On the Use of Marker Strategy Design to Detect Predictive Marker Effect in Cancer Immunotherapy and Targeted Therapy

Yan Han[1], Ying Yuan[2], Sha Cao[1], Muyi Li[3,4] and Yong Zang[1,*]

[1]Department of Biostatistics, Indiana University, USA

[2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA

[3]The Wang Yanan Institute for Studies in Economics (WISE), China

[4]Department of Statistics, School of Economics, Xiamen University, China

*Author for Correspondence: Yong Zang, PhD, Department of Biostatistics, Indiana University. 410 W 10th street, Indianapolis, Indiana, 46202; email: zangy@iu.edu

The marker strategy design (MSGD) has been proposed to assess and validate predictive markers for targeted therapies and immunotherapies. Under this design, patients are randomized into two strategies: the marker-based strategy, which treats patients based on their marker status, and the non-marker-based strategy, which randomizes patients into treatments independent of their marker status in the same way as in a standard randomized clinical trial. The strategy effect is then tested by comparing the response rate between the two strategies and this strategy effect is commonly used to evaluate the predictive capability of the markers. We show that this commonly used between-strategy test is flawed, which may cause investigators to miss the opportunity to discover important predictive markers or falsely claim an irrelevant marker as predictive. Then we propose new procedures to improve the power of the MSGD to detect the predictive marker effect. One is based on a binary response endpoint; the second is based on survival endpoints. We conduct simulation studies to compare the performance of the MSGD with the widely used marker stratified design (MSFD). Numerical studies show that the MSGD and MSFD has comparable performance. Hence, contrary to popular belief that the MSGD is an inferior design compared with the MSFD, we conclude that using the MSGD with the proposed tests is an efficient and ethical way to find predictive markers for targeted therapies.

This is the author's manuscript of the article published in final edited form as:

Han, Y., Yuan, Y., Cao, S., Li, M., & Zang, Y. (2020). On the Use of Marker Strategy Design to Detect Predictive Marker Effect in Cancer Immunotherapy and Targeted Therapy. Statistics in Biosciences, 12(2), 180–195. https://doi.org/10.1007/s12561-019-09255-1

1 Introduction

The emergence of immunotherapy and targeted therapy has revolutionized the era of clinical oncology[1,2]. One of the biggest challenges of immunotherapy is that it typically benefits only a subgroup of patients.[3] As a result, optimizing the treatment benefit of immunotherapy requires the identification of the predictive biomarker that can be used to foretell the differential efficacy of the immunotherapy based on the presence or absence of the marker, e.g., pembrolizumab is approved by the Food and Drug Administration for the treatment of advanced melanoma and metastatic squamous and nonsquamous non-small cell lung cancer (NSCLC) whose tumors express programmed death ligand-1 (PD-L1), i.e., PD-L1 positive patients.

Several novel biomarker-guided clinical trial designs have been proposed to achieve this goal[4,5,6,7,8,9,10]. Among them, the marker strategy design (MSGD) has been proposed as a useful trial design for identifying and validating predictive markers[4,5,6]. As shown in part (a) of Figure 1, the MSGD randomizes patients to two strategies, namely, the marker-based strategy and the non-marker-based strategy. Patients randomized to the marker-based strategy are treated (deterministically) based upon their biomarker statuses (e.g., patients with a marker-positive status receive the targeted treatment and those with a marker-negative status receive the standard treatment). Patients randomized to the non-marker-based strategy are further randomized to different treatments independent of their marker statuses. Although measuring the biomarker profiles of patients randomized to the non-marker-based strategy is not required, in practice we often do so, prospectively or retrospectively, for the purpose of biomarker discovery and other correlation studies. In this article, we assume that the biomarker is measured for all patients in the trial. A series of clinical trials[11,12,13] has adopted the MSGD for evaluating and validating predictive marker effects. For example, by using the MSGD, the excision repair cross-complementing 1 (ERCC1) trial[11] found that

the ERCC1 mRNA expression level might be a predictive marker for treating non-small cell lung cancer (NSCLC) patients with docetaxle plus gemcitabine (the p-value = 0.02) based on 444 patients with stage-IV NSCLC.

In addition to the MSGD, another biomarker-guided clinical trial design which has been widely used to identify and validate the predictive marker is the marker stratified design (MSFD)[4,5,6]. As shown in part (b) of Figure 1, the MSFD stratifies patients into different subgroups based on the patients' biomarker profile and then randomizes the patients to receive either the targeted treatment or the standard treatment within each subgroup. Under the MSFD, the predictive biomarker effect is typically evaluated by comparing the difference in the treatment effects within the marker-positive subgroup to those within the marker-negative subgroup[14]. Under the MSGD, however, the most common approach to test the predictive marker is to compare the response rate (or hazard for survival outcome) between the marker-based and non-marker-based strategies using a t test (or log-rank test). If the response rate of the marker-based strategy is significantly higher than that of the non-marker-based strategy, the marker is claimed as the predictive marker. Mandrekar and Sargent[5] and Freidlin et al.[15] noted that the between-strategy test has low statistical power to detect the predictive biomarker effect because a certain proportion of patients will receive the same treatment regardless of their assignment to the marker-based or non-marker-based strategies (e.g., some patients with a marker-positive status in both strategies will receive the targeted treatment), thereby diluting the differences between the two treatment strategies. Therefore, it is generally believed that the MSGD design is an inferior design compared with the MSFD[15].

In this article, we argue that the primary interest of MSGD is to evaluate the between-strategy effect, which does not necessarily equal to the predictive effect defined in the MSFD. Therefore, it is unfair to directly compare MSGD with MSFD as these two designs target for different objects. Actually, if the predictive marker effect rather than the between-

strategy effect is the primary interest of a clinical trial, we prove in the following content that the commonly used between-strategy test by MSGD is indeed problematic. After that, we propose a new test to evaluate the true predictive marker effect under MSGD. Finally, we conduct simulation studies to compare MSGD with MSFD under the same definition of predictive marker. Our simulation results reveal that contrary to popular belief, the MSGD is not an inferior design and has plausible performance compared with the MSFD.

