# Two-stage stratified designs with survival outcomes and adjustment for misclassification in predictive biomarkers

**Yanping Chen**[1], **Yong Lin**[2,3], **Shou-En Lu**[2,3], **Weichung J. Shih**[2,3], **Hui Quan**[4]

[1]Global Biometrics and Data Sciences, Bristol Myers Squibb, Berkeley Heights, New Jersey, USA

[2]Biostatistics and Epidemiology Department, School of Public Health, Rutgers University, Piscataway, New Jersey, USA

[3]Biometrics Division, Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA

[4]Biostatistics and Programming, Sanofi, Bridgewater, New Jersey, USA

## Abstract

Biomarker stratified clinical trial designs are versatile tools to assess biomarker clinical utility and address its relationship with clinical endpoints. Due to imperfect assays and/or classification rules, biomarker status is prone to errors. To account for biomarker misclassification, we consider a two-stage stratified design for survival outcomes with an adjustment for misclassification in predictive biomarkers. Compared to continuous and/or binary outcomes, the test statistics for survival outcomes with an adjustment for biomarker misclassification is much more complicated and needs to take special care. We propose to use the information from the observed biomarker status strata to construct adjusted log-rank statistics for true biomarker status strata. These adjusted log-rank statistics are then used to develop sequential tests for the global (composite) hypothesis and component-wise hypothesis. We discuss the power analysis with the control of the type-I error rate by using the correlations between the adjusted log-rank statistics within and between the design stages. Our method is illustrated with examples of the recent successful development of immunotherapy in nonsmall-cell lung cancer.

### Keywords

adjusted log-rank statistics; composite Hypothesis; predictive biomarker; sensitivity and specificity; survival outcome

**Correspondence**: Yong Lin, Biostatistics and Epidemiology Department, School of Public Health, Rutgers University, 683 Hoes Lane West, RM 214, Piscataway, NJ 08854, USA., linyo@rutgers.edu.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

## 1 | INTRODUCTION

Advancements in cancer biology and genomics-based technologies have revealed that patient responses to therapeutics are often heterogeneous as the genetic mechanism of cancer can be heterogeneous across patients. There has been a rapidly growing interest in biomarker-driven personalized cancer therapy, also known as targeted therapy.[1,2] Clinical trials utilizing biomarkers in selecting patients for targeted therapies have been developed in the past few decades.[3–5] Recently, Renfro et al[6] provided an excellent review and useful examples and references. Among biomarker-based clinical trials, the biomarker-stratified design is probably one of the most popular designs and has been used in many clinical trials.[7,8] Typically, patients are stratified into either a marker positive or a marker negative subgroup by their biomarker profiles, and then randomized to receive either the targeted therapy or the standard therapy within each marker stratum. These stratified designs can detect a statistically significant biomarker-by-treatment interaction effect, thereby statistically confirm the predictive ability of the biomarker. These designs often require a relatively large sample size, as its structure resembles multiple randomized trials conducted in parallel.[9] However, one practical impediment to the use of biomarker stratified designs is that, in some cases, the ascertainment of biomarker status may be subject to errors/misclassification, which makes it challenging to evaluate the treatment effect and identify/confirm predictive biomarker. Several papers have considered misclassification problems with biomarkers but in different contexts. For example, Krisam and Keiser[10] investigated decision rules for selecting target populations and methods for calculating sample size to achieve a specified selection probability. Friedlin et al[11] reported a Marker Sequential Test design and compared with the sequential subgroup-specific design that sequentially assesses the treatment effect in the biomarker subgroups. Wang et al,[12,13] focused on misclassification issues in noninferiority trials. Wang and Lim[14] applied a machine learning algorithm to validate biomarker predictive performance in assessing subpopulation-specific treatment effects. Liu et al[15] showed that marker misclassification may have profound adverse effects on the coverage of confidence intervals, power of the tests, and required sample sizes. Zang et al[16] proposed a two-stage design where the biomarkers are measured by both error-free and error-prone methods in patients enrolled in Stage I, and the biomarker is only measured on the error-prone method in patients enrolled in Stage II. They investigated the bias in the estimation of prognostic and predictive biomarker effects and treatment effects caused by biomarker misclassification in the presence of covariates. Zang and Guo[17] proposed a two-stage optimal enrichment design that utilizes the surrogate marker to correct for the biomarker misclassification. Lin et al[18] reported a two-stage adaptive enrichment design for continuous endpoints, with an adjustment for the misclassification of predictive biomarkers. Those reports focused on the misclassification adjustment for continuous and/or binary endpoints. To the best of our knowledge, adjustment for misclassification of biomarkers for assessing survival endpoint has not yet been reported.

