Published in final edited form as:
Stat Med. 2014 August 15; 33(18): 3100–3113. doi:10.1002/sim.6164.

Adjusting for Misclassification in a Stratified Biomarker Clinical Trial †

Chunling Liu[1], **Aiyi Liu**[2,*,†], **Jiang Hu**[3], **Vivian Yuan**[4], and **Susan Halabi**[5]

[1]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

[2]Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A

[3]Center for Biologics Evaluation and Research, Food and Drug Administration, U.S.A.

[4]Center for Drug Development and Research, Food and Drug Administration, U.S.A.

[5]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, 27708, U.S.A

Abstract

Clinical trials utilizing predictive biomarkers have become a research focus in personalized medicine. We investigate the effects of biomarker misclassification on the design and analysis of stratified biomarker clinical trials. For a variety of inference problems including marker-treatment interaction in particular, we show that marker misclassification may have profound adverse effects on the coverage of confidence intervals, power of the tests, and required sample sizes. For each inferential problem we propose methods to adjust for the classification errors.

Keywords

Biomarkers; classification error; correction for error; personalized medicine; power and sample size; prevalence; randomized controlled clinical trials; sensitivity and specificity

1. Introduction

Advances in understanding the genetics and biology of certain cancers have led to the successful development of novel therapies that target specific pathways. A convincing example is given in [1] that reported a statistically significant overall hazard ratio estimate from a randomized clinical trial in which women with ovarian cancer were treated with either pegylated liposomal doxorubicin or topotecan. The authors further reported that among patients with platinum-sensitive disease, a more significant hazard ratio was found. However, among patients with platinum-refractory disease, the hazard ratio was not

*Correspondence to: Aiyi Liu, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A.
†liua@mail.nih.gov

significant. The study results showed an evident interaction between treatment (pegylated liposomal doxorubicin or topotecan) and a biomarker (platinum).

When biomarker-treatment interaction is the primary research interest in a clinical trial, the stratified biomarker design is commonly used due to its fully taking the advantage of randomization and its ability to address various questions of interest; see, among others, [2–7]. In renal cell carcinoma, novel therapies that target the vascular endothelial growth factor (VEGF) and mammalian target of rapamycin (mTOR) pathways have been identified and are being used as treatment option for patients [8–9].

As another example, data from several centers have shown that retinoblastoma function may help differentiate if the androgen signaling pathway is viable. The loss of retinoblastoma status plays critical role in cell regulation and it suppresses androgen receptor expression and activity. It is estimated that 30% – 40% of prostate cancers will be androgen positive [10–12]. Investigators are interested in whether patients with advanced prostate cancer respond to treatment differently according to their retinoblastoma status.

Predictive markers for response have been shown to be important in patients with advanced renal cancer carcinoma. Furthermore, it has been reported that inhibition of the VEGF pathway prolong clinical outcomes, such as objective response, progression-free survival and overall survival. A statistically significant interleukin 6 (IL-6) by treatment interaction in predicting progression-free survival (PFS) was observed in patients with metastatic renal cell carcinoma (p-value=0.009) [13]. In patients with high IL-6, the median PFS was 33 weeks and 10 weeks in patients treated with pazopanib and placebo, respectively. On the other hand, the median PFS was 42 weeks and 24 weeks in low IL-6 patients treated with pazopanib and placebo, respectively [13].

We consider a two-arm trial (treatment versus standard) with T being the treatment indicator, where $T = 1$ if treatment and $T = 0$ if standard. We confine attention to a dichotomous predictive biomarker whose status is denoted by G (=1 if positive and =0 if negative). The prevalence of the biomarker is denoted by $\xi_G = \Pr(G = 1)$. Then in a stratified biomarker design, patients with the same biomarker status are randomized into treatment arm or standard arm, as shown in the following figure:

$$\text{True marker status} \Rightarrow \begin{cases} \text{positive}(G{=}1) \Rightarrow & \begin{cases} \text{treatment}(T{=}1), \\ \text{standard}(T{=}0); \end{cases} \\ \text{negative}(G{=}0) \Rightarrow & \begin{cases} \text{treatment}(T{=}1), \\ \text{standard}(T{=}0). \end{cases} \end{cases}$$

The primary interest of a stratified biomarker design is to investigate the marker-treatment interaction on a clinical endpoint, denoted by Y. Other questions that can be answered from the trial employing such a design include whether the treatments are different within the same marker status, or whether the clinical outcomes within the same treatment are different between marker status. These questions all involve inference on some function of the marker-by-treatment means of the clinical outcomes:

$$E(Y|G=g,T=t)=\mu_{gt},\quad VAR(Y|G=g,T=t)=\sigma_{gt}^2.$$

Define $\delta_g = \mu_{g1} - \mu_{g0}$ to be the mean outcome difference between treatments in the population with marker status $G = g$, and $\Delta_t = \mu_{1t} - \mu_{0t}$ to be the mean outcome difference between positive and negative marker status in the same treatment, as a measure of the marker effects in treatment arm $T = t$. We are interested in testing separately or simultaneously the null hypothesis $H_0 : \delta_g = 0$, ($g = 0, 1$), or $H_0^{'}:\Delta_t=0$, ($t = 0, 1$). The null hypothesis of no marker by treatment interaction is then $H_0^{''}:\gamma=0$, where

$$\gamma=\delta_1-\delta_0=\Delta_1-\Delta_0.$$

Because the independence between test statistics, the simultaneous null hypotheses can be tested by separately testing each individual null hypothesis with adequate allocation of the overall type I error rate, as demonstrated below for testing H_0.

Let W_g be a standardized test statistic for testing $H_{0g} : \delta_g = 0$. The null hypothesis $H_0 = H_{00} \cap H_{01}$ is rejected if $|W_g| > c_g$, for $g = 0$ or 1, where c_g are properly chosen critical values.

Assuming that W_0 and W_1 are independent, which is the case in most trial settings, the power of the test is given by

$$\omega(\delta_0,\delta_1)=\Pr(|W_0|>c_0 \text{ or } |W_1|>c_1)=\omega_0(\delta_0)+\omega_1(\delta_1)-\omega_0(\delta_0)\omega_1(\delta_1)$$

where $\omega_g(\delta_g) = \Pr(|W_g| > c_g)$ is the power of the test that rejects H_{0g} if $|W_g| > c_g$. The type I error rate is thus given by $\omega(0, 0) = \omega_0(0) + \omega_1(0) - \omega_0(0)\omega_1(0)$ where $\omega_g(0)$ is the type I error rate for testing H_{0g}.

