Research Article

Statistics in Medicine

Received 6 March 2013, Accepted 21 February 2014 Published online 18 March 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6146

Randomized reverse marker strategy design for prospective biomarker validation

Kevin H. Eng[*†]

We describe a novel study design for validating marker-based treatment strategies meant to select among possible therapeutic options using a biologic marker. Studying existing designs in realistic scenarios, we demonstrate that this design is more than four times more efficient for testing the interaction between a marker and its intended treatment. Our analysis employs a simple parametric framework that uncovers systematic biases in currently proposed designs and suggests how they may be accommodated or enumerated. In the context of markers for choosing a treatment for recurrent ovarian cancer, our proposal requires sample sizes on the order of recently completed phases II and III studies making validation studies for this clinical decision scenario viable.

Keywords: biomarker validation; interaction; ovarian cancer; randomized trial; trial design

1. Introduction

Acknowledging that tumor heterogeneity contributes to the variety in response to treatment [1] and noting the rise in the discovery of therapeutics whose response is limited to a subgroup (e.g., gefitinib and epidermal growth factor receptor mutant patients) [2] or in compounds whose beneficial effects are tied to a marker (e.g., tamoxifen and estrogen receptor, Herceptin and Her2/neu), there is a significant interest in finding biomarkers that can be used to target treatments.

Unfortunately, the clinical benefits of promising markers are rarely realized [3]. One reason may be the surprisingly large sample sizes required to test the superiority of a marker-based (MB) treatment versus nonmolecular treatment plans. One example from a prospective, randomized design [4] requires about 1000 patients to detect a hazard ratio of 0.70; in another optimal design, a 13% difference in response rate requires about 500 patients [5]. These large numbers add to the burden of discovery and may make trials prohibitively expensive for rarer cancers.

Even so, validation by multiple independent prospective trials is a necessary step in the development of newly characterized markers [6, 7]. An economical way to test a marker is to test for treatment effects in a specific marker subset following an enriched or targeted design [8]. However, it is also of interest to know whether the marker is relevant beyond a single stratum to quantify the practical, clinical impact of an MB strategy. A desirable design is both randomized and targeted [9]; subsequently, we mean targeted to say that the marker's predictions have been included in the assignment of treatments in the study. The ultimate goal of validation, then, is to interrogate a biomarker strategy: a predictive marker linked to treatments to form a predictive strategy.

For testing strategies, there exists a handful of study designs [4], which either directly test the clinical context of MB treatment or indirectly evaluate the interaction effect. The most efficient design is unclear

Roswell Park Cancer Institute, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14263, U.S.A.

**Correspondence to: Kevin H. Eng, Roswell Park Cancer Institute, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14263, U.S.A.*

†E-mail: kevin.eng@roswellpark.org

as investigations have determined that there is ambiguity about the situations where one design is more efficient than the other [10].

In this article, we will propose a new study design that we call a reverse marker (RM) strategy design. As in existing marker-strategy designs (viz. [4]), it employs a two-arm randomization scheme, provides a direct estimate of the marker-strategy response rate, and evaluates the interaction between the marker and possible treatments. The design appears to use one quarter of the observations required to test the same interaction in likely scenarios. Evaluated in the context of recurrent ovarian cancer, this efficiency makes it possible to conduct studies at effect sizes reported in the literature while previous studies could not. In deriving this new design, we employ a parametric framework to characterize the exact hypotheses under study. This analysis reveals, in all of the designs, an implicit bias that relates to the marginal difference in effect between candidate treatments. Fortunately, it is straightforward to adjust planned studies to accommodate these biases using this framework.

We first introduce three common designs for testing interactions as well as our proposed reverse-marker strategy design (Section 2). A nontechnical discussion of design considerations focuses on the contexts for and against the new design (Section 2.1). Section 3 reviews the parametric framework and makes more technical comparisons. Our ovarian cancer case study in Section 4 illustrates the relative efficiency of the designs and tests their sensitivity to marker prevalence.

2. Randomized designs for prospective biomarker validation

Throughout the article, we consider four designs whose schemas are given in Figure 1; the first three are described in [4], and sample sizes for a class including designs 2 and 3 are discussed in [5]. Design 1 is the stratified, marker-interaction (MI) study and is distinct from the others because it randomizes treatments stratified on marker status. Design 2, the MB strategy, and design 3, the modified MB (MMB)

