Published in final edited form as:

Biometrics. 2020 June ; 76(2): 643–653. doi:10.1111/biom.13161.

Designing precision medicine trials to yield greater population impact

Ying-Qi Zhao*, Michael L. LeBlanc

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109.

Summary:

Traditionally, a clinical trial is conducted comparing treatment to standard care for all patients. However, it could be inefficient given patients' heterogeneous responses to treatments, and rapid advances in the molecular understanding of diseases have made biomarker-based clinical trials increasingly popular. We propose a new targeted clinical trial design, termed as Max-Impact design, which selects the appropriate subpopulation for a clinical trial and aims to optimize population impact once the trial is completed. The proposed design not only gains insights on the patients who would be included in the trial, but also considers the benefit to the excluded patients. We develop novel algorithms to construct enrollment rules for optimizing population impact, which are fairly general and can be applied to various types of outcomes. Simulation studies and a data example from the SWOG Cancer Research Network demonstrate the competitive performance of our proposed method compared to traditional untargeted and targeted designs.

Keywords

Biomarkers; population impact; precision medicine; targeted clinical trial design

1. Introduction

Traditionally, a clinical trial is conducted comparing treatment to standard care for all patients. However, patient responses to treatments are often heterogeneous. For example, it is not uncommon that less than 50% of patients respond to a particular cancer therapy. In this case, a traditional clinical trial could be inefficient. New technologies such as DNA sequencing have provided evidence on new signatures and biomarkers (Rosenwald et al., 2002), and diseases are increasingly understood on a cellular and molecular level (Renfro et al., 2017). One follow-up task is to validate such biomarkers in the clinical trial setting. These biomarkers can usually be categorized into two classes. Prognostic markers are related to disease prognosis, and they give information on the disease outcome independent of the treatment received. Usually prognostic markers are informative for risk stratification. On the other hand, predictive markers are based on a specific treatment, and they indicate who might benefit from a targeted agent. These advances change the way we design and perform

* yqzhao@fredhutch.org.

Supporting Information

Web Appendices, Figures and simulation codes referenced in Sections 2 and 3 are available with this paper at the *Biometrics* website on Wiley Online Library.

clinical trials. A now established paradigm for the evaluation of new treatments for cancer (and proposals for prevention strategies) utilizes designs that enrich outcomes for patients most likely to exhibit maximum treatment effects and improves the efficiency of clinical trials (Simon and Maitournam, 2006; Mandrekar and Sargent, 2009).

There is a wealth of literature on enrichment designs that focus on the targeted group (Maitournam and Simon, 2005; Hoering et al., 2008). Usually people identify the desired combination and/or the cutoff among several candidate biomarkers to initiate an enrichment trial based on the criterion to maximize the treatment effect, and subsequently optimize the power of the trial. A successful cancer trial is considered by some in the pharmaceutical industry or government as study that shows there exists some subset of patients (even a very small subset) for which the new treatment offers some non-zero positive improvement in important clinical outcomes such as survival, progression-free survival, or clinical tumor response. Such strategies for evaluation of promising cancer treatments often do not serve the overall public health interest. In particular, such a targeted design tends to enroll a smaller number of patients to be randomized without insights gained on the group of excluded patients, so the impact on the overall population remains uninvestigated. In addition, by only randomizing to a small group of marker-positive patients who respond to the new treatment, we may ultimately miss and not use the new treatment in many patients who would benefit from a new treatment (or prevention strategy), especially when there is the possibility that the new treatment helps marker-negative patients. Other examples of biomarker-based design include marker-stratified and marker strategy designs. In marker-stratified design, all patients are first stratified by marker status and then randomized within their given marker cohort. This could be applied when there is insufficient evidence to exclude a subpopulation from randomization (Sargent and Allegra, 2002). The marker strategy design is usually used to validate the biomarker in treatment decision-making process. Patients are randomized to a treatment strategy accounting for biomarker status versus a treatment ignoring the biomarker. However, alternative markers and treatment strategies cannot be assessed (Sargent et al., 2005; Renfro et al., 2017).

We propose an enrichment design, termed as Max-Impact design, focusing on study designs that lead to a greater population impact by maximizing the expected future patient outcomes both on and off the trial patients. We focus on the situation where the underlying marker distribution is continuous, since markers are often based on levels of a specific chemical in the blood or on a combination of gene expression levels (Hoering et al., 2008). However, the best cutpoint on a single ordered biomarker or combination of biomarkers may be unknown. In the Max-Impact design, we formalize a novel objective function to reflect the new criterion based on biomarker values, i.e., the ultimate impact of a trial where a subgroup is selected for participation. Using the preliminary data from a phase II trial, we identify the biomarker combination and cutoffs to optimize the new objective function. The resulting biomarker-based rule can be utilized as the enrollment criterion for a future phase III enrichment design. We show that by conducting the follow-up phase III study using the proposed strategy, the population impact can be optimized. The Max-Impact design is applicable to various types of responses, including continuous, binary, and survival endpoints.