Our study is motivated by a colorectal cancer trial, which is being conducted at the Indiana University Melvin and Bren Simon Cancer Center. The biomarker used in this trial is the KRAS gene mutation. The MTA is a novel KRAS inhibitor and the standard treatment is radiotherapy. This trial is conducted under the MSGD. A total of 210 patients with colorectal cancer are equally randomized to either the non-marker-based strategy and marker-based strategy. Patients in the non-marker-based strategy are further equally randomized to receive either the MTA or the standard treatment. Patients in the marker-based strategy are treated according to their KRAS gene status. The patients without the KRAS gene mutation receive the standard treatment whereas the patients with the KRAS gene mutation receive the MTA. The purpose of this trial is to evaluate whether the KRAS gene is a predictive marker for patients with colorectal cancer. As the commonly used between-strategy test is problematic in detecting the predictive effect, novel test is required to evaluate such effect, which inspires the research for this article.

2 Deficiency of the between-strategy test

We first use two numerical examples to illustrate that the between-strategy test adopted by the MSGD is fundamentally flawed to detect the predictive marker effect. Suppose that the patient population of interest consists of 20% marker-positive ($M+$) patients and 80% marker-negative ($M-$) patients. Assume that for the standard treatment, the response

rates for the $M+$ and $M-$ patients are the same, at a value of 0.4; and for the targeted treatment under investigation, the response rates for the $M+$ and $M-$ patients are 0.8 and 0.5, respectively. Clearly, M is a predictive marker because the $M+$ patients respond to the targeted treatment substantially more favorably than the $M-$ patients. Now, we look at the response rate in the marker-based strategy and the non-marker-based strategy. As summarized in Table 1, in the marker-based strategy, $M+$ patients are assigned to the targeted treatment, and $M-$ patients are assigned to the standard treatment. Thus, the overall average response rate for the marker-based strategy is $20\% \times 0.8 + 80\% \times 0.4 = 0.48$. In the non-marker-based strategy, patients are equally randomized into the standard and targeted treatments. The average response rate is $(0.8 + 0.4)/2 = 0.6$ for the $M+$ patients, and $(0.5 + 0.4)/2 = 0.45$ for the $M-$ patients. Thus, the overall average response rate for the non-marker-based strategy treatment arm is $20\% \times 0.6 + 80\% \times 0.45 = 0.48$, which is the same as that of the marker-based strategy! This means that we will completely miss the predictive marker effect if we take the approach of the commonly used between-strategy test.

The between-strategy test can also mislead investigators to falsely conclude that a marker is predictive when it actually is not. To see this, consider a case similar to the above example, but now the marker is not predictive, with the response rate of the targeted treatment being the same (0.1) for both the $M+$ and $M-$ patients. In this case, as shown in Table 1, the overall average response rate in the marker-based strategy is 0.34, higher than the overall response rate in the non-marker-based strategy (i.e., 0.25). If we use the between-strategy test, we will draw an incorrect conclusion that the marker is predictive.

Mathematically, the deficiency of the between-strategy test stems from the fact that the treatment effect evaluated by the between-strategy test is actually not the predictive marker effect, except under certain restrictive conditions, as described in Theorem 1. The proof is provided in Appendix.

Theorem 1 Let ϕ be the marker positive prevalence. For the binary endpoint, the between-strategy Z test is valid for testing the predictive marker effect only when $\phi = 0.5$; and for the time-to-event endpoint, the between-strategy log-rank test is valid for testing the predictive marker effect only when (1) there is no treatment effect or (2) there is no prognostic effect and $\phi = 0.5$.

3 New tests for detecting the predictive marker effect

3.1 Binary endpoint

In this section, we describe new procedures that are generally valid for the MSGD to detect the predictive marker effect. We first consider the binary response outcome. Let p_{jk} denote the response rate for patients with marker status k who are receiving treatment j, where $k = +/-$ denotes marker-positive/-negative, and $j = 1/0$ denotes the targeted/standard treatment. The treatment effects of the targeted agent (with respect to the standard treatment as a control) are given by $p_{1+} - p_{0+}$ and $p_{1-} - p_{0-}$ for $M+$ and $M-$ patients, respectively. The predictive marker effect is defined as $\theta = (p_{1+} - p_{0+}) - (p_{1-} - p_{0-})$, i.e., the difference in the treatment effect between $M+$ and $M-$ patients, with $\theta = 0$ representing that the marker is not predictive. We notice that this definition has also been used by the MSFD to define the predictive marker effect[4,14]. Our goal here is to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We also aware that the definition of the predictive marker effect is not unique. Indeed, the predictive marker effect can also be defined as a treatment-marker interaction term in a logistic model[16,17], which is beyond the scope of this article.

Let $\hat{p}_{jk} = m_{jk}/n_{jk}$ denote the observed response rate for patients with marker status k who are receiving treatment j, where n_{jk} is the number of patients having marker status k who are receiving treatment j, and m_{jk} is the number of responses among n_{jk} patients. We

propose to evaluate the predictive marker effect for the MSGD using the following Z test,

$$Z = \frac{(\hat{p}_{1+} - \hat{p}_{0+}) - (\hat{p}_{1-} - \hat{p}_{0-})}{\sqrt{\frac{\hat{p}_{1+}(1-\hat{p}_{1+})}{n_{1+}} + \frac{\hat{p}_{0+}(1-\hat{p}_{0+})}{n_{0+}} + \frac{\hat{p}_{1-}(1-\hat{p}_{1-})}{n_{1-}} + \frac{\hat{p}_{0-}(1-\hat{p}_{0-})}{n_{0-}}}},$$

which asymptotically follows a standard normal distribution under the null that there is no predictive marker effect. Given a significance level of α, we declare that M is a predictive marker if $|Z| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution.