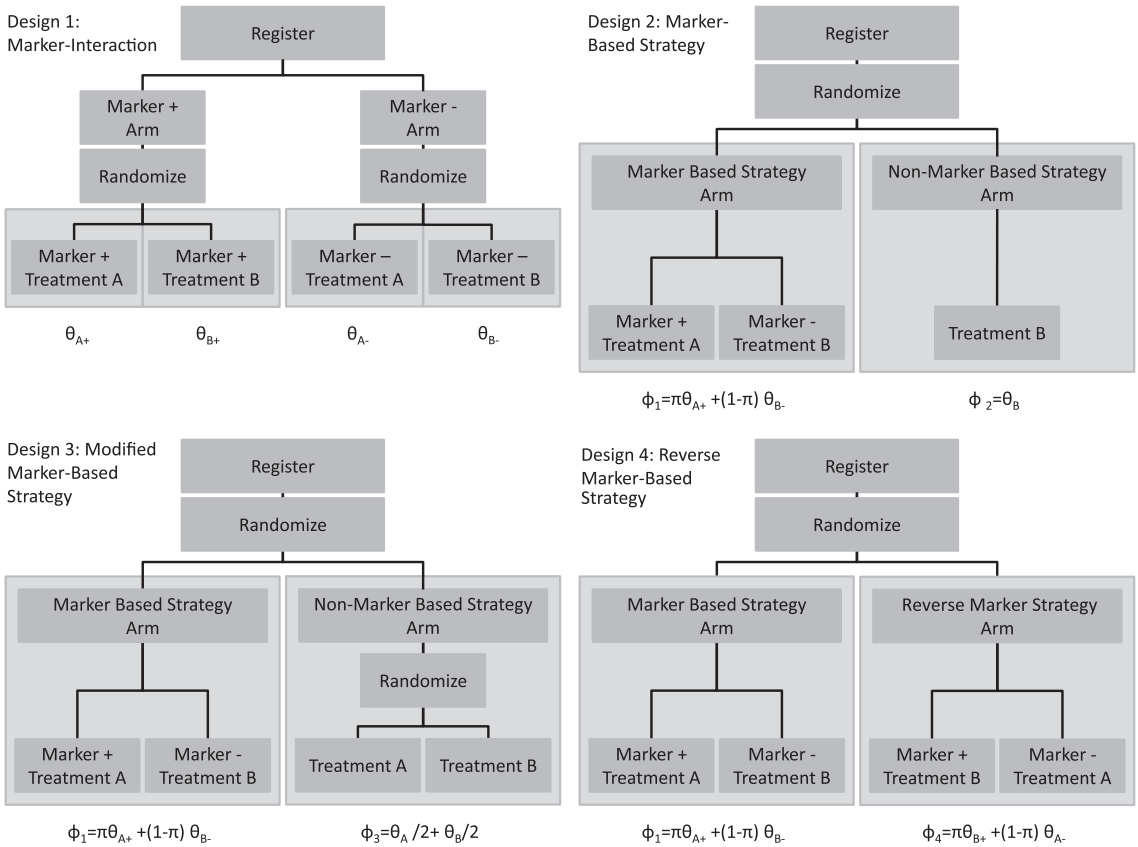

Figure 1. Four designs for marker validation studies. Shaded boxes indicate the arms used in the planned analysis. Parameters below each box are the expected response rate in each arm using the notation in Section 3. Designs 1–3 are described in [4]. Design 4 is a novel proposal.

strategy design, randomize patients to receive treatments following an MB assignment or a standard clinical decision workflow and thus *directly* test the marker-strategy hypothesis. For brevity, the reader is referred to the articles cited earlier for more details.

Design 4, the RM strategy design is a novel proposal. As in the MB and MMB designs, this is a direct marker strategy design with two arms where one arm follows the MB strategy. The complementary arm of the RM design, instead of a default treatment (MB) or a randomly drawn treatment (MMB), tests the reverse treatment hypothesis that M^+ patients should be assigned to treatment B and M^- patients should be assigned to treatment A.

Each of these designs is intended to test the utility of a marker strategy by conducting a simple test of proportions (MB, MMB, RM) or a test of the additive interaction (MI, defined in Appendix 1) between the major arms shaded in Figure 1. Therefore, any statistical guarantees given by the associated test reflect on the qualities of the design, and we will refer to test and design interchangeably.

2.1. *Considerations for testing marker strategies*

Before proceeding into the technical discussion, we summarize the key considerations for the trial designs, pointing to the relevant statistical discussion in Section 3 and case studies in Section 4.

Is our goal to test for a treatment effect in a marker positive subset only? The targeted or enriched designs assay patients first and then select only marker positive patients to study. Typically, the studied subset is based on a biological hypothesis or a drug's mechanism of effect and might be used in cases where a drug has shown little marginal effect in the general population. However, the trial design says nothing about the marker negative patients and may be less relevant in the clinical setting when the marker is rare. One notes that the power and sample sizes for enriched designs [8] must be similar to the situation where a marker is always present because the MI, RM, and MB designs become equivalent to the enriched design. These designs are not fully identical because of their own specific considerations.

Is our goal to test the interaction or the clinical strategy in all the patients? The MI design is designed to test whether the active treatment is unusually effective in marker positive patients. The MB, MMB, and RM designs are powered to test the joint deployment of the marker and active treatment as a strategy (Section 3.1). The difference is that the clinical strategy takes into account the marker prevalence as well as the interaction to characterize the potential impact of the marker strategy on clinical care.

A key difference between the MI (measuring the interaction) and the RM (testing the strategy) designs is whether the measurement of the marker implicitly represents an *intent to treat*. In the MI design, patients can be dropped from study between measurement and randomization. While the control of patients into arms is ideal because some patients may be discarded when the arm closes, this seems to have a mixed benefit because accrual rates may be low if the marker prevalence is extreme. In contrast, all patients are randomized, measured, treated, and evaluated in the RM design. Thus, a quantity of interest to some study designers (especially for rare diseases) may be the expected number of patients who must be assayed in order to reach the required accrual goal. The MB and MMB designs are in a gray area because the marker needs to only be measured in one arm of the trial.

The active treatment should be appropriate for all patients. The MI, MMB, and RM designs assume that the active treatment can be given to marker negative patients. In the MI and RM designs, the marker values for all patients are assessed, so the design might be modified to exclude contraindications: marker negative patients may be reassigned to the control treatment regardless of randomization. In the MB and MMB designs, patients randomized to the non-MB treatment arm have unknown status that may allow assignment to the contraindicated regimen (Section 3.2).