While an enrichment design usually requires fewer patients for randomization, savings in trial costs can be negated by additional costs from biomarker measurement in patient screening. Hence, we also consider practical constraints for the future trial such that at least a moderately sized group is eligible. The newly defined objective function, along with the constraints, is non convex, so it is difficult to optimize empirically when using data from a phase II trial. We will develop new algorithms to identify the optimal eligible group that leads to the greatest impact. Throughout, our methods are tied to our experience in cancer clinical trials primarily through the SWOG Cancer Research Network and more broadly through our involvement in the NCI National Clinical Trials Network (NCTN). In particular, we illustrate the proposed design using data from SWOG S0421 study, a prostate cancer clinical trial.

The following article is organized as follows. In Section 2 we introduce the concept of ultimate impact of a study, and we develop algorithms for optimizing ultimate impact with or without a constraint on the size of the eligible group. In Section 3 we conduct extensive simulation studies to examine the empirical performances of the proposed Max-Impact design, and other targeted and untargeted designs. A data example using data from the SWOG Cancer Research Network S0421 trial is presented in Section 4, and Section 5 provides a discussion.

2. Methodology

2.1 Ultimate impact of a study

Assume that the enrollment criterion for the enrichment trial is based on a marker measure(s) X , where patients should be enrolled if some function of X , $f(X)$, is positive.

$$f(X) > 0 : \text{enrolled},$$

$$f(X) \leq 0 : \text{not enrolled}.$$

Let Y denote the response, with larger outcomes being more preferable. Let $A \in \{0, 1\}$ denote the treatment options, where $A = 0$ represents standard care and $A = 1$ represents new treatment. Patients enrolled in the trial will be randomized to $A = 0$ or 1 . Let $Y(a)$ be the potential outcome that would be observed were the subject to receive treatment a , $a = 0, 1$ (Rubin, 1978; Splawa-Neyman et al., 1990). We will recommend treatment for a new individual X^* based on the trial result as follows: if the trial is successful, X^* with $f(X^*) > 0$ will receive new treatment and otherwise, they will be provided with standard care; for X^* with $f(X^*) \leq 0$, they will receive standard care. Assume the power of the study is P_f under the alternative hypothesis, based on an enrollment criterion of $f(X) > 0$. We denote the ultimate impact (UI) of the study as

$$UI(f) = E\{E\{Y(0) | f(X^*) \leq 0\} P\{f(X^*) \leq 0\} + [(1 - P_f)E\{Y(0) | f(X^*) > 0\} + P_f E\{Y(1) | f(X^*) > 0\}] P\{f(X^*) > 0\}\}. \quad (1)$$

A higher ultimate impact will benefit more individuals in the whole population of patients with that cancer. Under certain regularity conditions, including (i) strong ignorability, $\{Y(0), Y(1)\} \perp\!\!\!\perp A/X^*$ and (ii) consistency, $Y = Y(A)$, we have

$$E\{Y(a)|f(X^*) > 0\} = E[E\{Y(a)|X^*\}|f(X^*) > 0] = E\{E\{Y|X^*, A = a\}|f(X^*) > 0\}.$$

Hence, (1) equals to

$$\begin{aligned} UI(f) &= E(Y|A = 0) + P_f E\{E(Y|A = 1, X^*) - E(Y|A = 0, X^*)|f(X^*) > 0\}P\{f(X^*) > 0\} \\ &= E(Y|A = 0) + P_f E[\Delta(X^*)I\{f(X^*) > 0\}], \end{aligned} \quad (2)$$

where $\Delta(x) = E(Y|A = 1, X = x) - E(Y|A = 0, X = x)$. Therefore, one can describe the $UI(f)$ as the expected outcome for patients if they were treated with standard therapy plus the expected improvement in therapy for the fraction of patients in the study target group times the power of the study under the alternative hypothesis. This is in contrast to the traditional biomarker-driven enrichment design, where usually the goal is to maximize the power of the study in order to obtain a significant result. Our proposal proceeds from a population impact perspective, which takes into account benefits to the subjects regardless of their enrollment in the trial. This design could also mitigate the generalizability issue of clinical trials.

We illustrate the idea of ultimate impact using a simple marker treatment model that was investigated in Hoering et al. (2008). Assume the marker X is normally distributed $X \sim N(\mu, \sigma^2)$, $\mu = 0.5$, $\sigma^2 = 0.5$ with its density function denoted by $h(x)$. The treatment A is generated from $\{0, 1\}$ independently of X with $P(A = 1) = 1/2$. The outcome is binary and follows a logit link. Assume that c is the cutpoint that distinguishes the marker-positive patients from the marker-negative patients. The expected outcome for the subgroup of $x > c$ patients is

$$\int_{x > c} \frac{e^{a_0j + a_1jx}}{1 + e^{a_0j + a_1jx}} g(x)/v_{x > c},$$

where $v_{x > c} = \int_{x > c} h(x)dx$ and $j = 0, 1$ index the treatment group receiving $A = 0$ or $A = 1$.

We study four examples based on this marker treatment model, where the model parameters are presented in the web supplementary materials. The response distributions in both groups are presented for each example. In Example 1, the marker has no effect on the outcome. The marker is a predictive marker in Examples 2 and 3. In Example 2, the new treatment $A = 1$ does not help the marker-negative patients compared to the standard treatment $A = 0$. Conversely, it has additional benefit for marker-positive patients, which increases with increasing marker value. In Example 3, the new therapy benefits patients whose marker values are greater than 0.5, but has a detrimental effect otherwise. Finally, in Example 4, the marker is a prognostic marker, where $A = 1$ is overall better than $A = 0$, and both increase with the increasing marker value.