With significance level α to test the null hypothesis H_0 and power $1 - \beta$ to detect marker-specific treatment differences δ_0 and δ_1, one can allocate adequately the type I error rate and the power to test separately the two null hypotheses, H_{00} and H_{01}. Suppose the allocation for H_{0g} is α_g for type I error and $1 - \beta_g$ for power at δ_g, then these allocated errors must satisfy $\alpha = \alpha_0 + \alpha_1 - \alpha_0\alpha_1$, and $\beta = \beta_0\beta_1$. In practice one can assign smaller error rates to the more important hypotheses, e.g. H_{01} that concerns the treatment difference in the marker-positive group. With equal allocation of type I error rates and power, we have $\alpha_0 = \alpha_1 = 1 - (1 - \alpha)^{1/2}$ and $\beta_0 = \beta_1 = \beta^{1/2}$. The null hypotheses $H_0^{'}:\Delta_t=0$: $\Delta_t = 0$, ($t = 0, 1$) can be dealt with similarly.

In the present article, we investigate, both analytically and numerically, the adverse effects of biomarker classification errors on the design of a stratified biomarker clinical trial. For a variety of inference problems including marker-treatment interaction, we show that marker misclassification may have profound adverse effects on the coverage of confidence intervals, power of the tests, and required sample sizes. For each inference problem we

propose methods to adjust for the classification errors. Sample size calculations adjusting for misclassification are presented in particular for testing marker-treatment interactions.

The paper is organized as follows. In Section 2, we present notations and preliminary results concerning the design of a stratified biomarker trial in the presence of marker misclassification. We then discuss the effects of misclassification on estimating treatment means in each marker stratum, and present a method to correct for misclassification in Section 3. We investigate the effects of misclassification on estimating treatment differences in each marker stratum in Section 4, followed by a method to correct for misclassification. We evaluate the effects of misclassification on marker differences in each treatment arm in Section 5, with a method to correct for marker misclassification. In Section 6, we address the marker-treatment interaction, starting with the investigation of the effects on power and sample size of misclassification, followed by a method to correct for misclassification and an approach to compute sample sizes to warrant adequate power to detect potential interaction. We then present an example and then discuss the findings in Section 7.

2. The Design in Presence of Misclassification

We assume that a gold standard exists to determine the true status G of the biomarker, with $G = 1$ being positive and 0 if otherwise. Due to reasons such as cost, ethics or administration, an imperfect assay is used, resulting in classification errors in determining the biomarker status. This is common in assaying a diagnostic biomarker; see, among others, [14–16]. Wang et al. [16] demonstrated that misclassification can inflate type I error rates in a noninferiority trial with binary outcomes.

Let M be the observed status of G, with sensitivity $\pi_1 = \Pr(M = 1 \mid G = 1)$ and specificity $\pi_0 = \Pr(M = 0 \mid G = 0)$. For the biomarker to be practically useful, we assume that $1/2 < \pi_0, \pi_1 \leq 1$. It thus follows that the probability that the observed status of the marker is positive for a patient is

$$\xi_M = \pi_1 \xi_G + (1-\pi_0)(1-\xi_G). \quad (1)$$

We refer to ξ_M as the observed prevalence which is bounded by $1 - \pi_0$ and π_1 because $0 \leq \xi_G \leq 1$, and $\pi_0 + \pi_1 > 0$.

The actual stratified design is carried out according to the figure with the observed marker status M replacing the true status G.

Suppose that a total of N patients are enrolled into the trial. Let Y_i be the observed clinical outcome of the ith ($i = 1, \ldots, N$) patient with observed marker status $M_i (= 0, 1)$, in treatment arm $T_i (= 0, 1)$.

Let N_1 be the number of patients with observed marker status being positive. Note that N_1 is a random variable following a binomial distribution with size N and success probability ξ_M; thus $E(N_1) = N\xi_M$. Write $N_0 = N - N_1$, the number of patients with observed marker status being negative. Let $N_{mt} = \lambda_{mt} N_m$ the number of patients in the subgroup with $M = m$ and $T = t$, where the allocation proportions $\lambda_{mt} \in [0, 1]$ are usually pre-specified, and $\lambda_{m1} + \lambda_{m0} = 1$.

The allocation ratio of treatment to standard in the $M = m$ group is then $\lambda_{m1}/\lambda_{m0}$. Equal allocation between treatments in the $M = m$ group corresponds to $\lambda_{mt} = 1/2$. The targeted biomarker-strategy designs correspond to an extreme allocation with $\lambda_{01} = 0$; see, e.g., [4] and [16].

To simplify the notations, we assume that all the tests have significance level α and the confidence intervals have confidence level $1 - \alpha$. We will refer as "naive" procedures to those with no adjustment for classification errors, and as "error-adjusted" procedures to those that adjust for misclassification errors. Wherever there is no ambiguity, we will omit these distinctions.

The naive estimators of μ_{gt} and σ^2_{gt} are given by

$$\hat{\mu}_{gt}=\frac{1}{N_{gt}}\sum_{\{i:M_i=g,T_i=t\}} Y_i,\ \hat{\sigma}^2_{gt}=\frac{1}{N_{gt}-1}\sum_{\{i:M_i=g,T_i=t\}}(Y_i-\hat{\mu}_{gt})^2.$$

The naive confidence limits of μ_{gt} are calculated as

$$\hat{\mu}_{gt} \pm \hat{\sigma}_{gt} Z_{\alpha/2}/\sqrt{N_{gt}}, \quad (2)$$

where throughout Z_r is the rth upper quantile of the standard normal distribution, that is, $\Phi(Z_r) = 1 - r$, where Φ denotes the standard normal distribution function.

The naive testing procedure rejects the null hypothesis H_{0g} if

$$|\hat{\delta}_g/s_g| > Z_{\alpha/2}, \quad (3)$$

where

$$\hat{\delta}_g=\hat{\mu}_{g1}-\hat{\mu}_{g0},\ \ s^2_g=\frac{1}{N_{g1}}\hat{\sigma}^2_{g1}+\frac{1}{N_{g0}}\hat{\sigma}^2_{g0}. \quad (4)$$

Similarly the null hypothesis $\tilde{H_{0t}}: \Delta_t = 0$ is rejected if

$$|\hat{\Delta}_t/\tilde{s}_t| > Z_{\alpha/2},$$

where

$$\hat{\Delta}_t=\hat{\mu}_{1t}-\hat{\mu}_{0t},\ \ \tilde{s}^2_t=\frac{1}{N_{1t}}\hat{\sigma}^2_{1t}+\frac{1}{N_{0t}}\hat{\sigma}^2_{0t}.$$

If there are no classification errors, then the aforementioned estimates are unbiased, and, if N is large enough, the tests have significance level α and the confidence intervals have coverage probability $1 - \alpha$. In the presence of misclassification, however, these claims need to be carefully examined and corrections need to be made to account for classification error whenever necessary.