It can be shown that under the alternative hypothesis $H_1 : \theta = \theta_1$, Z asymptotically follows a non-central normal distribution $N(\tau, 1)$, where

$$\tau = \frac{2\sqrt{n}\theta_1}{\sqrt{3\phi p_{1+}(1-p_{1+}) + \phi p_{0+}(1-p_{0+}) + (1-\phi)p_{1-}(1-p_{1-}) + 3(1-\phi)p_{0-}(1-p_{0-})}}.$$

Given the type I error α, the power of the test under H_1 is given by

$$\begin{aligned} Pr(|Z| > |\Phi^{-1}(\alpha/2)|) &= Pr(Z > -\Phi^{-1}(\alpha/2)) + Pr(Z < \Phi^{-1}(\alpha/2)) \\ &= \Phi(\Phi^{-1}(\alpha/2) + |\tau|) + \Phi(\Phi^{-1}(\alpha/2) - |\tau|) \\ &\approx \Phi(\Phi^{-1}(\alpha/2) + |\tau|). \end{aligned}$$

Hence, to achieve the power of $1-\beta$, we require $\Phi^{-1}(\alpha/2) + |\tau| = \Phi^{-1}(1-\beta)$, leading to the following sample size formula

$$\begin{aligned} n &= \frac{1}{4\theta_1^2}[\Phi^{-1}(1-\beta) - \Phi^{-1}(\alpha/2)]^2[3\phi p_{1+}(1-p_{1+}) + \\ &\quad \phi p_{0+}(1-p_{0+}) + (1-\phi)p_{1-}(1-p_{1-}) + 3(1-\phi)p_{0-}(1-p_{0-})]. \end{aligned}$$

As most sample size calculations, the value of n depends on a variety of parameters, such as p_{1+}, p_{1-}, p_{0+} and p_{0-}. The values of these parameters can be estimated from historical data or provided by investigators based on their domain knowledge. If such prior information is not available, a pilot study may be needed to obtain initial estimates of the parameters.

3.2 Survival endpoint

We now turn to the survival endpoints (e.g., progression-free survival or overall survival). Let λ_{jk} denote the hazard rate for the patients with $D = j$ and $M = k$, and $\theta_+ = \log(\lambda_{1+}/\lambda_{0+})$ and $\theta_- = \log(\lambda_{1-}/\lambda_{0-})$ denote the log hazard ratio between the targeted treatment and standard treatment for the $M+$ patients and $M-$ patients, respectively. That is, θ_+ and θ_- respectively represent the treatment effect of the targeted agent (with respect to the standard treatment as the control) for the $M+$ patients and $M-$ patients. Then, the predictive marker effect can be defined as $\theta = \theta_+ - \theta_-$, with $\theta = 0$ representing no predictive marker effect. We are interested in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

Let $\tilde{Z}_+$ and $\tilde{Z}_-$ denote the standard log-rank test statistics of comparing the targeted treatment versus the standard treatment for $M+$ and $M-$ patients, respectively; and let ϕ denote the prevalence of $M+$ patients. We propose to test the predictive marker effect using the following weighted log rank test,

$$\tilde{Z} = \sqrt{1-\phi}\tilde{Z}_+ - \sqrt{\phi}\tilde{Z}_-.$$

The asymptotic distribution of $\tilde{Z}$ is described in Theorem 2. The proof is provided in Appendix.

Theorem 2 Let Δ be the total number of events. Test statistic $\tilde{Z}$ asymptotically follows $N(0,1)$ under $H_0 : \theta = 0$ (i.e., no predictive marker effect), and follows $N(\frac{\sqrt{3(1-\phi)\phi\Delta}\theta_1}{4}, 1)$ under $H_1 : \theta = \theta_1$.

Along the same line as the binary endpoint, given the type I error α and type II error β, it can be shown that the sample size formula for the survival endpoint is

$$\Delta = \frac{16[\Phi^{-1}(1-\beta) - \Phi^{-1}(\alpha/2)]^2}{3\theta_1^2\phi(1-\phi)}.$$

Under the MSGD, patients are randomized into two strategies. As a result, the between-strategy comparison is free of the influence of unmeasured confounders because the effects

of the confounders are balanced out between the two strategies through randomization. One may be concerned as to whether the proposed tests are subject to the influence of unmeasured confounders. The result is described in Theorem 3.

Theorem 3 The proposed tests are free from the influence of unmeasured confounders, and thus the predictive marker effect evaluated by the proposed tests is a causal effect.

The proof of Theorem 3 is provided in the Appendix.

4 Simulation Study

We carried out simulation studies to compare the performance of the proposed approaches with the commonly used between-strategy test under the MSGD. We considered three cases: (1) the marker has no predictive effect, which corresponds to the null scenario of no predictive marker effect; (2) the marker has only the predictive effect; and (3) the marker has both predictive and prognostic effects. The prognostic effect is a type of marker effect that is not affected by the treatment, e.g., tumor stage is often a prognostic marker, and patients with higher stages have poor outcomes, regardless of the treatment. Our purpose of the simulation was to evaluate the predictive marker effect only. Hence, case (1) was used to evaluate the empirical type I error rate, and cases (2) and (3) were used to evaluate the empirical power. Under each of the simulation configurations, we conducted 10,000 simulated trials to evaluate the type I error rate and power, with a nominal level of 5%.