Is there a marginal drug effect? For candidate drugs in search of marker-defined subgroups, it is likely that there is existing evidence of no beneficial effect at a previously tested population level. In general, if there is a significant difference, there seems to be little need to conduct a marker trial. In the situation that a marker trial is warranted, the framework in the next section provides a method for adjusting for expected differences (Section 3.3). In our computational study, the RM sample requirements were robust to variation in the marginal difference (Section 4.3).

How prevalent is the marker? Marker prevalence affects the clinical relevance of a marker strategy. In our computational example, we noted that the RM design was relatively robust to variation in the prevalence (Section 4.2). With respect to the logistics of accruing patients, if the marker is extremely rare, then one may consider enriched or targeted designs that study only the marker positive setting. During the design phase investigators, one may wish to compare the expected accrual rates among different designs.

Statistical efficiency. The power of the RM design dominates the MB and MMB designs in most reasonable cases (Section 3.4). In the simulation study, we consider tests based on the RM and MI designs that are similar but powered for different comparisons (Sections 4.1 Section 4.4).

3. Probability framework for marker strategies

Suppose that binary variable Y represents a patient's response to their randomly assigned treatment, $T \in \{A, B\}$. Let the binary marker under study, $M \in \{M^+, M^-\}$, have level M^+ associated with higher response rates and marker prevalence $P(M = M^+) = \pi$ where $0 < \pi < 1$. We assume that a patient's response depends only on their marker value and the treatment to which they are assigned. That is, $P(Y|T = i, M = j) = \theta_{ij}$ where

$$\theta_{ij} = \beta_0 + \beta_A I(T = A) + \beta_+ I(M = M^+) + \beta_I I(T = A, M = M^+) \tag{1}$$

Here, $I(\cdot)$ is the indicator function, β_0 represents a baseline effect, β_A represents the added effect of treatment A, β_+ the effect of a positive marker, and β_I is a nonadditive effect. The notation is summarized in Table I.

For completeness, we denote the marginal effect of treatment A as $\theta_A = P(Y|T = A)$, likewise for treatment B. It is generally assumed that B is a standard treatment and that A is under study for its increased effect ($\beta_A > 0$) or MB effect ($\beta_I > 0$). Under this framework, the MB strategy under consideration assigns M^+ patients to treatment A and M^- patients to treatment B.

Defining the marginal effect of treatment A over B to be

$$\gamma = P(Y|T = A) - P(Y|T = B) \tag{2}$$

$$= \beta_A + \pi\beta_I \tag{3}$$

where $\gamma = 0$ corresponds to no difference, Section 3.1 will show that existing designs implicitly make assumptions on γ that can lead to anti-conservative analyses.

We state that the marker-strategy validation designs (MB, MMB, RM) intend to test H: $\pi(1-\pi)\beta_I = 0$. We derive this quantity in Section 3.1 by considering exactly what hypothesis is tested by each design. Intuitively, this hypothesis contains both the marker prevalence π and an effect β_I. Notably, β_I focuses on the specificity of the marker effect: a marker that is only prognostic ($\beta_+ > 0$) will not aid treatment; a treatment that is independently superior ($\gamma > 0$) does not necessarily require a marker. The interaction (β_I) captures the information about whether the strategy to assign M^+ patients to treatment A has more merit than its individual components.

Targeted or enriched designs [8] that randomize within M^+ and M^- strata are similar to the MI design [4], except that the targeted design intends to examine only the M^+ arm, testing the null hypothesis that H: $\theta_{A+} = \theta_{B+}$. The MI design is really a test of both arms, H: $\{\theta_{A+} = \theta_{B+}, \theta_{A-} = \theta_{B-}\}$. We emphasize the adjective *marker-strategy* to denote designs meant to test H: $\pi(1-\pi)\beta_I = 0$.

Finally, while we consider binomial responses in this article, a similar argument can be made by considering the same additive model parametrization of the (negative) log hazard of a survival time. The corresponding log-rank test between arms and its sample size computation are based on an equivalent quantity [11], so many of the following results translate directly.

3.1. Expected response rates and hypotheses under consideration

In the MI design, the test of interaction is based on the comparison of treatment effect in each marker arm. It expects to test H: $\Delta_1 = 0$ where

Table I. Summary of mean parametrization and effect parametrization.

Treatment	Marker Status	Mean Notation	Effect Notation
A	M^+	θ_{A+}	$\beta_0 + \beta_A + \beta_+ + \beta_I$
A	M^-	θ_{A-}	$\beta_0 + \beta_A$
B	M^+	θ_{B+}	$\beta_0 + \beta_+$
B	M^-	θ_{B-}	β_0

$$\Delta_1 = (\theta_{A+} - \theta_{B+}) - (\theta_{A-} - \theta_{B-}) \tag{4}$$

$$= (\beta_A + \beta_I) - \beta_A \tag{5}$$

$$= \beta_I \tag{6}$$

We give a statistic testing the linear interaction in the Appendix. One might alternatively test the differences by constructing the 2×2 table of responders given treatment arm and marker status and then using Fisher's exact test, the test of proportions or a χ^2-test; the power of such a test depends on the odds ratio and is inconvenient in this notation.