Throughout, unless stated otherwise, distributions and their characteristics of estimators are unconditional, taking the randomness of the observed sample sizes N_m (m = 0, 1) into account. Such an unconditional approach will allow us to investigate the effects of the marker's prevalence ξ_G as well. Conditional inference given N_m can be obtained in the derivation by replacing N with N_1/ξ_M, where ξ_M is given in (18). To adjust for classification errors, we assume that the marker's prevalence ξ_G, sensitivity π_1, and specificity π_0 are known; this implies that the marker's positive and negative predictive values are also known, because of the well-know relationships:

$$\tau_1=\pi_1\xi_G/\xi_M,\ \tau_0=\pi_0(1-\xi_G)/(1-\xi_M). \quad (5)$$

3. Estimating Stratum-Specific Treatment Means μ_{gt}

3.1. Effects of Misclassification

If the true marker status of the ith patient is G_i, then by the conditional expectations arguments we have

$$\zeta_{mt}=E\,(Y_i|M_i=m,T_i=t)=\sum_{g=0}^{1}\mu_{gt}\Pr\,(G_i=g|M_i=m),$$

noting that the treatments play no role in determining the marker's status.

This leads to

$$\zeta_{1t}=\tau_1\mu_{1t}+(1-\tau_1)\mu_{0t},\ \ \zeta_{0t}=\tau_0\mu_{0t}+(1-\tau_0)\mu_{1t},$$

where $\tau_1 = \Pr\,(G_i = 1 \mid M_i = 1)$ and $\tau_0 = \Pr\,(G_i = 0 \mid M_i = 0)$ are the marker's positive predictive value and negative predictive value, respectively. Similarly we have

$$\nu^2_{mt}=VAR\,(Y_i|M_i=m,T_i=t)=\sum_{g=0}^{1}(\mu^2_{gt}+\sigma^2_{gt})\Pr\,(G_i=g|M_i=m)-\zeta^2_{mt},$$

and thus

$$\nu_{1t}^2=\tau_1(\mu_{1t}^2+\sigma_{1t}^2)+(1-\tau_1)(\mu_{0t}^2+\sigma_{0t}^2)-\zeta_{1t}^2, \quad (6)$$

$$\nu_{0t}^2=\tau_0(\mu_{0t}^2+\sigma_{0t}^2)+(1-\tau_0)(\mu_{1t}^2+\sigma_{1t}^2)-\zeta_{0t}^2. \quad (7)$$

Taking the marker classification errors into account, we have $E(\hat{\mu}_{gt}) = \zeta_{gt}$ and $E(\hat{\sigma}_{gt}^2)=\nu_{gt}^2$. The (unconditional) variances of the mean estimates are given by

$$VAR\,(\hat{\mu}_{1t})=E\,\{VAR\,(\hat{\mu}_{1t}|N_{1t})\}+VAR\,\{E\,(\hat{\mu}_{1t}|N_{1t})\}=E\left(\frac{\nu_{1t}^2}{N_{1t}}\right)\approx\frac{\nu_{1t}^2}{\lambda_{1t}\xi_M N} \quad (8)$$

and

$$VAR\,(\hat{\mu}_{0t})=E\left(\frac{\nu_{0t}^2}{N_{0t}}\right)\approx\frac{\nu_{0t}^2}{\lambda_{0t}(1-\xi_M)N}, \quad (9)$$

noting that N_1/N is a consistent estimator of ξ_M.

Therefore, in the presence of misclassification, the naive estimators, $\hat{\mu}_{gt}$ and $\hat{\sigma}_{gt}^2$, are no longer unbiased for the corresponding parameters (i.e., μ_{gt} and σ_{gt}) they estimate. The bias of the mean estimates is given by, respectively

$$E\,(\hat{\mu}_{1t})-\mu_{1t}=-(1-\tau_1)\Delta_t,\;\; E\,(\hat{\mu}_{0t})-\mu_{0t}=(1-\tau_0)\Delta_t. \quad (10)$$

If we assume that, in the same treatment group, larger clinical outcomes are more likely to occur in patients with positive marker status, then the treatment mean will be underestimated for marker positive patients, but overestimated for the marker negative patients.

For large sample, $\hat{\mu}_{1t}$ and $\hat{\mu}_{0t}$ are asymptotically normally distributed with ${N_{gt}}^{1/2}\,(\hat{\mu}_{gt}-\zeta_{gt})\sim\mathcal{N}(0,\nu_{gt}^2)$, where, throughout, "~" reads as "is distributed as". Then the coverage probability of the naive confidence interval of μ_{1t} in (2) is approximately

$$\Pr\left(\hat{\mu}_{1t}-\hat{\sigma}_{1t}Z_{\alpha/2}/N_{1t}^{1/2}\le\mu_{1t}\le\hat{\mu}_{1t}+\hat{\sigma}_{1t}Z_{\alpha/2}/N_{1t}^{1/2}\right)\approx\Phi\left(c_{1t}+Z_{\alpha/2}\right)-\Phi\left(c_{1t}-Z_{\alpha/2}\right) \quad (11)$$

where

$$c_{1t}=\frac{(1-\tau_1)\Delta_t(\lambda_{1t}N\xi_M)^{1/2}}{\nu_{1t}}.$$

The power, as a function of c_{1t}, strictly increases in $(-\infty, 0]$ and decreases in $[0, \infty)$. Therefore, when the true marker status can be correctly classified, (18) gives the coverage

probability approximately 100(1 − α)%. Otherwise, the asymptotic coverage probability of the naive confidence interval in (2) is always smaller than the nominal level of 1 − α. Indeed, the power can be substantially reduced; a particularly interesting observation is that the coverage probability approaches to zero when the sample size N gets larger.

3.2. Correction for Classification Error

From (10), unbiased estimators μ_{gt}^* of μ_{gt} can be derived by solving the equations:

$$\begin{cases} \hat{\mu}_{1t}=\tau_1\mu_{1t}^*+(1-\tau_1)\mu_{0t}^*, \\ \hat{\mu}_{0t}=(1-\tau_0)\mu_{1t}^*+\tau_0\mu_{0t}^*. \end{cases}$$

We have

$$\mu_{1t}^*=\frac{\tau_0\hat{\mu}_{1t}-(1-\tau_1)\hat{\mu}_{0t}}{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)},\ \ \mu_{0t}^*=\frac{\tau_1\hat{\mu}_{0t}-(1-\tau_0)\hat{\mu}_{1t}}{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)}.$$

It follows from (8) and (9) that the variances of the unbiased estimators are

$$VAR\,(\mu_{1t}^*) \approx \frac{\tau_0^2\nu_{1t}^2/(\lambda_{1t}\xi_M N)+(1-\tau_1)^2\nu_{0t}^2/\{\lambda_{0t}(1-\xi_M)N\}}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2}$$

and

$$VAR\,(\mu_{0t}^*) \approx \frac{\tau_1^2\nu_{0t}^2/\{\lambda_{0t}(1-\xi_M)N\}+(1-\tau_0)^2\nu_{1t}^2/(\lambda_{1t}\xi_M N)}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2}.$$

Recall that $E(\hat{\sigma}_{gt}^2)=\nu_{gt}^2$ where $\hat{\sigma}_{gt}^2$ are given in (4). Consistent estimate $\hat{VAR}\,(\mu_{1t}^*)$ of $VAR\,(\mu_{1t}^*)$ and $\hat{VAR}\,(\mu_{0t}^*)$ of $VAR\,(\mu_{0t}^*)$ are given by

$$\frac{\tau_0^2\hat{\sigma}_{1t}^2/N_{1t}+(1-\tau_1)^2\hat{\sigma}_{0t}^2/N_{0t}}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2},\ \frac{\tau_1^2\hat{\sigma}_{0t}^2/N_{0t}+(1-\tau_0)^2\hat{\sigma}_{1t}^2/N_{1t}}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2},$$

respectively.