In this paper, we propose a two-stage stratified design with survival outcomes. The stratification is based on the observed biomarker status, and specifically, the misclassification of a binary predictive biomarker is considered. Similar to Liu et al,[15] we assume that the prevalence rate of the biomarker and the sensitivity and specificity of the biomarker determination test are known. Compared to continuous and/or binary outcomes, the test statistics for survival outcomes with adjustment for biomarker misclassification is much more complicated and needs special care. We propose to use the information from the observed biomarker status strata to construct adjusted log-rank statistics for true biomarker status strata. These adjusted log-rank statistics are then used to develop sequential tests for the global (composite) hypothesis and component-wise hypothesis. We discuss the power analysis with the control of the type-I error rate by using the correlations between the adjusted log-rank statistics within and between the design stages.

This paper is organized as follows: In Section 2, we give the hypotheses of interest. In Section 3, we present the proposed two-stage stratified design. First, we describe our method to derive the log-rank statistics, adjusted for biomarker misclassification. The correlation matrix of the adjusted log-rank statistics within and between the design stages and their asymptotic joint distribution are presented. These correlations are crucial in calculating the critical values, type I error rates, and power in testing treatment effects. In Section 4, a numerical example is used to illustrate our method. Finally, discussion is followed in Section 5.

## 2 | PARAMETERS AND HYPOTHESIS

Consider a two-arm randomized controlled clinical trial to assess whether the effect of a test treatment is superior to control (standard of care) in terms of survival endpoint (eg, time to disease progression, or death). Suppose some baseline genomic biomarker may be predictive of a treatment effect. For simplicity, assume that the biomarker is dichotomized into either a positive or negative status. Denote the true marker status by $D = +$ for true positive and − for true negative status. Denote observed marker status by $M = *$ for observed positive and $\phi$ for observed negative status. The stratified randomization is performed based on observed marker status with sensitivity $\lambda_{\text{sen}}$ and specificity $\lambda_{\text{spec}}$. In this paper, we use observed biomarker positive (or negative) strata, observed positive (or negative) group interchangeably.

Let the prevalence of the true marker positive be $P(D = +) = p$. Then the proportion of observed marker positive is $q = P(M = *) = p\lambda_{\text{sen}} + (1 - p)(1 - \lambda_{\text{spec}})$, the positive predictive value (PPV) $= P(D = + \mid M = *)$ is $\tau = \frac{p\lambda_{\text{sen}}}{p\lambda_{\text{sen}} + (1 - p)(1 - \lambda_{\text{spec}})} = \frac{p\lambda_{\text{sen}}}{q}$, and the negative predictive value (NPV) $= P(D = - \mid M = \phi)$ is $\eta = \frac{(1 - p)\lambda_{\text{spec}}}{1 - P(M = *)} = \frac{(1 - p)\lambda_{\text{spec}}}{1 - q}$. Note that when $\lambda_{\text{sen}} = \lambda_{\text{spec}} = 1,\ q = p$ and $\tau = \eta = 1$.

Let randomization ratio be $r$ to active treatment and $1 - r$ to placebo treatment, where $0 < r < 1$, and $h_{ij}(t)$ denote the hazard at time $t$ with $i = +$ (positive) or − (negative) for true marker status, and $j = T$ for active treatment or $C$ for control treatment. Assuming proportional hazards between treatment and control, we denote the treatment effect for true marker positive stratum by $\delta_+ = \log\frac{h_{+T}(t)}{h_{+C}(t)}$, with the corresponding null hypothesis $H_{0+}: h_{+T}(t) = h_{+C}(t)$. Denote the treatment effect for true marker negative stratum by

$\delta_- = \log\frac{h_{-T}(t)}{h_{-C}(t)}$, corresponding to null hypothesis $H_{0-}: h_{-T}(t) = h_{-C}(t)$. The overall treatment effect is $\delta = w_1\delta_+ + w_0\delta_-$, where $w_1(\geq 0)$ and $w_0(\geq 0)$ are some weights for treatment effect.[9] The corresponding null hypothesis is $H_{0a}: \delta = 0$. A natural choice of weight is the marker positive prevalence, that is, $w_1 = p$ and $w_0 = 1 - p$. The quantity $\delta = p\delta_+ + (1-p)\delta_-$ is interpreted as “treatment utility effect”.[9] In our design, the weights are prespecified as $w_1 = p$ and $w_0 = 1 - p$, and this $p$ is a known value.

