(\theta_0)}$$

Therefore, we conclude:

$$\begin{aligned} \rho &\equiv \lim_{N,n \rightarrow \infty} Corr(\hat{\beta}_{z,all}, \hat{\beta}_{z,obs}) = \lim_{N,n \rightarrow \infty} \frac{Cov(\hat{\beta}_{z,all}, \hat{\beta}_{z,obs})}{\sqrt{Var(\hat{\beta}_{z,all})Var(\hat{\beta}_{z,obs})}} = \lim_{N,n \rightarrow \infty} \sqrt{\frac{n}{N}} \\ &\equiv \sqrt{\tau} \end{aligned}$$

References

1. Haller MJ, et al. ATG and G-CSF preserves beta cell function in established type 1 diabetes. *Journal of Clinical Investigation*. 2015; 125(1):448–455. [PubMed: 25500887]
2. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics, Part C-Sequential Analysis*. 1982; 1:207–219.
3. Lan KKG, Wittes J. The B-value: a tool for monitoring data. *Biometrics*. 1988; 44:579–585. [PubMed: 3390511]
4. Albert RK, et al. Azithromycin for prevention of exacerbations of COPD. *The New England Journal of Medicine*. 2011; 365:689–698. [PubMed: 21864166]
5. Bhatt DL, et al. Effect of platelet inhibition with cangrelor during PCI on ischemic events. *The New England Journal of Medicine*. 2013; 368:1303–1313. [PubMed: 23473369]
6. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*. 2005; 24:2747–2764. [PubMed: 16134130]
7. Shih WJ. Sample size re-estimation—journey for a decade. *Statistics in Medicine*. 2001; 20:515–518. [PubMed: 11223897]
8. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*. 2001; 55:314–321.
9. Ciolino JD, Martin RH, Zhao W, Jauch EC, Hill MD, Palesch YY. Continuous covariate imbalance and conditional power for clinical trial interim analyses. *Contemporary clinical trials*. 2014; 38(1):9–18. [PubMed: 24607294]
10. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*. 1989; 8(4):467–475. [PubMed: 2727470]
11. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984; 71:431–444.
12. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993; 71:431–444.

13. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric Efficiency and Its Implication on the Design and Analysis of Group-Sequential Studies. *Journal of the American Statistical Association*. 1997; 92(440):1342–1350.
14. Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics*. 1999; 55:853–857. [PubMed: 11315017]
15. Li G, Shih WJ, Lu J. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*. 2002; 3(2):277–287. [PubMed: 12933618]
16. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995; 51:1315–1324. [PubMed: 8589224]
17. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70(3): 659–663.
18. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63(3):581–592.
19. Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*. 1979; 7:1–26.
20. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*. 2013; 32:2837–2849. [PubMed: 23239115]
21. Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*. 2007; 26:5529–5544. [PubMed: 18058851]

Figure 1. Finite Sample Performance for Parameter Estimates of Gaussian Data

Figure 2. Finite Sample Performance for Parameter Estimates of Binary Data

Table 1
Simulation Results for Gaussian Data without Missing Data Scenario^a

Summary Statistic			Model Based	
n^b	Cp ^d	Emp. Type I Error ^e	Cp ^d	Emp. Type I Error ^e
Model Specified Correctly				
40	0.457	0.051	0.583	0.049
50	0.498	0.048	0.635	0.047
60	0.548	0.047	0.695	0.048
Model Specified Incorrectly (I)				
40	0.412	0.047	0.527	0.043
50	0.466	0.049	0.593	0.046
60	0.516	0.049	0.652	0.045
Model Specified Incorrectly (II)				
40	0.372	0.048	0.376	0.049
50	0.409	0.049	0.413	0.048
60	0.453	0.045	0.458	0.045
Skewed Error Distribution				
40	0.437	0.045	0.579	0.039
50	0.492	0.039	0.646	0.044
60	0.533	0.049	0.695	0.052

^aStudy design parameters $p_H=0.025$ and $p_F=0.15$

^bObserved sample size in each arm

^cTargeted final sample size in each arm

^dAverage conditional power achieved

^eEmpirical overall type I error while the theoretical overall type I error controlled at 0.05

Table 2
Simulation Results for Gaussian Data with Missing Data Scenario^a

Summary Statistic					Model Based		
$(n_1, n_1)^b$	$(N_1, N_0)^c$	Cp ^d	Emp. Type I Error ^e	Avg($\hat{\beta}_{obs}$) ^f	Cp ^d	Emp. Type I Error ^e	Avg($\hat{\beta}_{obs}$) ^f
Missing Data Scenario I							
	(68,72)	0.516	0.127	0.165	0.849	0.048	0.347
(51,54)	(85,90)	0.521	0.144	0.167	0.857	0.047	0.350
	(102,108)	0.563	0.152	0.167	0.883	0.052	0.350
Missing Data Scenario II							
	(72,68)	0.891	0.130	0.533	0.850	0.047	0.350
(54,51)	(90,85)	0.877	0.142	0.533	0.860	0.047	0.351
	(108,102)	0.879	0.153	0.528	0.881	0.048	0.347

^a Study design parameters $p=0.025$ and $p=0.15$

^b Actual observed sample size at treated and control arm

^c Targeted final samples at treated and control arm

^d Average conditional power achieved

^e Empirical overall type I error while theoretical overall type I error controlled at 0.05

^f Treatment effect estimated on the available observations (true $\beta=0.35$)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Simulation Results for Binary Outcomes^a

Summary Statistic				Model Based	
n^b	N^c	CP^d	Emp. Type I Error ^e	CP^d	Emp. Type I Error ^e
80	120	0.411	0.048	0.895	0.049
	160	0.530	0.050	0.915	0.049
200	200	0.683	0.050	0.957	0.045

^a Study design parameters $p_H=0.025$ and $p_I=0.15$
^b Observed sample size in each arm
^c Targeted final samples in each arm
^d Average conditional power achieved
^e Empirical overall type I error while theoretical overall type I error controlled at 0.05

Table 4

Simulation Results for Binary Data with Missing Data Scenario^a

$(n_1, n_0)^b$	Summary Statistic			Model Based	
	$(N_1, N_0)^c$	CP ^d	Emp. Type I Error ^e	CP ^d	Emp. Type I Error ^e
Missing Data Scenario I					
(68, 72)	(102, 108)	0.145	0.182	0.868	0.049
	(136, 144)	0.263	0.202	0.887	0.049
	(170, 180)	0.394	0.225	0.920	0.051
Missing Data Scenario II					
(72, 68)	(108, 102)	0.667	0.182	0.864	0.050
	(144, 136)	0.748	0.217	0.884	0.053
	(180, 170)	0.816	0.223	0.920	0.051

^aStudy design parameters $p_U=0.025$ and $p_I=0.15$

^bActual observed sample size at treated and control arm

^cTargeted final samples at treated and control arm

^dAverage conditional power achieved

^eEmpirical overall type I error while theoretical overall type I error controlled at 0.05

Table 5
Design Details of Sample Size Requirement for Extending T1D Study^{a,b}

Summary Statistic		Model Based				
		N_0^f	N_1^g	N_0^f	N_1^g	
n_0^c n_1^d	CPe	0.85	10	20	10	20
		0.90	13	25	10	20
		0.95	20	39	12	24
		0.99	61	121	39	77

^aStudy design parameters $pU=0.025$ and $pF=0.15$

^bTheoretical overall type I error controlled at $\alpha = 0.05$

^cObserved sample size in placebo arm

^dObserved sample size in drug arm

^eTargeted conditional power to achieve

^fFinal sample size required in placebo arm

^gFinal sample size required in drug arm