Note that in large sample $(\mu_{gt}^*-\mu_{gt})/\{\hat{VAR}(\mu_{gt}^*)\}^{1/2}\sim\mathcal{N}(0,1)$. Therefore, if $\lambda_{gt}\to$ *constant* when $N\to\infty$, then the error-adjusted confidence interval of μ_{gt} with limits $\mu_{gt}^*\pm Z_{\alpha/2}\{\hat{VAR}(\mu_{gt}^*)\}^{1/2}$ has asymptotic coverage probability of 1 − α.

4. Inference on Marker-Specific Treatment Differences

4.1. Effects of Misclassification

We confine our attention to the marker positive group $G = 1$. The marker negative group can be dealt with similarly. Consider testing the null hypothesis H_{01} based on the statistics in (3). Taking misclassification into consideration, we have

$$E\left(\hat{\delta}_1\right)=\zeta_{11}-\zeta_{10}=\tau_1\delta_1+(1-\tau_1)\delta_0,\ VAR\left(\hat{\delta}_1\right)\approx\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}. \quad (12)$$

In large sample, $\hat{\delta_1}$ asymptotically follows a normal distribution. Note that, under the simultaneous null hypothesis $H_0 : \delta_1 = \delta_0 = 0$, $E(\hat{\delta_1}) = 0$. The actual type I error rate is then given by

$$\begin{aligned}\Pr\left(|\hat{\delta}_1|>s_1 Z_{\alpha/2}\right)&\approx\Pr\left\{\frac{|\hat{\delta}_1|}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}>\frac{s_1 Z_{\alpha/2}}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}\right\} \quad (13)\\ &\approx 2\left\{1-\Phi\left(Z_{\alpha/2}\right)\right\}=\alpha,\end{aligned}$$

utilizing the fact that s_1^2 defined in (4) is a consistent estimate of $\nu_{11}^2/(\lambda_{11}\xi_M N)+\nu_{10}^2/(\lambda_{10}\xi_M N)$.

Therefore, under simultaneous null hypothesis H_0, the naive tests maintain the type I error at the nominal level, regardless of the marker misclassification. However, unlike the cases when there is no classification error, the type I error rate of the test for the individual hypothesis $H_{01} : \delta_1 = 0$ depends on δ_0, and thus is no longer controlled at the nominal level. Indeed, the power of the test at $\delta_1 > 0$ is given by

$$\Phi\left\{\frac{\tau_1\delta_1+(1-\tau_1)\delta_0}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}-Z_{\alpha/2}\right\}+\Phi\left\{-\frac{\tau_1\delta_1+(1-\tau_1)\delta_0}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}-Z_{\alpha/2}\right\} \quad (14)$$

as compared to

$$\Phi\left\{\frac{\delta_1}{\left(\frac{\sigma_{11}^2}{\lambda_{11}\xi_G N}+\frac{\sigma_{10}^2}{\lambda_{10}\xi_G N}\right)^{1/2}}-Z_{\alpha/2}\right\},$$

when there is no classification error.

The type I error rate follows by setting $\delta_1 = 0$ and is given by

$$\Phi\left\{\frac{(1-\tau_1)\delta_0}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}-Z_{\alpha/2}\right\},$$

which can be substantially inflated, and indeed approaches to 1 when $N \to \infty$ and $\delta_0 > 0$.

Reduction in power due to misclassification may also be sizable. The loss of power attributes to the following observations. First, if we assume that marker-positive patients benefit more from the treatment than marker-negative patients, that is, $\delta_1 > \delta_0$, then $\delta_1 > \tau_1\delta_1 - (1 - \tau_1)\delta_0$. Secondly, assuming that $\sigma_{1t} = \sigma_{0t} = \sigma_t$, that is, the variations of the outcomes in the same treatment arm are not affected by the marker status. Then from (6) and (7) we have

$$\nu_{1t}^2=\sigma_t^2+\tau_1(1-\tau_1)\Delta_t^2>\sigma_t^2,\ \ \nu_{0t}^2=\sigma_t^2+\tau_0(1-\tau_0)\Delta_t^2>\sigma_t^2$$

if $\Delta_t \neq 0.3$) It is possible that $\tau_1\delta_1 + (1 - \tau_1)\delta_0 \approx 0$, which may occur when only patients with marker positive status are benefited from the treatment, that is $\delta_1 > 0 > \delta_0$.

The classification error can also substantially affects the coverage probability of the naive confidence interval $\hat{\delta_1} \pm s_1 Z_{\alpha/2}$. Similar to the derivation of (14), we obtain

$$\Pr\left(\hat{\delta}_1-s_1Z_{\alpha/2}\le\delta_1\le\hat{\delta}_1+s_1Z_{\alpha/2}\right)=\Phi\left(\Upsilon_1+Z_{\alpha/2}\right)-\Phi\left(\Upsilon_1-Z_{\alpha/2}\right),$$

where

$$\Upsilon_1=\frac{(1-\tau_1)\gamma}{\left(\frac{\nu_{11}^2}{\lambda_{11}\xi_M N}+\frac{\nu_{10}^2}{\lambda_{10}\xi_M N}\right)^{1/2}}.$$

Again, in the presence of classification error, the coverage probability is always smaller, and often substantially so, than the nominal level of $1 - \alpha$; it approaches to zero if $N \to \infty$.