We assume a fixed screening size of 400 patients. Different cutoffs of the marker lead to enrichment clinical trials with different sizes. We plot the non centrality parameter curve and

the impact curve of the enrichment trial for different examples. We plot the non centrality parameter curve instead of the power curve because as sample sizes increase, the power curves become very close to one for a range of cutoff values, so they become difficult to assess the shape. But the non centrality parameter informs the power of the resulting study. For example, when designing a study at the level of 0.05, two sided,

$$Power \approx 1 - \Phi(-1.96 + \delta),$$

where δ is the non centrality parameter, and Φ is the standard normal distribution function. A design that enriches the population of patients who are thought to show the maximum treatment effect will select the cutoff value where the non centrality parameter curve reaches the maximum. Conversely, the proposed Max-Impact design will select the cutoff that corresponds to the highest point on the impact curve. As shown in Figure 1, as the eligibility criteria of the clinical trial vary over c , both curves change. In Example 1, since the treatment arm was uniformly better than the control arm, both designs selected the whole population for a future trial. In other examples, the treatment and the control arms had differential effects across biomarker values. While selection of the cutoff value in the design to maximize the power are mainly driven by the effect size, a trial would enroll more patients if it were designed to maximize the ultimate impact since it would also study the marker-negative patients. Since we may assess the effects more broadly to achieve the maximal impact, more patients will gain benefit from the trial if the trial is ultimately positive.

We additionally plot the curves for screening sizes of 200 and 2000, presented in web supplementary materials. We observe that the impact curve also peaks at a lower cut-off value. The shapes of the non centrality parameters curves remain the same regardless of the screening size, with varying magnitudes of the parameters. Hence, the cutoffs for enrolling patients under a Max-Power design are the same regardless of the screening size. The cutoffs under a Max-Impact design are exactly the points where treatment starts to outperform control. This is because when the screening size is large, the randomization size is large as well. Subsequently, the power of the study is close to 1. The ultimate impact $U(f)$ defined in (2) is equivalent to maximizing $E[(X^*)I\{f(X^*) > 0\}]$, which is maximized by selecting the patients with $(X^*) > 0$.

We also investigated the curves for enrichment trials with fixed randomization sample sizes, where eligibility criteria entering the trial can vary according to the cutoffs of the marker. The results are presented in the web supplementary materials. In general, trials designed to maximize the power tend to favor a more restrictive group where the enrolled patients have a larger effect size. For example, in Examples 2 and 3, the design for maximizing the power selected the cutoff as the largest marker value because the effect size increased over marker values. It is thus infeasible if the trial is solely designed to maximize the power. In particular, the downside to stricter eligibility is that a larger number of patients is needed for screening. This factor should be taken into account when selecting the cutoff, since the ability to screen patients is usually limited by the study cost and time required to complete accrual.

2.2 Max-Impact design

In this section, we introduce the proposed Max-Impact design, which identifies the subpopulation for a future study to maximize the ultimate impact. In particular, the trial could be designed to solely optimize the ultimate impact, or to optimize the impact with additional constraints on trial implementation. Throughout, we assume that $\pi(x)$ is the same between the trial population and the target population. That is, the treatment works the same in both populations conditional on x .

2.2.1 Optimizing ultimate impact.—The ultimate impact (2) depends on the power of the enriched study, P_F . Usually we design or target a larger effect size as we restrict patient enrollment, but would like to design the trial at a constant power regardless of the eligibility criterion. Thus, to maximize (2), it is equivalent to optimize

$$E[\Delta(X^*)I\{f(X^*) > 0\}].$$

This indicates that $I\{\rho^{\text{pt}}(X^*) > 0\} = I\{\hat{\rho}(X^*) > 0\}$, where ρ^{pt} is the optimizer.

Given the data from an existing phase II trial, we can estimate $\pi(x)$ with $\hat{\Delta}(x)$. Subsequently, for any x , a plug-in estimator for ρ^{pt} can be utilized as $I\{\hat{f}(x) > 0\} = I\{\hat{\Delta}(x) > 0\}$. If we can observe the triplet (Y, A, X) completely for each subject, a general approach to estimate $\pi(x)$ is to model the mean for both treatment groups, $E(Y|X, A)$, using parametric and nonparametric methods such as regression or statistical learning methods. For example, linear regression including covariate by treatment interactions can be used for continuous outcomes, and logistic regression can be applied for binary outcomes. Then $\hat{\Delta}(x) = \hat{E}(Y|X = x, A = 1) - \hat{E}(Y|X = x, A = 0)$.