We are interested in testing the treatment effect in the overall patient population

$$H_{0a}: \delta = 0 \text{ vs } H_{1a}: \delta < 0,$$

as well as the treatment effect in the true marker-positive cohort

$$H_{0+}: \delta_+ = 0 \text{ vs } H_{1+}: \delta_+ < 0.$$

Since we are pursuing either treatment effect in the overall population or only the true marker-positive cohort (ie, $\delta < 0$ or $\delta_+ < 0$), we are testing the following composite (global) hypotheses:

$$H_0: \delta = \delta_+ = 0 (H_{0a} \cap H_{0+}) \text{vs}$$

$$H_1: \delta < 0 \text{ or } \delta_+ < 0 (H_{1a} \cup H_{1+}).$$

### 2.1 | Asymptotic distribution of log-rank statistic

When the biomarker assay is perfect, that is, $\lambda_{\text{sen}} = \lambda_{\text{spec}} = 1$, the treatment effect can be tested using the conventional log-rank test. Specifically, let $Y_{ijk}$ denote the number of subjects at risk at the $k$th distinct event time in the $i$th true marker stratum and the $j$th treatment group, where $i = +$ for true marker positive and $-$ for true marker negative, and $j = T$ for active treatment or $C$ for control treatment. The unstandardized log-rank statistic for testing the treatment effect in the $i$th marker status stratum can be expressed as

$$Q_i = \sum_{k=1}^{K}\left(d_{iTk} - \frac{d_{i\cdot k}Y_{iTk}}{Y_{i\cdot k}}\right),$$

where $i = +$ or,$-$, $d_{i\cdot k}$ and $Y_{i\cdot k}$ denote the total number of events and total number of subjects at the $k$th event time, respectively, and $K$ is the total number of different event times in the $i$th stratum, then

$$\widehat{\mathrm{Var}}(Q_i) = \sum_{k=1}^{K}\widehat{\mathrm{Var}}(d_{iTk}) = \sum_{k=1}^{K}\frac{Y_{iTk}Y_{iCk}d_{i\cdot k}(Y_{i\cdot k} - d_{i\cdot k})}{Y_{i\cdot k}^2(Y_{i\cdot k} - 1)},$$

is a consistent estimator of the asymptotic variance $\mathrm{Var}(Q_i)$ of $Q_i$. It is well known that under the null hypothesis: $h_{iT}(t) = h_{iC}(t)$, the standardized log-rank statistic

$$Z_i = \frac{Q_i}{\sqrt{\widehat{\mathrm{Var}}(Q_i)}} \sim \mathrm{AN}(0,1)\,.$$

In Appendix B, based on Schoenfeld's method,[19] under alternative hypothesis $h_{iT}(t) \neq h_{iC}(t)$ satisfying $\log(h_{iT}(t)/h_{iC}(t)) = O\left(n^{-1/2}\right)$, we have

$$Z_i = \frac{Q_i}{\sqrt{\widehat{Var}(Q_i)}} \sim \mathrm{AN}(\psi_i, 1),$$

with

$$\psi_i = \sqrt{r(1-r)D_i}\log\frac{h_{iT}}{h_{iC}},$$

and, asymptotically,

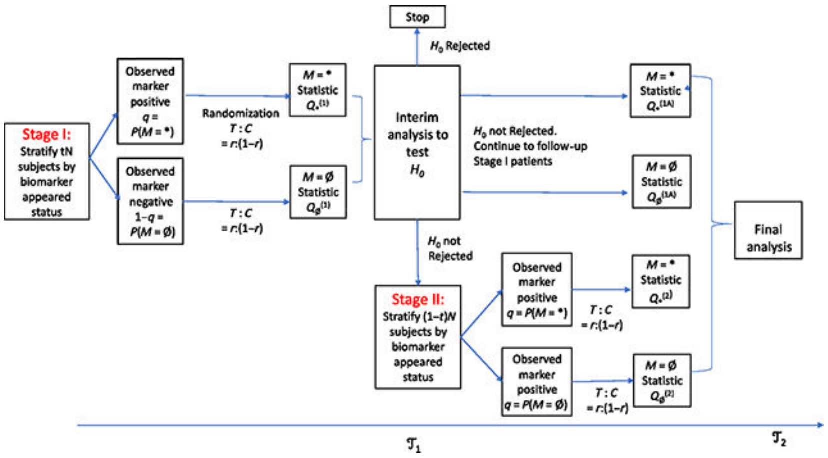
$$E(Q_i) = r(1-r)D_i\log\frac{h_{iT}}{h_{iC}} = r(1-r)D_i\delta_i,$$

and

$$\mathrm{Var}(Q_i) = r(1-r)D_i,$$

where $D_i$ is the expected total number of events in the $i$th marker status stratum.