4.2. Correction for Classification Error

Similar to (12), we can show that

$$E\left(\hat{\delta}_0\right)-(1-\tau_0)\delta_1+\tau_0\delta_0,\ VAR\left(\hat{\delta}_0\right)\approx\frac{\nu_{01}^2}{\lambda_{01}(1-\xi_M)N}+\frac{\nu_{00}^2}{\lambda_{00}(1-\xi_M)N}.$$

Therefore, unbiased estimates δ_g^* of δ_g can be obtained by solving the following equations:

$$\begin{cases} \hat{\delta}_1=\tau_1\delta_1^*+(1-\tau_1)\delta_0^*, \\ \hat{\delta}_0=(1-\tau_0)\delta_1^*+\tau_0\delta_0^*. \end{cases}$$

It follows that

$$\delta_1^*=\frac{\tau_0\hat{\delta}_1-(1-\tau_1)\hat{\delta}_0}{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)},\ \ \delta_0^*=\frac{\tau_1\hat{\delta}_0-(1-\tau_0)\hat{\delta}_1}{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)}.$$

The variance $VAR\,(\delta_1^*)$ of the unbiased estimator δ_1^* is approximately

$$\frac{\tau_0^2\,\{\nu_{11}^2/(\lambda_{11}\xi_M N)+\nu_{10}^2/(\lambda_{10}\xi_M N)\}+(1-\tau_1)^2\,[\nu_{01}^2/\{\lambda_{01}(1-\xi_M)N\}+\nu_{00}^2/\{\lambda_{00}(1-\xi_M)N\}]}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2},$$

which can be estimated consistently by

$$\hat{VAR}\,(\delta_1^*)=\frac{\tau_0^2\,(\hat{\sigma}_{11}^2/N_{11}+\hat{\sigma}_{10}^2/N_{10})+(1-\tau_1)^2\,(\hat{\sigma}_{01}^2/N_{01}+\hat{\sigma}_{00}^2/N_{00})}{\{\tau_1\tau_0-(1-\tau_1)(1-\tau_0)\}^2}.$$

Note that in large sample $(\delta_1^*-\delta_1)/\{\hat{VAR}(\delta_1^*)\}^{1/2}\sim\mathcal{N}(0,1)$. Therefore, the error-adjusted confidence interval of δ_1 with limits $\delta_1^*\pm Z_{\alpha/2}\{\hat{VAR}(\delta_1^*)\}^{1/2}$ has asymptotic coverage probability of $1-a$. Furthermore, the error-adjusted test that rejects $H_{01}:\delta_1=0$ if $|\delta_1^*|/\{\hat{VAR}(\delta_1^*)\}^{1/2}>Z_{\alpha/2}$ has type I error approximately a, regardless of the value of δ_0. The power is given by $\Phi[\delta_1/\{VAR(\delta_1^*)\}^{1/2}-Z_{\alpha/2}]$.

5. Inference on Treatment-Specific Marker Effects

5.1. Effects of Misclassification

Consider the naive test procedure given in Section 2. Taking the classification errors into account we have

$$E\left(\hat{\Delta}_t\right)=(\tau_0+\tau_1-1)\Delta_t,\ \ VAR(\hat{\Delta}_t)\approx\frac{\nu_{1t}^2}{\lambda_{1t}\xi_M N}+\frac{\nu_{0t}^2}{\lambda_{0t}(1-\xi_M)N}. \quad (15)$$

Therefore $\hat{\Delta_t}$ is no longer unbiased for Δ_t. Indeed, it always underestimates Δ_t if $\Delta_t > 0$ and overestimates Δ_t if $\Delta_t < 0$. In large sample, $\hat{\Delta_t}$ asymptotically follows a normal distribution. Similar to the derivations of (13) and (14) we conclude that the naive test asymptotically maintains the type I error at the nominal level, and the power of the test at some $\Delta_t > 0$ is given by

$$\Pr\left(\hat{\Delta}_t > \tilde{s}_t Z_{\alpha/2}\right) = \Phi\left[\frac{(\tau_0+\tau_1-1)\Delta_t}{\left\{\frac{\nu_{1t}^2}{\lambda_{1t}\xi_M N}+\frac{\nu_{0t}^2}{\lambda_{0t}(1-\xi_M)N}\right\}^{1/2}} - Z_{\alpha/2}\right]$$

which can be substantially smaller than

$$\Phi\left[\frac{\Delta_t}{\left\{\frac{\sigma_{1t}^2}{\lambda_{11}\xi_G N}+\frac{\sigma_{0t}^2}{\lambda_{0t}(1-\xi_G)N}\right\}^{1/2}} - Z_{\alpha/2}\right],$$

the power when there is no classification error.

Furthermore, the coverage probability of the naive confidence interval $\hat{\Delta}_t \pm s_t \tilde{Z}_{\alpha/2}$ of Δ_t is given by

$$\begin{aligned}\Pr\left(\hat{\Delta}_t - \tilde{s}_t Z_{\alpha/2} \le \Delta_t \le \hat{\Delta}_t + \tilde{s}_t Z_{\alpha/2}\right) &= \Phi\left(\tilde{\Upsilon}_t + Z_{\alpha/2}\right) - \Phi\left(\tilde{\Upsilon}_t - Z_{\alpha/2}\right) \\ &\le 1-\alpha, (\text{and} \to 0 \,\text{if}\, N \to \infty),\end{aligned}$$

where

$$\tilde{\Upsilon}_t = \frac{(2-\tau_0-\tau_1)\Delta_t}{\left\{\frac{\nu_{1t}^2}{\lambda_{1t}\xi_M N}+\frac{\nu_{0t}^2}{\lambda_{0t}(1-\xi_M)N}\right\}^{1/2}}.$$

5.2. Correction for Classification Error

Correction for misclassification follows from the fact that

$$\Delta_t^* = \frac{\hat{\Delta}_t}{\tau_0+\tau_1-1}$$

is an unbiased estimator of Δ_t. The variance and its consistent estimator are given respectively by

$$VAR\left(\Delta_t^*\right) \approx \frac{\nu_{1t}^2/(\lambda_{1t}\xi_M N)+\nu_{0t}^2/\{\lambda_{0t}(1-\xi_M)N\}}{(\tau_0+\tau_1-1)^2}, \quad \hat{VAR}\left(\Delta_t^*\right) = \frac{\hat{\sigma}_{1t}^2/N_{1t}+\hat{\sigma}_{0t}^2/N_{0t}}{(\tau_0+\tau_1-1)^2}.$$

In large sample $(\Delta_t^*-\Delta_t)/\left\{\hat{VAR}(\Delta_t^*)\right\}^{1/2} \sim \mathcal{N}(0,1)$. Assume that $\lambda_{gt} \to$ *constant* when $N \to \infty$. Then, the error-adjusted confidence interval of Δ_t with limits

$\Delta_t^* \pm Z_{\alpha/2}\{\hat{VAR}(\Delta_t^*)\}^{1/2}$ has asymptotic coverage probability of $1 - \alpha$. The error-adjusted test that rejects $\tilde{H_{0t}}: \Delta_t = 0$ if $|\Delta_t^*|/\{\hat{VAR}(\Delta_t^*)\}^{1/2} > Z_{\alpha/2}$ is equivalent to the naive test.