Now, consider the case that Y may not be fully observed due to the censoring. Assume that the data consists of n independent identically distributed subjects, $\{U_i = T_i \wedge C_i, \delta_i = I(T_i \leq C_i), X_i, A_i\}$ $i = 1, \dots, n$, where T denotes the true survival time, and $\delta = I(T \leq C)$ denotes the censoring indicator. A common approach to model the event time is a Cox proportional hazards model (Cox, 1972). For example, we impose the working model for the hazard function of failure time $\lambda(t)$,

$$\lambda(t) = \lambda_0(t) \exp\{\beta^T h(X, A)\},$$

where $\lambda_0(t)$ is the baseline hazard function for the failure time, and $h(X, A)$ is a known vector function of (A, X) . β can be estimated by maximizing the partial likelihood. Suppose the outcome of interest is $Y = I(T \geq t_0)$, where t_0 is a given time point. Then $\pi(x)$ can be estimated using

$$\hat{\Delta}(X) = g\{\log \hat{\Lambda}_0(t_0) + \hat{\beta}^T h(X, 1)\} - g\{\log \hat{\Lambda}_0(t_0) + \hat{\beta}^T h(X, 0)\},$$

Where $g(x) = e^{-e^x}$, $\Lambda_0(t)$ is the baseline cumulative hazard function, and can be estimated by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n I(U_j \geq s) e^{\hat{\beta}^T h(X_j, A_j)}}$$

with $N_i(t) = I(U_i \leq t) \delta_i$. If we are interested in the restricted mean survival time, where $Y = \min(T, \tau_0)$, then $E(X)$ can be estimated using

$$\hat{\Delta}(X) = \int_0^{\tau_0} g \left\{ \log \hat{\Lambda}_0(t) + \hat{\beta}^T h(X, 1) \right\} - g \left\{ \log \hat{\Lambda}_0(t) + \hat{\beta}^T h(X, 0) \right\} dt.$$

2.2.2 Optimizing ultimate impact under constraints.—For practical reasons, studies are usually designed with certain considerations. For example, we may only provide the new treatment to patients when the benefit outweighs, or at least offsets its extra risk. To incorporate this consideration, we can modify the impact criterion to,

$$E[\Delta(X^*) I\{\Delta(X^*) > \delta\} I\{f(X^*) > 0\}].$$

where $\delta > 0$ quantifies the tradeoff between treatment efficacy and cost or risk associated with new treatment.

Sometimes, the value of δ is unknown. We can impose a desired condition that the overall impact in terms of treatment risk, for instance experience of serious adverse events, is controlled at certain level to actually conduct a randomized phase III study. Let R denote a risk outcome, and $R(x) = E(R|A = 1, X = x) - E(R|A = 0, X = x)$. Such a constraint can be expressed as

$$UI_R(f) = E(R|A = 0) + P_f E[\Delta_R(X^*) I\{f(X^*) > 0\}] \leq v'_0,$$

where v'_0 is a prespecified constant such that the overall impact on risk is well controlled.

Suppose that we design the trial at a constant power. Given the condition, the objective can be written as

$$\begin{aligned} & \max_f E[\Delta(X^*) I\{f(X^*) > 0\}] \\ & \text{s.t. } E[\Delta_R(X^*) I\{f(X^*) > 0\}] \leq v_0, \end{aligned} \quad (3)$$

where $v_0 = \{v'_0 - E(R|A = 0)\} / P_f$.

Recall that the optimal solution $I\{f^{\text{opt}}(X^*) > 0\} = I\{X^* > 0\}$ if we do not impose any constraints. Hence, if $E[\Delta_R(X^*) I\{f(X^*) > 0\}] \leq v_0$, then the optimal solution to (3), denoted by $f_c^{\text{opt}}(x)$, is also $I\{f_c^{\text{opt}}(X^*) > 0\} = I\{X^* > 0\}$. Otherwise, we solve for

$$\max_f E[\Delta(X^*) I\{f(X^*) > 0\}], \text{ s.t. } E[\Delta_R(X^*) I\{f(X^*) > 0\}] = v_0.$$

Following the arguments provided in Wang et al. (2018), we apply the Lagrange multiplier, $f_c^{\text{opt}}(x^*)$ maximizes

$$E[\{\Delta(X^*) - \lambda \Delta_R(X^*)\} I\{f(X^*) > 0\}]$$

subject to $E[\Delta_R(X^*) I\{f(X^*) > 0\}] = v_0$ and $\lambda > 0$. Clearly, $I\{f_c^{\text{opt}}(X^*) > 0\} = I\{\Delta(X^*) - \lambda \Delta_R(X^*) > 0\}$, where λ solves the equation,

$$E[\Delta_R(X^*) I\{\Delta(X^*) - \lambda \Delta_R(X^*) > 0\}] = v_0 \quad (4)$$

Denote the unique solution of (4) by λ^* , and $I\{f_c^{\text{opt}}(X^*) > 0\} = I\{\Delta(X^*) - \lambda^* \Delta_R(X^*) > 0\}$

Let $\mathbb{P}_n$ denote the empirical average of the data. We estimate $\Delta(X^*)$ by $\hat{\Delta}(X)$ using similar strategies as outlined in Section 2.2.1. $\Delta_R(X^*)$ can be estimated by $\hat{\Delta}_R(X)$ using the appropriate methods depending on the type of the outcome. The optimal rule for enrollment with a risk constraint can be estimated by $I\{\hat{f}(X) > 0\} = I\{\hat{\Delta}(X) - \hat{\lambda} \hat{\Delta}_R(X) > 0\}$, where $\hat{\lambda} = 0$ if $\mathbb{P}_n[\hat{\Delta}_R(X) I\{\hat{\Delta}(X) > 0\}] \leq v_0$, and $\hat{\lambda}$ is the solution to $\mathbb{P}_n[\hat{\Delta}_R(X) I\{\hat{\Delta}(X) - \hat{\lambda} \hat{\Delta}_R(X) > 0\}] = v_0$ if $\mathbb{P}_n[\hat{\Delta}_R(X) I\{\hat{f}(X) > 0\}] > v_0$.