Table 2 shows the results for the binary response outcome. The between-strategy test generally led to inflated type I error rates except when $\phi = 0.5$. For example, when $\phi = 0.3$, the response rate of the standard treatment is 0.1 for $M+$ and $M-$ patients, the response rate of the targeted treatment is 0.4 for $M+$ and $M-$ patients, and the type I error rate was inflated to 17.8%. In contrast, the proposed Wald test consistently yielded type I error

rates around the nominal level of 5%. In terms of power, the proposed test significantly outperformed the between-strategy test. The power gain ranged from 30% to 50%, depending on the size of the predictive marker effect. For example, when the true response rate of the standard treatment is 0.2 for $M+$ and $M-$ patients and the true response rates of the targeted treatment are 0.6 and 0.1 for $M+$ and $M-$ patients, given that the prevalence of the $M+$ status is 30%, the power of the proposed test is 88.8%, while that of the between-strategy test is merely 33.5%.

Table 3 shows the results for the survival endpoint. We use the exponential distribution to generate the survival endpoint and specify a 20% censoring rate for each patient. The simulation results for the survival endpoint were similar to those for the binary outcomes. That is, the between-strategy test inflated the type I error rate except when the $M+$ prevalence was $\phi = 0.5$, while the proposed test consistently yielded reasonable type I error rates close to the nominal value of 5%. Compared to the between-strategy comparison, the power of the MSGD often more than doubled when using the proposed test.

In addition to the MSGD, the MSFD can also be used to evaluate the predictive marker effect and it is popular belief that MSFD is much more powerful than the MSGD. However, we argue that such conclusion is arbitrary because the original MSGD actually evaluate the between-strategy effect. Therefore, to make a fair comparison, we conducted simulation studies to compare the MSGD with MSFD by using the same test proposed in this paper. That is, both designs were targeted for the same predictive marker effect. Also, the between-strategy test was also used for the MSGD for the purpose of power comparison. In addition to the power evaluation, we also reported the number of response (for the binary response) and the median survival month (for the survival outcome) to investigate the individual ethics of these two designs.

Table 4 summarizes the simulation results for the binary response outcome. In terms of power comparison, the MSGD is less powerful mainly because the between-strategy test

used. For example, given $\phi = 0.3$ and $n = 200$, when the true response rate of the standard treatment is 0.2 for $M+$ and $M-$ patients and the true response rates of the targeted treatment are 0.4 and 0.1 for $M+$ and $M-$ patients, if the between-strategy test is used, the MSGD is 42.9% less powerful than the MSFD. On the other hand, if the proposed test is used, then the MSGD is only 7% less powerful. Moreover, although the MSGD was still 5% to 10% less powerful than the MSFD with the proposed method, this design gets around 6 to 12 more patients response to the treatment, indicting the MSGD a more ethical design. This is because the MSGD allocates patients to more effective treatments based on their biomarker profiles in the marker-based strategy arm, thereby enhancing the ethics of the trial. As a tradeoff, the randomization in the MSGD is less balanced than the MSFD, resulting in a slight power loss. The simulation results in Table 5 for the survival outcome were similar to those in Table 4. When the proposed method is used, the MSGD was only slightly less powerful than the MSFD, but the median survival month for the MSGD was 2 to 10 months longer. Hence, these two designs yield comparable performance and the MSGD is particularly useful when the predictive marker effect is large. That is because, with a large effect size, both the MSGD and MSFD should be able to identify the predictive marker but the MSGD can benefit more patients enrolled in the trial.

As a side note, our results also indicate that the criticism that the MSGD is an inefficient design with low power to detect predictive markers[5,15] is not completely valid. Low statistical power is not an inherent deficiency of the MSGD design itself, but simply caused by the use of an inappropriate statistical method (i.e., between-strategy test). When adopting the proposed test procedures, the MSGD can have significantly higher power to detect predictive markers.

5 Conclusion

The MSGD has been used in clinical trials to evaluate predictive marker effects. In this article, we show that, under the MSGD, the commonly used between-strategy test for assessing the predictive marker effect is fundamentally flawed. Such an approach not only suffers from low statistical power, but also potentially misleading results, e.g., falsely declaring that a marker is predictive when it is actually not. We propose new tests to be used with the MSGD for detecting the predictive marker effects. Numerical studies show that the proposed tests are generally valid and substantially more powerful than the between-strategy tests. Equipping the MSGD with the proposed tests provides clinicians a powerful design to detect predictive marker effects. Our simulation results also show that compared with the MSFD, the true power reduction by using the MSGD is at most 10% but the MSGD is a more ethical design. Therefore, we conclude that the MSGD is not an inferior design and is especially useful when the predictive marker effect is large. The choice between the MSFD and MSGD depends on the trial setting and objectives. If power is of the biggest concern, the MSFD might be preferred. If investigators are interested in evaluating the real-world effect of the targeted therapy (i.e., the benefit of personalizing treatment by patient's biomarkers versus treating patients without using their biomarkers), the MSGD is clearly the choice. In addition, as the personalized treatment component of the MSGD may increase patient enrollment and retention, the MSGD is an attractive option when patient accrual is difficult, in particular given that the power loss of the MSGD is generally minor.

We have focused on the case in which the marker is measured for all patients prospectively or retrospectively. In principle, the MSGD does not require the measurement of the marker for the patients randomized to the non-marker-based strategy. If this is the case, we can extend our methods to accommodate the missing marker information, for example, using the expectation-maximum algorithm. These extensions are statistically more involved and

will be discussed elsewhere.

In conclusion, on the basis of the results of our study, the common approach of using the between-strategy test to detect predictive markers is problematic and has caused the misconception that the MSGD is an inefficient design with low statistical power. By using the proposed testing procedures, the MSGD provides a powerful and ethical clinical trial design to detect predictive markers.