The planned analyses of the MB, MMB, and RM designs each test the difference in response rates in the two arms. Let $\phi_1 = \pi E(Y|M = M^+, T = A) + (1-\pi)E(Y|M = M^-, T = B)$ be the expected response to an MB strategy. The non-marker strategy arm has expected response rate $\phi_2 = E(Y|T = B)$ in the MB design; $\phi_3 = E(Y|T = A)/2 + E(Y|T = B)/2$, in the MMB design; and $\phi_4 = \pi E(Y|M = M^+, T = B) + (1-\pi)E(Y|M = M^-, T = A)$, in the RM design.

By comparing the response rate across these two arms, the designs test $H : \Delta_k = 0$ where $\Delta_k = \phi_1 - \phi_k$ for $k = 2, 3, 4$ (MB, MMB, and RM). It can be shown that the expected differences are as follows:

$$\Delta_2 = \pi(1-\pi)\beta_I + \pi\gamma \tag{7}$$

$$\Delta_3 = \pi(1-\pi)\beta_I + (\pi - 1/2)\gamma \tag{8}$$

$$\Delta_4 = 2\pi(1-\pi)\beta_I + 2(\pi - 1/2)\gamma \tag{9}$$

In the case that $\gamma = 0$, it is sensible that the three direct designs test the MB strategy effect, $\pi(1-\pi)\beta_I$, as this depends on the prevalence of the marker as well as the expected interaction effect.

The distinction between the indirect and direct designs is evident here: the direct designs account for the clinical utility of the marker and treatment. In the case that the marker is very common or very rare, the likelihood of a situation that may be adjudicated by a marker is low and $\pi(1-\pi)\beta_I$ reflects this.

We expect that given the same number of patients, the RM design will have more power to detect deviations from $H : \pi(1-\pi)\beta_I = 0$ than the MMB design, because it has twice the signal ($\Delta_4 = 2\Delta_3$); simulation studies easily demonstrate this effect in supplemental material. Relative sample sizes calculations follow in Section 3.4.

This amplification of effect comes from observing that $\phi_3 = \phi_1/2 + \phi_4/2$. This means that between the two arms in the modified MB strategy design, *ceteris paribus*, half the patients would have received the same treatment regardless of randomization. The RM design arises by recognizing that we can adjust the randomization point to minimize redundancy.

3.2. Treatment assignment frequency and balance

An informative comparison is to consider how each design assigns patients to the four possible groups: AM^+, AM^-, BM^+, and BM^-. The expected fractions are summarized in Table II. We observe that all combinations of treatment and marker are possible in the MI, MMB, and RM designs. In these cases, the investigator must be prepared to use all of the possible levels.

The MI and RM designs have identical assignment rates and can be applied in similar situations. We will see later that the designs are not fully equivalent because the MI design is powered to test treatment effects in each marker arm separately, while the RM design ought to be powered for the interaction hypothesis directly.

Note that the marginal frequency of treatment assignments in the MB and MMB designs depends on marker prevalence. When π is very small, this dependence may lead to inefficiency: if M^+ is rare, the MB strategy is largely concordant with the non-MB strategy. In contrast, the MI and RM designs show a marginal sense of balance by assigning treatments in equal weight, invariant to π. For the MI design, this is a result of the stratified approach that treats marker groups as cohorts and does not use the marker to supervise treatment assignment. The RM design's balance comes because marker values should be evenly randomized across arms.

Table II. Expected fraction of patients assigned to treatment groups A and B by design.

	Fraction Assigned						
	Marginal		Marker and Treatment				Probability of
Design	A	B	A, M^+	A, M^-	B, M^+	B, M^-	same treatment
1 (MI)	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\pi}{2}$	$\frac{1-\pi}{2}$	$\frac{\pi}{2}$	$\frac{1-\pi}{2}$	n/a
2 (MB)	$\frac{\pi}{2}$	$1-\frac{\pi}{2}$	$\frac{\pi}{2}$	0	$\frac{\pi}{2}$	$1-\pi$	$1-\pi$
3 (MMB)	$\frac{1}{4}+\frac{\pi}{2}$	$\frac{3}{4}-\frac{\pi}{2}$	$\frac{3\pi}{4}$	$\frac{1-\pi}{4}$	$\frac{\pi}{4}$	$\frac{3(1-\pi)}{4}$	$\frac{1}{2}$
4 (RM)	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\pi}{2}$	$\frac{1-\pi}{2}$	$\frac{\pi}{2}$	$\frac{1-\pi}{2}$	0

π is the prevalence of M^+ markers. The last column refers to the probability that the same treatment is assigned regardless of randomization to marker-strategy arm or not.
MI, marker interaction; MB, market based; MMB, modified marker based; RM, reverse marker.

3.3. *Unequal treatment effects*

When $\gamma \neq 0$, the guarantee on type I control in the test of H: $\pi(1-\pi)\beta_I = 0$ may be affected. Simply, the test may be invalid if $\beta_I = 0$ does not always imply that $\Delta_k = 0$. Without loss of generality, we may consider the $\gamma > 0$ case.

In the MB design, $\{\beta_I = 0, \gamma > 0\}$ implies that $\pi\gamma > 0$, so the design is always anti-conservative and will falsely identify an interaction more often than it should. This occurs because only the marker-strategy arm receives the superior ($\gamma > 0$) treatment, which is aliased with marker effect [4].