6. Inference on Marker-Treatment Interaction

6.1. Effects of Misclassification

Recall that the marker-treatment interaction effect is measured by $\gamma = \Delta_1 - \Delta_0$. It follows from (15) that the naive estimate of the interaction $\hat{\gamma} = \hat{\Delta_1} - \hat{\Delta_0}$ has mean and variance, given respectively by $E(\hat{\gamma}) = (\tau_0 + \tau_1 - 1)\gamma$ and $VAR(\hat{\gamma}) \approx \theta_1^2/N$ where

$$\theta_1^2 = \frac{\nu_{10}^2}{\lambda_{10}\xi_M} + \frac{\nu_{00}^2}{\lambda_{00}(1-\xi_M)} + \frac{\nu_{11}^2}{\lambda_{11}\xi_M} + \frac{\nu_{01}^2}{\lambda_{01}(1-\xi_M)}.$$

Therefore the naive estimator of the marker-treatment interaction is biased and under-(over-)estimates the interaction if $\gamma > (<) 0$. The naive test for interaction rejects the null hypothesis $H_0'': \gamma = 0$ if

$$|\hat{\gamma}|/(\tilde{s}_0^2 + \tilde{s}_1^2)^{1/2} > Z_{\alpha/2}.$$

The power of the test at some $\gamma > 0$ is given by

$$\Pr\left(\hat{\gamma} > Z_{\alpha/2}(\tilde{s}_0^2 + \tilde{s}_1^2)^{1/2}\right) = \Phi\left\{\frac{(\tau_0 + \tau_1 - 1)\gamma N^{1/2}}{\theta_1} - Z_{\alpha/2}\right\}. \quad (16)$$

It follows from (16) that the naive test maintains the type I error rate at the nominal level of α, regardless of the classification errors. However, the power of the test can be substantially adversely affected as compared to the power of the test with no misclassification, that is, $\Phi(\gamma N^{1/2}/\theta_0 - Z_{\alpha/2})$, where θ_0 is such that

$$\theta_0^2 = \frac{\sigma_{10}^2}{\lambda_{10}\xi_G} + \frac{\sigma_{00}^2}{\lambda_{00}(1-\xi_G)} + \frac{\sigma_{11}^2}{\lambda_{11}\xi_G} + \frac{\sigma_{01}^2}{\lambda_{01}(1-\xi_G)}.$$

The coverage probability of the naive confidence interval $\hat{\gamma} \pm z_{\alpha/2}(\tilde{s}_0^2 + \tilde{s}_1^2)^{1/2}$ of γ is given by

$$\begin{aligned}\Pr&\left\{\hat{\gamma}-z_{\alpha/2}(\tilde{s}_0^2+\tilde{s}_1^2)^{1/2}\le\gamma\le\hat{\gamma}+z_{\alpha/2}(\tilde{s}_0^2+\tilde{s}_1^2)^{1/2}\right\}\\&=\Phi\left\{\frac{(2-\tau_0-\tau_1)\gamma N^{1/2}}{\theta_1}+Z_\alpha/2\right\}\\&-\Phi\left\{\frac{(2-\tau_0-\tau_1)\gamma N^{1/2}}{\theta_1}-Z_\alpha/2\right\},\end{aligned}$$

which can be substantially lower (approaching 0 if $N\to\infty$) than the nominal level of $1-\alpha$.

For $\alpha = 0.05$, $\sigma_{gt} = 1$, $\lambda_{mt} = 1$, $\gamma = 0.936$, and selected values of π_0, π_1, ξ_G, and N, Table 1 presents coverage probability of the naive confidence interval and the power of the naive test. In all cases the actual coverage probability is smaller that the nominal level of 0.95, many are of more than 25% reduction. The actual power is also substantially lower than that with no classification errors, some with more than 50% reduction in power. The coverage probability and the power increase as the classification accuracy improves. An increased sample size yields increased power but decreased coverage probability. For example, with 90% sensitivity and specificity respectively, and 40% marker prevalence, the naive coverage probability is 0.90 and the power is 0.71 if the sample size is $N = 200$. These two measures change to 0.84 and 0.95 respectively when the sample size doubles.

6.2. Correction for Classification Error

An unbiased estimator of the interaction effect γ can be given by

$$\gamma^*=\frac{\hat{\gamma}}{\tau_0+\tau_1-1}.$$

The variance and its consistent estimator are given respectively by

$$VAR(\gamma^*)\approx\frac{\theta_1^2}{N(\tau_0+\tau_1-1)^2},\ \hat{VAR}(\gamma^*)=\frac{\hat{\theta}_1^2}{N(\tau_0+\tau_1-1)^2}$$

where

$$\hat{\theta}_1^2=\frac{\hat{\sigma}_{10}^2}{\lambda_{10}\xi_M}+\frac{\hat{\sigma}_{00}^2}{\lambda_{00}(1-\xi_M)}+\frac{\hat{\sigma}_{11}^2}{\lambda_{11}\xi_M}+\frac{\hat{\sigma}_{01}^2}{\lambda_{01}(1-\xi_M)}.$$

In large sample $(\gamma^*-\gamma)/\left\{\hat{VAR}(\gamma^*)\right\}^{1/2}\sim\mathcal{N}(0,1)$. Hence, the error-adjusted confidence interval of γ with limits $\gamma^*\pm Z_{\alpha/2}\{\hat{VAR}(\gamma^*)\}^{1/2}$ has asymptotic coverage probability of $1-$

a. The error-adjusted test that rejects $H_0'':\gamma=0$ if $|\gamma^*|/\{\hat{VAR}(\gamma^*)\}^{1/2} > Z_{\alpha/2}$ is equivalent to the naive test.

6.3. Sample Size Adjustment

For the stratified biomarker design, the sample size N needs to be sufficiently large to ensure adequate power of $1-\beta$ to detect a meaningful marker-treatment interaction γ. From (16) the sample size is given by

$$N=\frac{(Z_{\alpha/2}+Z_\beta)^2\theta_1^2}{(\tau_0+\tau_1-1)^2\gamma^2}. \quad (17)$$

On the other hand, in the absence of misclassification the required sample size is

$$N'=\frac{(Z_{\alpha/2}+Z_\beta)^2\theta_0^2}{\gamma^2}.$$

It follows from (18) and (5) that

$$1 \geq \tau_0+\tau_1-1=\frac{(1-\xi_G)(2\pi_0-1)}{\xi_M} \geq \frac{(1-\xi_G)(2\pi_0-1)}{\pi_1}>0.$$

Furthermore, as pointed out in Section 4.1, the variance ν is usually larger than its counterpart σ. Therefore, a much larger sample size may be required to achieve the desirable power when classification errors exist.

Under the same specifications of parameters' values (except for N) used for Table 1, Table 2 presents the actual sample size needed and its ratio to the sample size when there is no classification error. It shows that the sample size can be more than twice that required when there is no misclassification of the marker status.