Appendix

(A) Proof of Theorem 1

We consider the binary endpoint first. For the equally randomized MSGD, the response rate for the marker-based strategy is $p_{1+}\phi + p_{0-}(1-\phi)$ and the response rate for the non-marker-based strategy is $0.5[(p_{1+} + p_{0+})\phi + (p_{1-} + p_{0-})(1-\phi)]$. Hence, defining θ^* as the between-strategy difference, it can be expressed as $\theta^* = 0.5[(1-\phi)\theta + (2\phi-1)(p_{11}-p_{01})]$ and the conventional between-strategy method indeed tests the hypothesis $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$, since in general we have $p_{11} \neq p_{01}$. As a result, when $\theta = 0$, $\theta^* = 0$ only if $\phi = 0.5$. That is, the between-strategy test is statistically valid only when the restrictive condition $\phi = 0.5$ holds.

Similarly, for the survival endpoint, the hazard ratio at time t under the marker-based strategy is

$$\frac{\lambda_{00}e^{-\lambda_{00}t}(1-\phi) + \lambda_{11}e^{-\lambda_{11}t}\phi}{e^{-\lambda_{00}t}(1-\phi) + e^{-\lambda_{11}t}\phi}$$

and the hazard ratio at time t under the non-marker-based strategy is

$$\frac{0.5\lambda_{00}e^{-\lambda_{00}t}(1-\phi) + 0.5\lambda_{01}e^{-\lambda_{01}t}\phi + 0.5\lambda_{10}e^{-\lambda_{10}t}(1-\phi) + 0.5\lambda_{11}e^{-\lambda_{11}t}\phi}{0.5e^{-\lambda_{00}t}(1-\phi) + 0.5e^{-\lambda_{01}t}\phi + 0.5e^{-\lambda_{10}t}(1-\phi) + 0.5e^{-\lambda_{11}t}\phi}.$$

When $\theta = 0$, these two hazard ratios are equivalent only if (1) there is no treatment effect or (2) there is no prognostic effect and $\phi = 0.5$. Therefore, for the survival endpoints, the

between-strategy test is valid to detect the predictive marker effect only if one of these two restrictive conditions hold.

(B) Proof of Theorem 2

For the survival endpoint, defining Δ as the total number of events, according to Schoenfeld[18], we have $\tilde{Z}_- \sim N(\frac{\sqrt{3(1-\phi)\Delta}\theta_-}{4}, 1)$ and $\tilde{Z}_+ \sim N(\frac{\sqrt{3\phi\Delta}\theta_+}{4}, 1)$. Hence, $\tilde{Z} = \sqrt{1-\phi}\tilde{Z}_+ - \sqrt{\phi}\tilde{Z}_-$ has the following asymptotic distribution

$$\tilde{Z} \sim N(\frac{\sqrt{3(1-\phi)\phi\Delta(\theta_+ - \theta_1)}}{4}, 1) = N(\frac{\sqrt{3(1-\phi)\phi\Delta\theta}}{4}, 1).$$

Then, under the null hypothesis $H_0 : \theta = 0$, we have $\tilde{Z} \sim N(0, 1)$, and under $H_1 : \theta = \theta_1$, $\tilde{Z} \sim N(\frac{\sqrt{3(1-\phi)\phi\Delta}\theta_1}{4}, 1)$.

(C) Proof of Theorem 3

Note that the predictive marker effect $\theta = (p_{1+} - p_{0+}) - (p_{1-} - p_{0-})$, where $(p_{1+} - p_{0+})$ and $(p_{1+} - p_{0+})$ are respectively the treatment effects of the targeted treatment for the $M+$ and $M-$patients, with respect to the standard treatment as a control. In order to show that θ is free of the influence of unmeasured confounders, we just need to show that both $(p_{1+} - p_{0+})$ and $(p_{1-} - p_{0-})$ are free of the influence of unmeasured confounders. We first look at $(p_{1+} - p_{0+})$. Because patients are randomized into two strategies, the $M+$ patients assigned to the marker-based strategy (denoted as $\mathcal{P}_1$) should be comparable to the $M+$ patients assigned to the non-marker-based strategy (denoted as $\mathcal{P}_0$) in the sense that all unmeasured confounders are balanced between $\mathcal{P}_1$ and $\mathcal{P}_0$. We use $\mathcal{P}_1 \sim \mathcal{P}_0$ to denote that $\mathcal{P}_1$ and $\mathcal{P}_0$ are comparable. In the non-marker-based strategy, patients are further randomized into the targeted treatment and standard treatment. That is, $\mathcal{P}_0$ is further randomized into the targeted treatment and standard treatment, denoted as $\mathcal{P}_{0t}$ and $\mathcal{P}_{0s}$, respectively, with $\mathcal{P}_{0t} \cup \mathcal{P}_{0s} = \mathcal{P}_0$. Because of randomization, $\mathcal{P}_{0t} \sim \mathcal{P}_{0s} \sim \mathcal{P}_0$. Because

$\mathcal{P}_1 \sim \mathcal{P}_0$, it follows that $\mathcal{P}_{0t} \cup \mathcal{P}_1 \sim \mathcal{P}_{0s}$. As a result, the comparison of the response between the targeted and standard treatments based on $\mathcal{P}_{0t} \cup \mathcal{P}_1$ (i.e., $M+$ patients received the targeted treatment) versus $\mathcal{P}_{0s}$ (i.e., $M+$ patients received the standard treatment), i.e., $(\hat{p}_{1+} - \hat{p}_{0+})$, is free from the influence of unmeasured confounders. By the same argument, the comparison of the response between the targeted and standard treatments based on $M-$ patients, i.e., $(\hat{p}_{1-} - \hat{p}_{0-})$, is also free from the influence of unmeasured confounders. Thus, the proposed Z test is free from the influence of unmeasured confounders. Along the same line, it follows that the proposed weighted log-rank test is also free from the influence of unmeasured confounders.