Remark 1: Enrichment usually leads to smaller trials, which are less expensive. However, savings in trial costs are usually accompanied by additional costs for biomarker measurement in the screening step. Hence, it is appealing to control the total costs for an enrichment trial, including patient screening costs and per-patient costs for patients in the trial. We can extend the proposed Max-Impact design to accommodate such a constraint, in addition to the risk constraint we discussed above. For example, let c_s denote the cost per screened patient, c_r denote the cost per patient randomized in the study, n_r denote the number of patients randomized in the study and n_s denote the number of patients screened for the study. Assume the rule for selecting a patient into the trial is $I\{f(X^*) > 0\}$, then $1/P\{f(X^*) > 0\}$ patients must be screened to identify one eligible patient. The total cost of the trial can be calculated as $TC = n_s c_s + n_r c_r = n_s c_s + n_s P\{f(X^*) > 0\} c_r$. Subsequently, we can add this additional constraint to (3) and solve the resulting problem.

2.3 Sample size determination for the future study

In practice, we do not usually know the effect size curves with respect to different enrollment criteria. To determine the sample size for a future study once we derive the rule for enrolling patients using the Max-Impact design, we need to estimate the effect size within the eligible group using the available data.

We first consider the continuous endpoints. Assume that the estimated rule for enrollment is $I\{\hat{f}(X^*) > 0\}$ which can be derived with or without constraints when optimizing ultimate impact. Using the current data, the mean difference between different arms can be estimated by

$$\hat{\theta}(\hat{f}) = \frac{\mathbb{P}_n[YI\{\hat{f}(X) > 0\}I(A=1)]}{\mathbb{P}_n[I\{\hat{f}(X) > 0\}I(A=1)]} - \frac{\mathbb{P}_n[YI\{\hat{f}(X) > 0\}I(A=0)]}{\mathbb{P}_n[I\{\hat{f}(X) > 0\}I(A=0)]}.$$

The variance within the group can also be estimated correspondingly, denoted by $\hat{\sigma}^2(\hat{f})$. To design a study with power $1 - \beta$ at the level of α , two-sided, the sample size is

$$\frac{(z_\beta + z_{\alpha/2})^2 \hat{\sigma}^2(\hat{f})}{\hat{\theta}^2(\hat{f})},$$

where $z_{\alpha/2}$ and z_β are standard normal percentiles.

For the binary endpoints, we estimate the success probability in each arm, where for $a = 0, 1$,

$$\hat{p}_a(\hat{f}) = \frac{\mathbb{P}_n[YI\{\hat{f}(X) > 0\}I(A=a)]}{\mathbb{P}_n[I\{\hat{f}(X) > 0\}I(A=a)]}.$$

The sample size can be determined by

$$\frac{(z_\beta + z_{\alpha/2})^2 [\hat{p}_1(\hat{f})\{1 - \hat{p}_1(\hat{f})\} + \hat{p}_0(\hat{f})\{1 - \hat{p}_0(\hat{f})\}]}{\{\hat{p}_1(\hat{f}) - \hat{p}_0(\hat{f})\}^2}.$$

With survival endpoints subject to censoring, we estimate the hazard rate in each arm. Let $\hat{S}_0(t_0; \hat{f})$ and $\hat{S}_1(t_0; \hat{f})$ be the estimated probability of surviving beyond t_0 for the eligible group taking treatment 0 or 1 respectively. These probabilities can be estimated similarly as outlined in Section 2.2.1. Assuming an exponential distributed survival time, we estimate the hazard function as

$$\hat{\lambda}_a(\hat{f}) = -\log\{\hat{S}_a(t_0; \hat{f})\}/t_0.$$

The required number of events is

$$n_{events} = \frac{(z_\beta + z_{\alpha/2})^2}{\pi_1 \pi_0 \log^2\{\hat{\lambda}_0(\hat{f})/\hat{\lambda}_1(\hat{f})\}},$$

where π_0 and π_1 are the proportions to be allocated to arms 0 and 1. The total sample size can be calculated as

$$\frac{n_{events}}{1 - \{\pi_1 \hat{S}_1(\tau; \hat{f}) + \pi_0 \hat{S}_0(\tau; \hat{f})\}},$$

where τ_0 is the planned length of study.

The above development of sample size calculations assumed estimated treatment effect of the marker is based on observed data. However, in some cases there will not be available phase II or other data for this estimation. In that case models of the outcome as a function of marker for each group can be hypothesized and incorporated into sample size formulas given above.