In the MMB and RM designs, $\{\beta_I = 0, \gamma > 0\}$ implies that $\Delta_k = 0$ only when $\pi = 1/2$, corresponding to even chance of assignment to the superior treatment. Because they decrease Δ_3, rarer markers ($\pi < 1/2$) will make the test more conservative, while more prevalent markers will make the test anti-conservative. While concerning, the latter case is less relevant: highly prevalent markers in the presence of unequal treatments represents a case where biomarker-mediated treatment is likely to be redundant.

Thus, we have identified the bias, which can be accounted for during study design by estimates from prior studies. If we adjust for the bias, the test becomes the two-sample test that the response proportions differ by the bias, namely, H: $\Delta_k = (\pi - 1/2)$ for designs k=3, 4 (MMB, RM). Note that the MB design cannot be adjusted (if $\gamma \neq 0$ and $\Delta_2 = 0$, then π must be zero).

3.4. *Relative efficiency of designs for testing a marker-strategy hypothesis*

We consider the relative sample sizes required by each design. Because the MI design is intended to test treatment effects in each stratified arm before testing for the interaction, it is recommended to power each arm separately [8, 12]. The formula is listed in the Appendix.

Given a particular set of response rates, $\theta_{A+}, \theta_{A-}, \theta_{B+}, \theta_{B-}$, a target level ($\alpha$) and power ($1-\beta$), the sample sizes for the MB, MMB, and RM designs are computed as follows. Under one-to-one randomization, let n_k be the number of patients required in each arm in design $k > 1$. As in the previous section, suppose that ϕ_1 is the expected response rate in the marker arm and $\phi_k, k > 1$ is the rate in other arm under design k, where $\Delta_k = \phi_1 - \phi_k$. The required sample size for each arm in a test of proportions between the two randomization arms is

$$n_k = \frac{(z_\alpha + z_{1-\beta})^2[\phi_1(1-\phi_1) + \phi_k(1-\phi_k)]}{\Delta_k^2} \tag{10}$$

Where z_α is the αth quantile of a standard normal distribution. The relative sample size required by designs 3 and 4 is

$$\frac{n_3}{n_4} = \frac{\Delta_4^2}{\Delta_3^2}\left(\frac{\phi_1(1-\phi_1) + \phi_3(1-\phi_3)}{\phi_1(1-\phi_1) + \phi_4(1-\phi_4)}\right) \tag{11}$$

$$= 4\left(\frac{\phi_1 + \phi_3 - \phi_1^2 - \phi_3^2}{\phi_1 + \phi_4 - \phi_1^2 - \phi_4^2}\right) \tag{12}$$

So, if $\phi_4 = \phi_3$ or $\phi_4 = 1 - \phi_3$, then $n_3/n_4 = 4$; the MMB design uses four times more subjects than the RM design. It can be shown that in general, $\phi_4 < \min\{\phi_3, 1-\phi_3\}$ and $\phi_4 > \max\{\phi_3, 1-\phi_3\}$ assure $n_3/n_4 > 4$.

When ϕ_1 is close to zero or one, the relative efficiency is most sensitive to ϕ_3 and ϕ_4. Inversely, $\phi_1 = 1/2$ minimizes their effect. So, the parenthetical term in Equation (12) may be bounded by two ratios:

$$r_0 = \left(\frac{\phi_3 - \phi_3^2}{\phi_4 - \phi_4^2}\right) \tag{13}$$

$$r_1 = \left(\frac{1/4 + \phi_3 - \phi_3^2}{1/4 + \phi_4 - \phi_4^2}\right) \tag{14}$$

As a function of ϕ_4, the ratio r_0 is lowest at $\phi_4 = 1/2$ by concavity, so the zeroes of the parabola $\phi_3^2 - \phi_3 + 1/16 = 0$ mark a boundary: for $1/2 - \sqrt{3}/4 < \phi_3 < 1/2 + \sqrt{3}/4$ (approximately, $0.07 < \phi_3 < 0.93$) and $0 < \phi_4 < 1$, $r_0 \geq 1/4$ and $n_3/n_4 \geq 1$. In the other direction, considering $\phi_1 = 1/2$ and r_1, a similar argument shows that $r_1 > 1/4$ for all values of ϕ_3 and ϕ_4. These represent a conservative boundary as there are more extreme cases where the ratio is larger than 1/4.

In summary, the RM design is more efficient than the MMB design for $\left\{(\phi_3, \phi_4) : 1/2 - \sqrt{3}/4 < \phi_3 < 1/2 + \sqrt{3}/4, 0 < \phi_4 < 1\right\}$ and is more than four times more efficient for $\phi_4 < \min\{\phi_3, 1-\phi_3\}$ or $\phi_4 > \max\{\phi_3, 1-\phi_3\}$.

4. Recurrent ovarian cancer study planning

Our specific motivation comes from the validation of biomarkers meant to guide maintenance treatment of recurrent, advanced ovarian cancer. These cancers respond to initial platinum treatment but commonly relapse and, through a cycle of serial treatments, become increasingly platinum resistant [13]. While several approved chemotherapies are available [14], the best recurrent alternative to platinum treatment is unclear [15].

This treatment decision is presently guided by previous response to therapy that is coarse, is variable, and requires an intervention to evaluate [1]. The use of genomic biomarkers offers an individually relevant guide [16], but these quantities will need to be evaluated through prospective study. As such, our general intent is to consider the characteristics of study designs that will be employed in the next phases of research.