Acknowledgments

The authors thank three referees for their valuable comments. The research of Yong Zang is partial supported by the design and biostatistics program pilot grant, Indiana CTSI and NIH P30 grant CA082709. The research of Muyi Li is partial supported by the NSFC (No. 71671150) and Fujian Key Laboratory of Statistical Science.

Table 1: Examples to illustrate the deficiency of the between-strategy test.

	Example 1: marker is predictive				Example 2: marker is not predictive			
	Marker-based strategy		Non-marker-based strategy		Marker-based strategy		Non-marker-based strategy	
Treatment	M+ (20%)	M- (80%)	M+ (20%)	M- (80%)	M+ (20%)	M- (80%)	M+ (20%)	M- (80%)
Targeted	0.8	N/A	**0.8**	**0.5**	0.1	N/A	**0.1**	**0.1**
Standard	N/A	0.4	0.4	0.4	N/A	0.4	0.4	0.4
Average	0.8	0.4	0.6	0.45	0.1	0.4	0.25	0.25
Overall	0.48		0.48		0.34		0.25	

Table 2: Type I error rate and power (%) of the MSGD for evaluating the predictive marker effect under the proposed approach (pro.) and between-strategy (str.) comparison when the outcome is a binary response endpoint and $n = 200$.

True response rate				Prevalence of $M+$									
Standard		Targeted		30%		50%		70%		45%		55%	
M+	M-	M+	M-	str.	pro.	str.	pro.	str.	pro.	str.	pro.	str.	pro.
No predictive (type I error rate)													
0.1	0.1	0.2	0.2	7.2	4.9	5.3	5.1	7.3	5.1	5.5	5.2	5.2	4.9
0.1	0.1	0.3	0.3	11.6	5.2	5.4	5.1	10.9	5.2	5.5	5.1	5.9	5.3
0.1	0.1	0.4	0.4	17.8	4.8	5.2	4.9	15.9	5.3	6.0	5.0	6.0	5.3
Predictive only (power)													
0.2	0.2	0.4	0.1	20.1	52.1	23.3	63.3	26.7	58.4	22.3	61.1	24.0	62.6
0.2	0.2	0.5	0.1	26.7	73.8	35.2	83.8	43.7	80.6	33.2	83.1	36.3	84.4
0.2	0.2	0.6	0.1	33.5	88.8	46.9	95.2	62.4	92.4	43.5	94.6	50.9	95.4
0.2	0.2	0.7	0.1	40.1	95.9	60.4	98.9	77.8	98.1	55.1	98.9	64.2	99.1
Predictive + prognostic (power)													
0.2	0.4	0.3	0.1	46.7	74.4	34.5	83.5	23.4	76.8	37.7	81.7	32.4	82.7
0.2	0.4	0.4	0.1	52.9	88.1	47.1	94.3	41.2	90.6	48.9	93.7	45.8	94.6
0.2	0.4	0.5	0.1	61.1	95.7	59.3	98.7	58.8	97.1	60.1	98.6	58.8	98.5
0.2	0.4	0.6	0.1	68.9	98.9	72.0	99.8	75.9	99.3	70.7	99.6	73.2	99.7

Table 3: Type I error rate and power (%) of the MSGD for evaluating the predictive marker effect under the proposed approach (pro.) and between-strategy (str.) comparison when the outcome is a survival endpoint and $n = 200$.

True hazard				Prevalence of $M+$									
Standard		Targeted		30%		50%		70%		45%		55%	
M+	M-	M+	M-	str.	pro.	str.	pro.	str.	pro.	str.	pro.	str.	pro.
No predictive (type I error rate)													
0.5	0.5	0.25	0.25	13.7	4.7	4.8	4.6	12.7	5.3	4.9	4.5	5.5	5.3
0.5	0.5	0.15	0.15	26.1	5.0	5.0	5.1	22.4	5.2	6.1	4.8	5.8	4.5
0.3	0.3	0.10	0.10	26.7	5.4	5.4	5.2	17.5	4.9	6.2	4.5	6.6	5.3
Predictive only (power)													
0.5	0.5	0.38	0.75	17.9	41.4	17.4	49.0	13.8	42.3	15.7	48.1	16.3	46.4
0.5	0.5	0.25	0.75	24.1	79.7	35.0	87.5	35.5	78.6	30.7	85.2	31.3	85.7
0.5	0.5	0.19	0.75	32.1	93.1	42.8	96.8	51.2	93.2	41.0	96.5	47.6	96.0
0.5	0.5	0.15	0.75	38.9	98.0	48.2	99.2	62.1	97.6	48.0	99.1	54.9	99.3
Predictive + prognostic (power)													
0.5	0.4	0.36	0.60	21.6	46.2	20.3	53.2	17.6	43.0	20.0	50.9	20.0	52.4
0.5	0.4	0.30	0.60	22.0	65.5	23.8	73.0	27.5	62.2	27.4	70.1	27.1	70.9
0.5	0.4	0.24	0.60	28.8	79.2	35.8	89.1	44.2	81.7	34.3	89.5	37.3	86.3
0.5	0.4	0.16	0.60	39.2	96.6	50.0	99.3	58.2	96.7	45.4	98.8	52.6	98.6

Table 4: Type I error rate, power (%) and the number of response (in brackets) of the MSGD and MSFD for evaluating the predictive marker effect under the between-strategy test and proposed test when the outcome is a binary response endpoint.