3. Simulation studies

We conducted simulation studies to assess the proposed Max-Impact design. In all scenarios, we generated multiple biomarkers, where $X = (X_1, X_2)$, with $X_p \sim \mathcal{N}(0, 1)$, $p = 1, 2$. Binary treatments were generated from $\{0, 1\}$ with a probability of 0.5. In Scenario 1, both efficacy and risk outcomes were continuous. The efficacy outcome $Y = 1 - 2X_1 + X_2 + (1 - X_1 - X_2) * A / 3 + \epsilon_Y$, and the risk outcome $R = 2 + X_1 - 2X_2 + (1 + X_1 - X_2)A + \epsilon_R$, where $\epsilon_Y, \epsilon_R \sim \mathcal{N}(0, 1)$. Approximately 56% of patients would gain benefit from treatment. However, among these marker positive patients, 60% would experience an elevated risk. In Scenario 2, both outcomes were binary. The outcomes followed the model $\text{logit}\{P(Y = 1|X, A)\} = \beta_0 + \beta_1 A + X^T \gamma_1 + AX^T \gamma_2$, where $\text{logit}(t) = \log t(1 - t)$, $\beta_0 = -1.5$, $\beta_1 = 0.125$, $\gamma_1 = (0.125, 0.125)^T$ and $\gamma_2 = (-0.5, -1)^T$ for the efficacy outcomes, and $\beta_0 = -1$, $\beta_1 = 1$, $\gamma_1 = (0.125, 0.125)^T$ and $\gamma_2 = (1, -1)^T$ for the risk outcomes. Approximately 54% of patients would gain benefit from treatment, and 78% of them would experience an elevated risk. In Scenario 3, survival efficacy outcomes were generated. The survival time T was generated with hazard rate function

$$\lambda_{\tilde{T}}(t|A, X) = \exp\left\{X_1 + (-0.225 + \text{sign}(-2X_1 + X_2)\sqrt{-0.75X_1 + 0.375X_2})A\right\}.$$

The censoring time C was generated from a uniform distribution on $[0, 25/4]$. We set $t_0 = 1$. The risk outcomes were generated from the binary model with $\beta_0 = -0.5$, $\beta_1 = 1$, $\gamma_1 = (0.125, 0.125)^T$ and $\gamma_2 = (4, -1.5)^T$. In this scenario, the treatment effect was nonlinear, and 66% of the patients would benefit from the treatment. 60% of the marker positive patients would however, had a higher probability of developing risk outcomes under the new treatment.

The models were correctly specified for continuous and binary endpoints, but misspecified for the survival endpoint. The data were generated from cox model with a nonlinear basis, but we fitted the cox model using a linear basis. To assess the model misspecification in survival data fitting, we calculate the mean absolute differences between the estimated and the true survival function over the time period $[0, 1]$, defined by

$$\int_{u \in [0, 1]} \int_x \{|\hat{S}(u|x, a) - S(u|x, a)|\} du dx,$$

when using the linear cox model. The obtained value is 0.050, which is not substantially different from the value we obtained using correct model (a value of 0.043). As a ratio, this

is about a 16% increase in relative error. Hence, the degree of model misspecification is not severe.

The Max-Impact design was compared with the Randomize-All design and the Max-Power design. In the Randomize-All design, we would enroll all patients into a future phase III study without any screening procedures. This design does not depend on whether there are other constraints to consider. The Max-Power design is the targeted design that aims to maximize the power of the phase III study given the same pool of screening patients. Hence, the enrollment rule is the one maximizing $\widehat{ES}(f)^2 \times P\{f(X) > 0\}$ where $\widehat{ES}(f)$ denotes the estimated effect size of the selected group by the rule $I\{f(X) > 0\}$. In particular, $\widehat{ES}(f) = \hat{\theta}(f)$ for continuous outcomes, $\widehat{ES}(f) = \hat{p}_1(f) - \hat{p}_0(f)$ for binary outcomes, and $\widehat{ES}(f) = \log\{\hat{\lambda}_0(\hat{f})/\hat{\lambda}_1(\hat{f})\}$ for survival outcomes, with quantities defined in Section 2.3. If there is an additional imposed constraint, we choose the best enrollment rule among all the rules that satisfy the constraint. To implement the proposed design, we need to estimate $\widehat{ES}(f)$ using the trial data. We utilized linear regression for a continuous endpoint, logistic regression for a binary endpoint, and Cox regression for a survival endpoint.

We investigated different situations where the constraints on the ultimate risk impact varied. Note that the situation without any constraints can be treated as a special case, since the risk constraint becomes ineffective if we set it to a large value. By modifying the level of risk constraints, the derived enrollment rule under the proposed Max-Impact design and the Max-Power design also changed, which subsequently affected the ultimate impact in the population.

We repeatedly generated data from the phase II study 500 times, where the sample size was set to 200. For each dataset, we identified the enrollment rule under different criteria. We also generated a large validation data set of size 10000. Based on this validation set, we calculated the ultimate impact that would be achieved in the population if we carried out a phase III design using the enrollment rule. There were 500 ultimate impact values on the validation set, and we took an average of these values.