A review of platinum-resistant cancers in phases II and III studies without markers [17] reports sample sizes ranging from 27 to 254 total patients in single and double arm trials with response rates ranging from 0.06 to 0.18 for single agent and 0.22 to 0.40 for double agent therapies (each representing a different clinical context). These numbers are consistent with reviews citing a 0.10–0.20 response, regardless of treatment, in previously treated platinum-resistant cancers [1].

Subsequently, we study the required sample sizes (Equations (10) and (16)) as a function of β_I and π, and we discuss tests of interaction versus stratification. The intention is to illustrate the use of the designs in study planning and is not meant to be comprehensive of all (β_I, π) scenarios. Throughout the section, we consider sample sizes for tests at level $\alpha = 0.05$ and power $1 - \beta = 0.80$.

4.1. Sample sizes for interaction tests

Denoting typical frontline treatment, a platinum and taxane, as treatment B, we imagine that we have a marker with prevalence $\pi = 0.5$, which is predictive in platinum/taxol treated patients: high-marker values have a response rate of 0.10, and low-marker values have a response rate of 0.50 (the marginal rate is 0.30). Obviously, a marker that simply tells us that some patients will not respond to treatment has an important, but limited, value. Thus, we imagine a search for a treatment combination that improves response in the high-marker patients.

Table IIIA parameterizes our scenario. The platinum/taxol combination is listed with fixed response rates. The response rate of an alternative active treatment (treatment A) is parameterized by β_I. The marginal rates are fixed at 0.30, so $\gamma = 0$ for all β_I. Note that while the response rate for AM^+ patients

Table III. Recurrent ovarian cancer treatment example scenarios used in Section 4.

Treatment	Response rate M^+	M^-	Population
A. Parametrization under $0 < \beta_I < 0.5$, $\pi = 0.5$, $\gamma = 0$			
A	$0.10 + \beta_I$	$0.50 - \beta_I$	0.30
B	0.10	0.50	0.30
B. $\beta_I = 0.20$ used for $\pi \neq 0.5$ study			
A	0.30	0.30	$0.30\pi + 0.30(1-\pi)$
B	0.10	0.50	$0.10\pi + 0.50(1-\pi)$
C. $\beta_I = 0.20$, $\pi = 0.6$ used for $-0.3 < \gamma < 0.7$			
A	$0.30 + \gamma$	$0.30 + \gamma$	$0.30 + \gamma$
B	0.10	0.50	0.30

Fixed values are taken from literature.

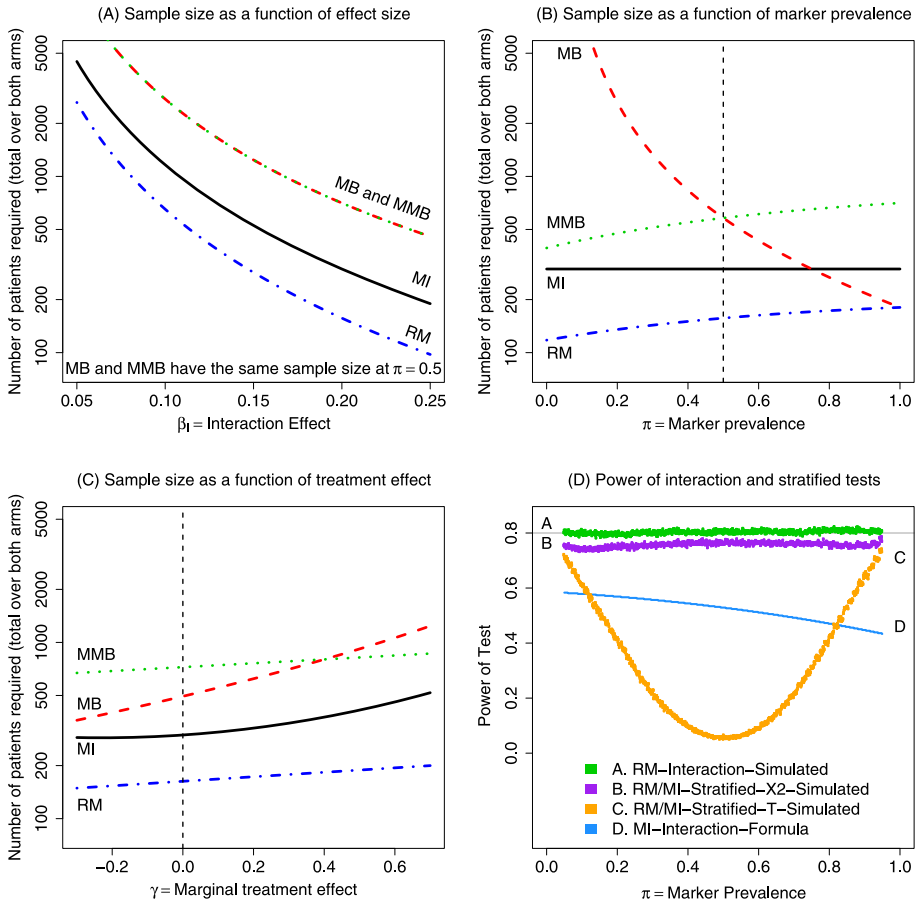

Figure 2. (A)–(C) Required sample sizes for ovarian cancer scenarios outlined in Table III. The vertical lines highlight the $\beta_I = 0.2$, $\pi = 0.5$, and $\gamma = 0$ scenarios where the marker is uninformative in one and predictive in the other treatment. (D) Power of interaction and stratification tests given the computed sample size for the reverse marker design.

may be lower than AM$^-$ patients, as long as $\beta_I > 0$, marker positive patients respond to the active treatment better than treatment B. Figure 2(A) shows the required sample sizes as a function of $0 < \beta_I < 0.25$. We observe that the MB and MMB designs have the same sample size requirement when $\pi = 0.5$ and are less efficient than the MI and RM designs (as expected per Section 3.4). The RM design is about

twice as efficient as the MI design that may be attributed to the fact that the MI design divides patients four ways (AM^+, AM^-, BM^+, BM^-), while the RM design requires only two (marker strategy and the reverse).