True response rate				Prevalence of $M+$								
Standard		Targeted		30%			50%			70%		
M+	M-	M+	M-	MSGD		MSFD	MSGD		MSFD	MSGD		MSFD
				strategy	proposed		strategy	proposed		strategy	proposed	
n=200, No predictive												
0.3	0.2	0.4	0.3	7.3	5.2(56.0)	5.2(54.0)	5.0	5.3(60.1)	5.1(59.9)	7.4	4.9(65.9)	5.0(63.9)
n=200, Predictive only												
0.2	0.2	0.4	0.1	19.4	55.3(45.5)	62.3(38.9)	23.4	62.5(52.5)	73.7(45.0)	25.9	57.9(59.6)	70.7(51.0)
0.2	0.2	0.5	0.1	27.2	73.8(49.9)	83.6(41.8)	34.6	83.4(60.2)	93.1(50.0)	44.9	80.2(70.0)	91.5(58.0)
0.2	0.2	0.6	0.1	33.2	88.1(54.4)	95.5(44.9)	47.6	95.2(67.5)	98.9(54.9)	61.9	93.0(80.4)	98.4(64.9)
n=200, Predictive + prognostic												
0.4	0.2	0.6	0.1	17.9	44.2(57.5)	57.3(51.1)	20.0	56.5(72.4)	68.2(64.9)	24.1	54.7(87.5)	66.2(78.9)
0.4	0.2	0.7	0.1	23.7	67.6(62.0)	81.3(53.9)	30.9	80.0(80.0)	91.3(69.9)	40.4	77.3(98.1)	90.3(86.0)
0.4	0.2	0.8	0.1	30.3	86.2(66.5)	95.8(57.0)	43.2	94.4(87.5)	98.8(74.9)	60.5	91.6(108.6)	98.4(92.8)
n=100, No predictive												
0.3	0.2	0.4	0.3	6.3	4.9(28.0)	4.8(26.9)	5.7	5.2(29.9)	5.2(29.9)	6.3	5.5(33.0)	4.9(32.0)
n=100, Predictive only												
0.2	0.2	0.4	0.1	13.5	30.8(22.7)	35.7(19.5)	15.0	36.5(26.3)	44.2(22.5)	16.2	31.7(29.7)	42.0(25.5)
0.2	0.2	0.5	0.1	16.1	46.0(25.0)	55.5(20.9)	20.9	55.3(30.0)	67.1(25.0)	25.6	51.4(35.1)	64.4(29.0)
0.2	0.2	0.6	0.1	19.6	60.7(27.3)	73.8(22.4)	27.6	72.6(33.7)	84.6(27.4)	37.1	67.7(40.2)	83.4(32.5)
n=100, Predictive + prognostic												
0.4	0.2	0.6	0.1	11.7	26.1(28.8)	32.6(25.5)	13.2	31.2(36.2)	40.4(32.5)	15.1	29.6(43.8)	39.3(39.5)
0.4	0.2	0.7	0.1	14.2	39.2(30.9)	52.7(27.0)	18.3	51.0(40.0)	64.8(35.0)	24.2	47.9(49.0)	62.0(43.0)
0.4	0.2	0.8	0.1	18.3	56.8(33.2)	74.6(28.5)	25.4	69.0(43.8)	85.2(37.5)	36.4	66.3(54.2)	82.7(46.4)

Table 5: Type I error rate, power (%) and median survival month (in brackets) of the MSGD and MSFD for evaluating the predictive marker effect under the between-strategy test and proposed test when the outcome is a survival endpoint.

True hazard				Prevalence of $M+$								
Standard		Targeted		30%			50%			70%		
M+	M-	M+	M-	MSGD		MSFD	MSGD		MSFD	MSGD		MSFD
				strategy	proposed		strategy	proposed		strategy	proposed	
n=200, No predictive												
0.5	0.5	0.25	0.25	14.0	5.0(27.1)	4.8(25.2)	5.2	4.8(27.1)	5.2(26.9)	12.1	4.9(29.0)	5.3(26.9)
n=200, Predictive only												
0.5	0.5	0.38	0.75	18.2	42.1(19.2)	51.8(17.3)	16.2	47.6(20.4)	60.0(18.7)	15.5	42.3(21.8)	51.6(19.9)
0.5	0.5	0.25	0.75	27.9	78.6(20.6)	89.0(18.2)	32.1	85.6(23.5)	93.8(20.4)	37.4	78.7(26.6)	88.3(22.8)
0.5	0.5	0.19	0.75	35.1	93.2(21.8)	98.3(19.0)	43.3	97.1(25.9)	98.6(21.6)	52.0	93.6(31.0)	98.0(25.0)
n=200, Predictive + prognostic												
0.4	0.5	0.24	0.60	14.5	42.9(22.1)	52.1(20.6)	18.0	48.4(25.2)	58.7(23.0)	22.3	42.8(28.8)	51.3(25.9)
0.4	0.5	0.16	0.60	22.9	79.8(23.8)	89.1(21.6)	33.6	85.9(28.8)	93.8(25.0)	43.8	79.7(35.5)	89.2(29.5)
0.4	0.5	0.12	0.60	29.4	93.6(25.0)	98.2(22.1)	43.9	96.8(31.4)	98.6(26.4)	57.2	93.3(41.0)	98.1(31.9)
n=100, No predictive												
0.5	0.5	0.25	0.25	9.2	4.3(27.0)	4.5(25.3)	4.5	5.1(27.1)	4.6(26.6)	9.0	5.0(29.2)	4.6(27.0)
n=100, Predictive only												
0.5	0.5	0.38	0.75	9.4	26.8(19.4)	28.0(17.5)	10.5	27.8(20.1)	31.7(18.6)	8.7	25.7(21.8)	26.3(20.0)
0.5	0.5	0.25	0.75	17.9	53.1(20.4)	60.4(18.2)	20.6	62.3(23.6)	66.9(20.6)	23.1	55.5(26.6)	61.3(22.7)
0.5	0.5	0.19	0.75	18.4	70.3(21.8)	77.0(19.0)	26.1	80.5(25.6)	87.1(21.3)	31.5	72.9(31.1)	78.5(24.8)
n=100, Predictive + prognostic												
0.4	0.5	0.24	0.60	8.3	28.9(22.2)	29.6(20.4)	12.4	31.4(25.3)	34.5(23.1)	13.2	25.8(28.9)	28.4(25.9)
0.4	0.5	0.16	0.60	14.4	56.9(23.9)	61.1(21.4)	19.2	61.7(28.6)	67.7(24.8)	26.6	52.7(35.3)	62.1(29.2)
0.4	0.5	0.12	0.60	17.8	70.4(25.1)	78.2(22.3)	28.0	78.8(31.4)	86.6(26.4)	32.0	71.1(41.2)	81.5(32.1)