6.4. Example

We sought to design a phase III trial where patients with metastatic renal cell carcinoma will be randomized to sunitinib (standard of care) or sunitinib plus an experimental drug stratified by the IL-6 status. The primary endpoint is progression-free survival (PFS) rate at 6 months. IL-6 is a continuous variable with high IL-6 status defined as a value greater than or equal to 13 pg/mL; this cut-point value is based on the observed median as was reported in one study [13]. Based on observed data, the PFS rate at 6-months in low and high IL-6 patients treated with sunitinib is 48% and 18%, respectively. The hypothesized effect in low and high IL-6 patients treated with the experimental drug is 66% and 59%, respectively. The assay has 95% sensitivity and 90% specificity. Assuming equal allocation and 40% prevalence of high IL-6. Assuming further that a power of 0.85 is desirable to detect a marker-treatment interaction effect of $\gamma = (0.59 - 0.18) - (0.66 - 0.48) = 0.23$ in PFS rates.

Using equation (17), the required sample size is about 1,020, or 255 patients are needed in each stratum of IL-6 by treatment. If on the other hand, the prevalence of high IL-6 status is 30%, then the required sample size is much larger, about 1,244, or 311 patients in each stratum. In contrast, the sample sizes are about 177 and 202 respectively per stratum when there is no classification errors for the two scenarios. Note that similar to the comparison of two independent proportions, in the calculation the stratum-specific variances σ_{gt}^2 are set to be $\bar{\mu}(1-\bar{\mu})$ where $\bar{\mu}$ is the average of stratum-specific rates, that is,

$$\bar{\mu}=(\mu_{11}+\mu_{01}+\mu_{10}+\mu_{00})/4 \ (=(0.59+0.66+0.18+0.48)/4=0.4775),$$

yielding $\sigma_{gt} = 0.25$.

7. Discussion

In the present paper we demonstrated both analytically and numerically that the misclassified biomarker status can have profound negative impact on various inference problems in a stratified biomarker trial. The methods developed are based on asymptotic theory and are suitable for most biomarker stratified trials that usually require relatively large sample sizes; however, caution needs to be taken for small-size trials.

It is worth noting that, as a result of the randomization, the naive test for marker-treatment interaction maintains the required type I error rates, but suffers considerably from loss of power due to misclassification, which in turn, results in larger sample sizes required for the trial.

Our investigation assumes that the marker's prevalence ξ_G, sensitivity π_1, and specificity π_0 are all known. When the N patients are a representative sample of the targeted population, $\hat{\xi}_M=\sum_{i=1}^{N} M_i/N$ is an unbiased estimate of ξ_M. Then from (18), it follows that

$$\hat{\xi}_G=\frac{\hat{\xi}_M-(1-\pi_0)}{\pi_0+\pi_1-1}$$

is an unbiased estimate of the marker's prevalence ξ_G. If sensitivity π_1, and specificity π_0 are unknown, then a preliminary study can be conducted to estimate π_0 and π_1.

The technical developments employed in the present paper can be readily extended to other biomarker-driven designs, for example, the biomarker enrichment strategy design in which only marker positive patients are randomized to receive treatments. However, as shown in the developments, data from all marker by treatment strata are needed to adjust for classification errors. For a review of useful biomarker based clinical designs, see, e.g. [3, 6, 18,19]. Although the choice of these various designs depends on the trial aims, the impact of biomarker misclassification can be substantial in each design, and needs further evaluations. For example, some designs involve testing multiple hypotheses concerning the various

aspects of the marker-treatment effects. It is then important to investigate how the classification errors adversely affect the allocation of type I error rates and the power of the study. Such investigation is also warranted for adaptive and Bayesian biomarker designs.

Throughout, testing marker-treatment effects is formulated based on stratum-specific means, e.g. means of normal distributions or proportions of a dichotomous endpoint. The methods developed in the present paper could be generalized, with some tedious algebraic manipulations, to ordinal/categorical and longitudinal/repeated endpoints with stratum means as the primary interest. We are currently working on extending the method for time-to-event endpoints and longitudinally measured endpoints with hazards ratio and rates of change as the primary comparison, respectively. As can be expected, these types of endpoint require different and more complicated technical handling of the assumptions.

Increased advances in understanding the roles of molecular and genetic pathways in carcinogenesis are leading to the development of novel therapies that target the disease pathways. As a result of these advances, the landscape for performing clinical trials with biomarkers in cancer is evolving and becoming complex. Despite the large sample size required for the stratified biomarker design, we believe that this approach is realistic and worth it as it accounts for misclassification errors.

Acknowledgments

Research of A. Liu was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health. Research of S. Halabi was supported by grants R01-CA155296 and U01-CA157703.

References

1. Gordon AN, Tonda M, Sun S, Rackoff W. Doxil study 30–49 investigators. Long-term survival advantage for women treated with pegylated liposomal doxorubicin compared with topotecan in a phase 3 randomized study of recurrent and refractory epithelial ovarian cancer. Gynecologic Oncology. 2004; 95:1–8. [PubMed: 15385103]
2. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research. 2004; 10:6759–6763. [PubMed: 15501951]
3. Mandrekar SJ, Sargent DJ. Clinical Trial Designs for Predictive Biomarker Validation: Theoretical Considerations and Practical Challenges. Journal of Clinical Oncology. 2009; 27:4027–4034. [PubMed: 19597023]
4. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: Design issues. Journal National Cancer Institute. 2010; 102:152–160.
5. Joo J, Geller NL, French B, Kimmel SE, Rosenberg Y, Ellenberg JH. Prospective alpha allocation in the clarification of optimal anticoagulation through genetics (COAG) trial. Clinical Trials. 2010; 7:597–604. [PubMed: 20693186]
6. Simon R. Clinical trials for predictive medicine: new challenges and paradigms. Clinical Trials. 2010; 7:516–524. [PubMed: 20338899]
7. Lai TL, Lavori PW. Innovative clinical trial designs toward a 21st-century health care system. Statistics in Bioscience. 2011; 3:145–168.
8. Motzer RJ, Hutson TE, Tomczak P, Michaelson MD, Bukowski RM, Rixe O, Oudard S, Negrier S, Szczylik C, Kim ST, Chen I, Bycott PW, Baum CM, Figlin RA. Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. New England Journal of Medicine. 2007; 356:115–124. [PubMed: 17215529]