We consider what effect sizes can be detected using the roughly 200 patients accrued in the previously described second-line ovarian studies. For the MI design, 200 patients require $\beta_I = 0.24$ (a 0.34 response rate in M^+ and 0.26 in M^- patients); similarly, the RM design with 200 patients could detect a $\beta_I = 0.18$ scenario (0.28 vs 0.32 response rates). These effects are not inconsistent with the meta-reviews [17]. Subsequently, we consider the $\beta_I = 0.20$ case where the MI design calls for 298 and the RM design calls for 158 patients, which is close to the largest reported study sizes [18].

4.2. *Sample sizes as a function of prevalence*

We select the $\beta_I = 0.2$ scenario for further study because it calls for a realistic number of patients ($n = 158$) and has a reasonable magnitude of effect. Table IIIB reparameterizes the scenario in terms of the marker prevalence π. Note that in this baseline scenario, the marker has no effect in treatment A, but it does indicate that for M^+ patients, A is the better treatment.

In this case, the RM design dominates all of the others for all prevalence values so the efficiency gain seen in Figure 2(A) is invariant to π. Noting that the MB and RM designs have the same sample size requirement at $\pi = 1$, by design, these must be equivalent to the enriched or targeted design where only M^+ patients are randomized to treatment A or B.

The sample size for the MI design does not vary as we may simply close the arm when the required number of patients is accrued. The MB design requires fewer patients as π increases, concordant with the idea that more patients are being assigned to the active treatment. Further, there does exist a range of π where the MB design is more economical than the MI design (as reported in [10]), although this is likely to be an unrealistic clinical scenario.

4.3. *Sample sizes as a function of marginal treatment effect*

We reparameterize the $\beta_I = 0.2$ case again in Table IIIC to depend on $-0.3 < \gamma < 0.7$, the marginal treatment effect. Because Δ_4 depends on $\gamma(\pi - 1/2)$, we select the $\pi = 0.6$ case to avoid the insensitivity of RM and MMB designs on γ. In Figure 2(C), the MMB and RM sample sizes are still surprisingly insensitive to γ, while the MI design's sample size requirement increases mildly as the marginal effect increases.

4.4. *Power to test stratified differences*

We consider the $\beta_I = 0.2$ scenario as a function of π again (Table IIIB) and fix the sample size at the targeted value required for the RM design (n_4). Under this scenario, we compute the power for various RM-based and MI-based tests relative to the 80% RM interaction (marker strategy) test and plot the results in Figure 2(D). To compute power by simulation, we generated 10,000 datasets and reported the fraction of tests significant at level $\alpha = 0.05$.

Power for the MI interaction test can be evaluated by formula: fixing $n = n_4$, we invert the sample size formula to obtain the power. Notably, the test has consistently lower power than the RM interaction test at n_4 for all values of π.

We have implemented both a stratified t-test and a χ^2-test (goodness of fit). The former tests that one treatment is superior in both arms (separately), so the $\pi = 0.5$ case is a null scenario (and power approaches the appropriate target 0.05). The goodness of fit test is powered against the range of π, but both are consistently underpowered versus the interaction test.

5. Conclusion

We have proposed a new design for a randomized prospective marker validation study that is significantly more efficient for testing marker strategies than existing designs in scenarios motivated by our ovarian cancer work. Using literature-based estimates of the available sample sizes, we determine that this design is a step toward making these studies possible where previous designs have made them logistically

infeasible. Pragmatically, for situations where a randomly selected treatment has a better than 7% response rate, the RM design is more efficient than the MMB designs, and we have given bounds for when it is more than four times more efficient.

This design is balanced: randomization frequencies for each treatment are equal independent of marker prevalence. While the MI design balances treatments without using the marker in treatment assignment, the RM design maintains balance and implements an MB strategy mimicking the clinical workflow. Both of these properties have a place in phased biomarker development.

While it is difficult to match specific designs to specific situations abstractly, we have provided some guidance on the considerations for marker strategy versus interaction designs as well as a set of parameters to consider during the design phase. It is of primary importance that the study designer is clear on what quantity best represents their research question: the interaction, the marker strategy, or the subgroup treatment effect.

Appendix A

A.1. Interaction test for marker interaction design

There are a couple of inequivalent choices for testing the interaction in the MI design. We have opted to use the direct test under an additive model as in the succeeding text; however, the interaction might also be tested by laying the responders into a 2×2 table and conducting any of the usual tests of independence: effectively testing whether the responders accrue at a differential rate. The latter approach has power determined by the multiplicative odds ratio and not the additive interaction and is harder to compare with the two sample proportion tests used in the other designs.