We show three comparisons in Tables 1 to 3. We first present the relative increase or decrease of Max-Impact, Max-Power and Optimal design in the average ultimate impact on efficacy compared to the Randomize-All design, against different thresholds on the risk impact constraints. Mathematically, they are calculated as

$$\frac{\mathbb{P}_n[\hat{\Delta}(X)I\{f(X) > 0\}] - \mathbb{P}_n[\hat{\Delta}(X)]}{\mathbb{P}_n[\hat{\Delta}(X)]},$$

where f denotes a general enrollment rule derived using Max-Impact or Max-Power design. In addition, we present the number of patients to be randomized and total cost, with those of the Randomize-All design as a baseline. The total cost is combined of screening and randomization cost, calculated using the estimates from the US NCI's National Clinical Trial Network (NCTN), of which the SWOG Cancer Research Network is a member. Typically, per-patient screening institutional reimbursement is typically approximately 1/4 of

the amount received for patients randomized to treatment, where more intensive requirements including the more extensive data collection are usually needed. In Figure 2, we show the corresponding risk impacts achieved by different methods, where we examine if the desired constraint levels were well controlled at.

The optimal design always led to the greatest impact in efficacy. However, this is accompanied by higher risk in the overall population, and the resulting risk impact usually exceeded the desired level. When the constraint on risk impact was imposed, the proposed design yielded the best result in most situations. The impact decreased as the risk impact constraint became more restrictive. In all scenarios, Max-Power design tended to over control the impact on risk, which in turn led to a lower impact on the efficacy. In certain scenarios, it could reduce the population impact compared with the Randomize-All design. Although the sample sizes and total costs using the Max-Impact design were higher than the Max-Power design as expected, it significantly reduced the requirement from Randomize-All design.

4. Data application

We applied the proposed method to data from SWOG S0421 study, a North American Intergroup phase III trial for men with metastatic castration-resistant prostate cancer. Patients were randomly assigned in a blinded fashion in a 1:1 ratio to docetaxel administered every 21 days at a dose of 75 mg/m² with or without the bone-targeted oral agent atrasentan taken daily for up to 12 cycles (Quinn et al., 2013). 1038 eligible patients were enrolled. Among these patients, 855 submitted serum for the bone biomarkers and 778 patients had usable specimens at baseline. After removing missing observations, the sample size for our analysis was 751, where 371 patients are in the docetaxel + atrasentan arm and 380 patients are in the docetaxel + placebo arm.

We used overall survival as the primary outcome, which could be censored due to loss of follow-up. The risk outcome we considered was the indicator of any adverse event of degree 3 or higher. 7 baseline covariates were considered to inform the enrollment rule, including age (range 40–92), serum prostate-specific antigen (PSA, range 0.1–10414.1), indicator of performance status (2–3 versus 0–1, 56% 2–3), and bone marker levels. Four bone markers were bone alkaline phosphatase (BAP, range 1.9–1761.0 u/L), C-terminal of type 1 collagen (CICP, range 1.4–273.6 ng/mL), N-telopeptides of type 1 collagen (NTx, range 1.4–480.0 nM), and pyridinoline (PYD, range 0.3–15.0 nmol/L). Log transformations were utilized for bone marker concentrations and serum PSA since their distributions were skewed.

We examined the impact that could be achieved using different designs, including the Max-Impact, Max-Power, and Randomize-All designs. The data was partitioned into 3 roughly equalized parts. 2 parts of the data were used as the preliminary phase II data, on which we established the enrollment criterion for a future study using various designs. The impact was computed on the remaining third part using the constructed enrollment rule. The averages of the cross-validation estimated impact from 250 replicates were presented in Figure 3. Since the adverse event rates were similar between two arms, the reduction in risk impacts was limited. The proposed design led to a better overall impact in general. The impact curve was

flat once the constraint on risk impacts exceeded 0.795, because the risk impact achieved by the optimal enrollment rule without any constraints was under 0.795.

5. Discussion

We propose a novel biomarker-based empirically constructed design focusing on optimizing the population impact if a phase III study were subsequently conducted using this design. The design selects the most appropriate subpopulation for a future study, and takes into account the benefits that would be obtained for both on- and off- study patients. The method applies to both continuous and categorical design factors, and can handle various types of outcomes.

We also consider the practical constraint such that the overall impact on the additional risk outcome is controlled, while maintaining the overall impact on the treatment efficacy. A key aspect to consider for an enrichment trial is the total costs, including patient screening costs and per-patient costs for patients in the trial. We can introduce additional conditions on controlling the total cost of the study. The resulting optimization problem will involve nonlinear cost constraints, where we may use nonlinear optimization technique for the solution.

Often, several candidate biomarkers are available, and phase II information could be inadequate to reliably select among them. Hence, there are proposals of phase III designs that adaptively change eligibility criteria during the trial. In other words, the clinical trial begins without restricting entry, and sequentially restricts entry based on the candidate biomarkers in an adaptive manner (Simon and Simon, 2013). We will consider extending the proposed strategy to an adaptive manner, which offers more flexibility in designing the phase III trial.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The work is supported by grants R01DK108073, U10CA180819 and P30 CA015704 awarded by the National Institutes of Health and Colman Early Career Fellowship awarded by Hope Foundation.