9. Rini BI, Halabi S, Rosenberg JE, Stadler WM, Vaena DA, Ou SS, Archer L, Atkins JN, Picus J, Czaykowski P, Dutcher J, Small EJ. Bevacizumab plus interferon-alpha versus interferon-alpha monotherapy in patients with metastatic renal cell carcinoma: Results of CALGB 90206. Journal of Clinical Oncology. 2008; 26:5422–5428. [PubMed: 18936475]
10. Bosco EE, Wang Y, Xu H, Zilfou JT, Knudsen KE, Aronow BJ, Lowe SW, Knudsen ES. The retinoblastoma tumor suppressor modifies the therapeutic response of breast cancer. Journal of Clinical Investigation. 2007; 117:218–228. [PubMed: 17160137]
11. Sharma A, Comstock CE, Knudsen ES, Cao KH, Hess-Wilson JK, Morey LM, Barrera J, Knudsen KE. Retinoblastoma tumor suppressor status is a critical determinant of therapeutic response in prostate cancer cells. Cancer Research. 2007; 67:6192–6203. [PubMed: 17616676]
12. Sharma A, Yeow WS, Ertel A, Coleman I, Clegg N, Thangavel C, Morrissey C, Zhang X, Comstock CE, Witkiewicz AK, Gomella L, Knudsen ES, Nelson PS, Knudsen KE. The retinoblastoma tumor suppressor controls androgen signaling and human prostate cancer progression. Journal of Clinical Investigation. 2010; 120:4478–4492. [PubMed: 21099110]
13. Tran HT, Liu Y, Zurita AJ, Lin Y, Baker-Neblett KL, Martin AM, Figlin RA, Hutson TE, Sternberg CN, Amado RG, Pandite LN, Heymach JV. Prognostic or predictive plasma cytokines and angiogenic factors for patients treated with pazopanib for metastatic renal-cell cancer: a retrospective analysis of phase 2 and phase 3 trials. Lancet Oncology. 2012; 13:827–837. [PubMed: 22759480]
14. Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. European Journal of Human Genetics. 2001; 9:130–134. [PubMed: 11313746]
15. Hao K, Li C, Rosenow C, Wong WH. Estimation of genotype error rate using samples with pedigree informationan application on the GeneChip Mapping 10K array. Genomics. 1992; 84:623–630. [PubMed: 15475239]
16. Wang SJ, Hung HMJ, O'Neill RT. Genomic classifier for patient enrichment: Misclassification and type I error issues in pharmacogenomics noninferiority trial. Statistics in Biopharmaceutical Research. 2011; 3:310–319.
17. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine. 2005; 24:329–339. [PubMed: 15551403]
18. Freidlin B, Korn EL. Biomarker-adaptive clinical trial designs. Pharmacogenomics. 2010; 11:1679–1682. [PubMed: 21142910]
19. Gosho M, Nagashima K, Sato Y. Study Designs and Statistical Analyses for Biomarker Research. Sensors. 2012; 12:8966–8986. [PubMed: 23012528]

Appendix: Some Technical Details

Proof of Eq. (1)

$$\begin{aligned}\xi_M &= \Pr(M=1) = \Pr(M=1, G=1) + \Pr(M=1, G=0) \\ &= \Pr(M=1|G=1)\Pr(G=1) + \Pr(M=1|G=0)\Pr(G=0) \\ &= \pi_1 \xi_G + (1-\pi_0)(1-\xi_G).\end{aligned}$$

Proof of Eq. (11)

$$\begin{aligned}\Pr\left(\hat{\mu}_{1t} - \hat{\sigma}_{1t} Z_{\alpha/2}/N_{1t}^{1/2} \le \mu_{1t} \le \hat{\mu}_{1t} + \hat{\sigma}_{1t} Z_{\alpha/2}/N_{1t}^{1/2}\right) &\approx \Pr\left(\hat{\mu}_{1t} - \nu_{1t} Z_{\alpha/2}/N_{1t}^{1/2} \le \mu_{1t}\right) - \Pr\left(\hat{\mu}_{1t} + \nu_{1t} Z_{\alpha/2}/N_{1t}^{1/2} \le \mu_{1t}\right) \\ &= E\left\{\Phi\left(\frac{N_{1t}^{1/2}(1-\tau_1)\Delta_t}{\nu_{1t}} + Z_{\alpha/2}\right)\right\} - E\left\{\Phi\left(\frac{N_{1t}^{1/2}(1-\tau_1)\Delta_t}{\nu_{1t}} - Z_{\alpha/2}\right)\right\} \\ &\approx \Phi(c_{1t} + Z_{\alpha/2}) - \Phi(c_{1t} - Z_{\alpha/2})\end{aligned}$$

Note that in the third expression the expectation is taken with respect to the random number N_{1t}.

Table 1

Coverage probability of the naive confidence interval and power of the naive test for marker-treatment interaction:

	(N = 200, ξ_G = 0.4)†			
	π_1=0·80	0·85	0·90	0·95
π_0= 0·80	0·74/0·46	0·78/0·52	0·82/0·58	0·85/0·64
0·85	0·80/0·53	0·83/0·59	0·86/0·65	0·89/0·70
0·90	0·85/0·61	0·88/0·66	0·90/0·71	0·91/0·76
0·95	0·90/0·69	0·91/0·73	0·93/0·78	0·94/0·82
	(N = 200, ξ_G = 0.6)†			
	π_1=0 ·80	0·85	0·90	0·95
π_0= 0·80	0·74/0·46	0·80/0·53	0·85/0·61	0·90/0·69
0·85	0·78/0·52	0·83/0·59	0·88/0·66	0·91/0·73
0·90	0·82/0·58	0·86/0·65	0·90/0·71	0·93/0·78
0·95	0·85/0·64	0·89/0·70	0·91/0·76	0·94/0·82
	(N = 400, ξ_G = 0.4)††			
	π_1=0·80	0·85	0·90	0·95
π_0= 0·80	0·54/0·75	0·61/0·81	0·68/0·86	0·75/0·91
0·85	0·65/0·82	0·71/0·87	0·76/0·91	0·82/0·94
0·90	0·75/0·88	0·80/0·92	0·84/0·95	0·88/0·97
0·95	0·85/0·93	0·88/0·96	0·90/0·97	0·92/0·98
	(N = 400, ξ_G = 0.6)††			
	π_1=0·85	0·90	0·95	0·99
π_0= 0·80	0·54/0·75	0·65/0·82	0·75/0·88	0·85/0·93
0·85	0·61/0·81	0·71/0·87	0·80/0·92	0·88/0·96
0·90	0·68/0·86	0·76/0·91	0·84/0·95	0·90/0·97
0·95	0·75/0·91	0·82/0·94	0·88/0·97	0·92/0·98

† power=0.90 if no misclassification;

†† power=0.99 if no misclassification.

Table 2

Required sample size and its ratio to the sample size (= 200) when there is no misclassification

		ξ_G = 0.4		
	π_1=0·80	0·85	0·90	0·95
π_0= 0·80	612/3·06	522/2·61	449/2·24	388/1·94
0·85	508/2·54	442/2·21	386/1·93	339/1·69
0·90	423/2·11	373/1·87	331/1·65	295/1·47
0·95	350/1·75	314/1·57	283/1·41	255/1·27
		ξ_G = 0.6		
	π_1=0·80	0·85	0·90	0·95
π_0= 0·80	612/3·06	508/2·54	423/2·11	350/1·75
0·85	522/2·61	442/2·21	373/1·87	314/1·87
0·90	449/2·24	386/1·93	331/1·65	283/1·41
0·95	388/1·94	339/1·69	295/1·47	255/1·27