In design 1, we accrue n_+ marker M^+ patients and n_- marker M^- patients separately. Each arm is randomized separately. Let the number of responders in each treatment arm be $(Y_{A+}, Y_{A-}, Y_{B+}, Y_{B-})$ where each of these counts is binomial: $Y_{ij} \sim$Binomial$(n_j/2, \theta_{ij})$ for $i \in \{A, B\}$, $j \in \{-, +\}$. Defining $p_{ij} = Y_{ij}/(n_j/2)$, the interaction test is simply

$$Z_n = \frac{(p_{A+} - p_{B+}) - (p_{A-} - p_{B-})}{\sqrt{\frac{p_{A+}(1-p_{A+})+p_{B+}(1-p_{B+})}{n_+/2} + \frac{p_{A-}(1-p_{A-})+p_{B-}(1-p_{B-})}{n_-/2}}} \tag{15}$$

Following a normal approximation of binomial proportions argument, the distribution of Z_n can be read in a standard normal table.

A.2. Sample size for design 1

The required sample size to separately power the marker $+$ and marker $-$ arms of the MI design (design1) is as follows:

$$n_1 = 2(z_\alpha + z_{1-\beta})^2 \left\{ \frac{\theta_{A+}(1-\theta_{A+}) + \theta_{B+}(1-\theta_{B+})}{(\beta_A + \beta_I)^2} + \frac{\theta_{A-}(1-\theta_{A-}) + \theta_{B-}(1-\theta_{B-})}{(\beta_A)^2} \right\} \tag{16}$$

noting that factor 2 arises because of the two treatment arms for each marker level. This formula does not reflect the rate at which M^+ and M^- patients accrue, so when $\pi \neq 0.5$, one arm may close before the other.

A.3. Comment on survival time response

In the case of survival times, assuming proportional hazards and the same interaction notation as before for the log hazard, we can directly apply the results as Δ_k represents the difference in log hazard ratio. As in [11], where patients are randomized 1:1 to each arm, the total number of events required is

$$2n_k = \frac{4(z_\alpha + z_{1-\beta})^2}{\Delta_k^2} \tag{17}$$

and so is similar to the binary case.

Acknowledgements

The author thanks Rick Chappell for his encouragement and comments and an anonymous reviewer for their comments. This work was supported by Roswell Park Cancer Institute and National Cancer Institute grant P30 CA016056.

References

1. Vaughan S, Coward J, Bast R, Berchuck A, Berek J, Brenton J, Coukos G, Crum C, Drapkin R, Etemadmoghadam D, Friedlander M, Gabra H, Kaye SB, Lord CJ, Lengyel E, Levine DA, McNeish IA, Menon U, Mills GB, Nephew KP, Oza AM, Sood AK, Stronach EA, Walczak H, Bowtell DD, Balkwill FR. Rethinking ovarian cancer: recommendations for improving outcomes. *Nature Reviews Cancer* 2011; **11**(10):719–725.
2. Thatcher N, Chang A, Parikh P, Rodrigues Pereira J, Ciuleanu T, von Pawel J, Thongprasert S, Tan E, Pemberton K, Archer V, Carroll K. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (iressa survival evaluation in lung cancer). *The Lancet* 2005; **366**(9496):1527–1537.
3. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer* 2005; **5**(11):845–856.
4. Sargent D, Conley B, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; **23**(9):2020–2027.
5. Tang L, Zhou X. A general framework of marker design with optimal allocation to assess clinical utility. *Statistics in Medicine* 2013; **32**(4):620–630.
6. Pepe M, Feng Z, Janes H, Bossuyt P, Potter J. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *Journal of the National Cancer Institute* 2008; **100**(20):1432–1438.
7. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology* 2009; **27**(24):4027–4034.
8. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 2005; **24**(3):329–339.
9. Freidlin B, McShane L, Korn E. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute* 2010; **102**(3):152–160.
10. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal of Biopharmaceutical Statistics* 2009; **19**(3):530–542.
11. Schoenfeld D. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**(2):499–503.
12. Mandrekar S, Grothey A, Goetz M, Sargent D. Clinical trial designs for prospective validation of biomarkers. *American Journal of Pharmacogenomics* 2005; **5**(5):317–325.
13. Cooke S, Brenton J. Evolution of platinum resistance in high-grade serous ovarian cancer. *The Lancet Oncology* 2011; **12**(12):1169–1174.
14. Bookman M. Standard treatment in advanced ovarian cancer in 2005: the state of the art. *International Journal of Gynecological Cancer* 2005; **15**:212–220.
15. Bhoola S, Hoskins W. Diagnosis and management of epithelial ovarian cancer. *Obstetrics & Gynecology* 2006; **107**(6):1399–1410.
16. Na Y, Farley J, Zeh A, del Carmen M, Penson R, Birrer M. Ovarian cancer: markers of response. *International Journal of Gynecological Cancer* 2009; **19**(11):S21–S29.
17. Ushijima K. Treatment for recurrent ovarian cancer – at first relapse. *Journal of Oncology* 2010; **2010**:ID497429.
18. O'Malley D, Azodi M, Makkenchery A, Tangir J, McAlpine J, Kelly M, Schwartz P, Rutherford T. Weekly topotecan in heavily pretreated patients with recurrent epithelial ovarian carcinoma. *Gynecologic Oncology* 2005; **98**(2):242–249.