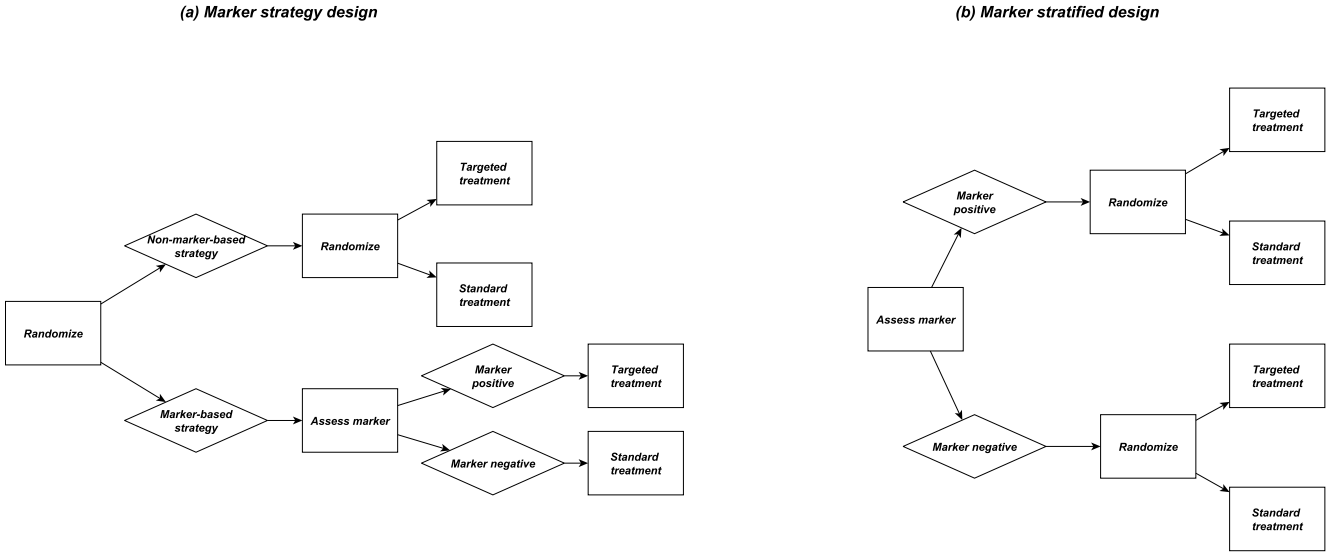

Figure 1: Diagram of the marker strategy design (MSGD) and marker stratified design (MSFD).

References

[1] Couzin-Frankel J. (2013), Cancer immunotherapy. *Science*, 324, 1432-1433.

[2] Sawyers C: Targeted cancer therapy. *Nature* 432: 294-297, 2004.

[3] Kaufman, H.L. (2015), Precision immunology: the promise of immunotherapy for the treatment of cancer. *Journal of Clinical Oncology*, 33, 1315-1317.

[4] Sargent DJ, Conley BA, Allegra C, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 23: 2020-2027, 2005.

[5] Mandrekar SJ, Sargent DJ: Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol* 27: 4027-4034, 2009.

[6] Sargent DJ, Allegra C: Issues in clinical trial design for tumor marker studies. *Semin Oncol* 29: 222-230, 2002.

[7] Simon R, Maitournam A: Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 10: 6759-6763, 2004.

[8] Freidlin B, Simon R: Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 11: 7872-7878, 2005.

[9] Jiang W, Freidlin B, Simon, R: Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst* 99: 1036-1043, 2007.

[10] Freidlin B, Jiang W, Simon, R: The cross-validated adaptive signature design. *Clin Cancer Res* 16: 691-698, 2009.

[11] Cobo M, Isa D, et al: Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *J Clin Oncol* 25: 2747-2754, 2007.

[12] Cree IA, Kurbacher CM, Lamont A, et al: A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physicians choice in patients with recurrent platinum-resistant ovarian cancer. *Anticancer Drugs* 18: 1093-1101, 2007

[13] Rosell R, Vergnenegre A, Fournel P, et al: Pharmacogenetics in lung cancer for the lay doctor. *Target Oncol* 3: 161-171, 2008.

[14] Lee JJ, Gu X and Liu S: Bayesian adaptive randomization designs for targeted agent development. *Clin Tri* 7: 584-596, 2010.

[15] Freidlin B, McShane LM, Korn EL: Randomized clinical trials with biomarkers: design issues. *J Natl Cancer Inst* 102: 152-160, 2010.

[16] Liu C, Liu A, Hu J, Yuan V and Halabi S: Adjusting for misclassification in a stratified biomarker clinical trial. *Statist Med* 33: 3100-3113, 2014.

[17] Zang Y, Lee JJ and Yuan Y: Two-stage marker-stratified clinical trial design in the presence of biomarker misclassification. *Journal of the Royal Statistical Society-Series C* 65: 585-601, 2016.

[18] Schoenfeld D: The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 68: 316-319, 1981.