References

- Cox D (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, 87–22.
- Hoering A, LeBlanc M, and Crowley JJ (2008). Randomized phase iii clinical trial designs for targeted agents. *Clinical Cancer Research* 14, 4358–4367. [PubMed: 18628448]
- Maitournam A and Simon R (2005). On the efficiency of targeted clinical trials. *Statistics in medicine* 24, 329–339. [PubMed: 15551403]
- Mandrekars SJ and Sargent DJ (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology* 27, 4027–4034. [PubMed: 19597023]

- Quinn DI, Tangen CM, Hussain M, Lara PN, Goldkorn A, Moinpour CM, Garzotto MG, Mack PC, Carducci MA, Monk JP, et al. (2013). Docetaxel and atrasentan versus docetaxel and placebo for men with advanced castration-resistant prostate cancer (swog s0421): a randomised phase 3 trial. *The Lancet Oncology* 14, 893–900. [PubMed: 23871417]
- Renfro LA, An M-W, and Mandrekar SJ (2017). Precision oncology: A new era of cancer clinical trials. *Cancer letters* 387, 121–126. [PubMed: 26987624]
- Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltmane JM, et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine* 346, 1937–1947. [PubMed: 12075054]
- Rubin DB (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6, 34–58.
- Sargent D and Allegra C (2002). Issues in clinical trial design for tumor marker studies In *Seminars in oncology*, volume 29, pages 222–230. Elsevier. [PubMed: 12063675]
- Sargent DJ, Conley BA, Allegra C, and Collette L (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 23, 2020–2027. [PubMed: 15774793]
- Simon N and Simon R (2013). Adaptive enrichment designs for clinical trials. *Biostatistics* 14, 613–625. [PubMed: 23525452]
- Simon R and Maitournam A (2006). Evaluating the efficiency of targeted designs for randomized clinical trials (vol 10, pg 6759, 2004). *Clinical Cancer Research* 12, 3229–3229.
- Splawa-Neyman J, Dabrowska D, and Speed T (1990). On the application of probability theory to agricultural experiments (engl. transl. by d.m. dabrowska and t.p. speed). *Statistical Science* 5, 465–472.
- Wang Y, Fu H, and Zeng D (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association* 113, 1–13. [PubMed: 30034060]

Figure 1. Power curve and the ultimate impact curve of the enrichment clinical trial based on eligibility criterion $X > c$ for a screening size of 400 ($ns = 400$).

Figure 2.
Control of ultimate risk impact under different designs.

Figure 3.
Ultimate impact under different designs.

Table 1

Increment or decrement in proportion relative to Randomize-All design for Scenario 1

	Design	0.4	0.5	0.6	v_0 0.7	0.8	0.9	1
Impact	Max-Impact	0.08	0.12	0.14	0.15	0.16	0.16	0.15
	Max-Power	-0.27	-0.21	-0.17	-0.15	-0.14	-0.13	-0.13
	Optimal	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Sample size	Max-Impact	-0.72	-0.69	-0.66	-0.64	-0.62	-0.61	-0.60
	Max-Power	-0.87	-0.86	-0.85	-0.84	-0.84	-0.84	-0.84
Cost	Max-Impact	-0.60	-0.57	-0.54	-0.51	-0.49	-0.47	-0.47
	Max-Power	-0.76	-0.75	-0.74	-0.74	-0.74	-0.74	-0.73

Table 2

Increment or decrement in proportion relative to Randomize-All design for Scenario 2

	Design	0.08	0.1	0.12	v_0 0.14	0.16	0.18	0.2
Impact	Max-Impact	0.47	0.55	0.61	0.64	0.66	0.67	0.67
	Max-Power	0.25	0.35	0.42	0.45	0.47	0.48	0.48
	Optimal	0.82	0.82	0.82	0.82	0.82	0.82	0.82
Sample size	Max-Impact	-0.85	-0.84	-0.84	-0.83	-0.82	-0.82	-0.81
	Max-Power	-0.93	-0.92	-0.92	-0.91	-0.91	-0.91	-0.91
Cost	Max-Impact	-0.75	-0.75	-0.75	-0.74	-0.73	-0.73	-0.73
	Max-Power	-0.85	-0.85	-0.85	-0.84	-0.84	-0.84	-0.84

Table 3

Increment or decrement in proportion relative to Randomize-All design for Scenario 3

	Design	0.18	0.19	0.20	v_0 0.21	0.22	0.23	0.24
Impact	Max-Impact	0.04	0.08	0.11	0.14	0.17	0.19	0.21
	Max-Power	-0.21	-0.19	-0.18	-0.17	-0.17	-0.17	-0.16
	Optimal	0.41	0.41	0.41	0.41	0.41	0.41	0.41
Sample size	Max-Impact	-0.62	-0.63	-0.64	-0.65	-0.66	-0.67	-0.67
	Max-Power	-0.94	-0.94	-0.93	-0.93	-0.93	-0.93	-0.93
Cost	Max-Impact	-0.46	-0.48	-0.5	-0.52	-0.53	-0.54	-0.54
	Max-Power	-0.88	-0.88	-0.88	-0.88	-0.88	-0.88	-0.88