## 3 | STRATIFIED DESIGN


Stop
$H_0$ Rejected
Stage I:
Stratify tN subjects by biomarker appeared status
Observed marker positive $q = P(M = *)$
Observed marker negative $1-q = P(M = \emptyset)$
Randomization
$T:C = r:(1-r)$
$T:C = r:(1-r)$
$M = *$ Statistic $Q_*^{(1)}$
$M = \emptyset$ Statistic $Q_\emptyset^{(1)}$
Interim analysis to test $H_0$
$H_0$ not Rejected. Continue to follow-up Stage I patients
$M = *$ Statistic $Q_*^{(1A)}$
$M = \emptyset$ Statistic $Q_\emptyset^{(1A)}$
$H_0$ not Rejected
Stage II:
Stratify $(1-t)N$ subjects by biomarker appeared status
Observed marker positive $q = P(M = *)$
Observed marker positive $q = P(M = \emptyset)$
$T:C = r:(1-r)$
$T:C = r:(1-r)$
$M = *$ Statistic $Q_*^{(2)}$
$M = \emptyset$ Statistic $Q_\emptyset^{(2)}$
Final analysis
$\mathcal{T}_1$
$\mathcal{T}_2$

Now, consider $0 < \lambda_{\text{sen}} \leq 1,\ 0 < \lambda_{\text{spec}} \leq 1$. The above diagram shows the proposed two-stage stratified design based on observed marker status. Let $N$ denote the total number of patients of the study. At Stage I, $tN$ patients, $0 < t < 1$, are stratified according to their observed marker status, and within each observed marker stratum, patients are randomized to either the active treatment or the control group with ratio of $r$ to $1 - r$. At the interim analysis time $\mathscr{T}_1$, and the information fraction $I_n$, the composite null hypothesis $H_0$ will be tested to determine whether to declare superiority early. If $H_0$ is not rejected, we will continue following up Stage I patients and enrolling $(1 - t)N$ Stage II patients and perform the final analysis at time $\mathscr{T}_2$.

### 3.1 | Adjusted log-rank statistics at Stage I

For Stage I, let $n_{mi}^{(1)}$ denote the expected number of patients in the observed marker group $k$ with true marker status $i$, and $n_m^{(1)}$ denote the expected number of subjects with observed marker status $m$, where $m = *$ or $\phi$ and $i = +$ or $-$. Then $n_{*+}^{(1)} = \tau n_*^{(1)}$ and $n_{*-}^{(1)} = (1 - \tau)n_*^{(1)}$. Similarly, $n_{\phi+}^{(1)} = (1 - \eta)n_\phi^{(1)}$ and $n_{\phi-}^{(1)} = \eta n_\phi^{(1)}$. At the interim analysis time $\mathscr{T}_1$, let the probability to observe an event for a true marker $i$ (including both active and control treatment in that stratum) be $\pi_i^{(1)}, i = +$ or $-$ for true biomarker status; and let the probability to observe an event for the observed marker stratum $m$ (including both active and control treatment in that stratum) be $\pi_m^{(1)}, m = *$ or $\phi$ for observed biomarker status. Then, the expected number of events at interim analysis time $\mathscr{T}_1$ for observed biomarker positive and negative groups are

$$n_*^{(1)}\pi_*^{(1)} = \tau n_*^{(1)}\pi_+^{(1)} + (1 - \tau)n_*^{(1)}\pi_-^{(1)} \text{ and } n_\phi^{(1)}\pi_\phi^{(1)} = (1 - \eta)n_\phi^{(1)}\pi_+^{(1)} + \eta n_\phi^{(1)}\pi_-^{(1)}.$$

When the biomarker assay is imperfect, that is, $0 < \lambda_{\text{sen}} < 1$ and/or $0 < \lambda_{\text{spec}} < 1$, the true biomarker status is unknown and only the marker appeared status can be observed. To assess the treatment effect for the true biomarker status, we start with the conventional log-rank statistic and apply it to each marker observed stratum for subjects recruited at Stage I. Let $Q_*^{(1)}$ and $Q_\phi^{(1)}$ be the log-rank statistics at the interim analysis for observed marker positive and observed marker negative stratum, respectively, then both $Q_*^{(1)}$ and $Q_\phi^{(1)}$ take the same form as the log-rank statistic $Q_i$ given in Section 2.1. Using Schoenfeld[19]‘s methods (Appendix A), asymptotically we have

$$E(Q_*^{(1)}) = r(1 - r)\tau n_*^{(1)}\pi_+^{(1)}\log\frac{h_{+T}}{h_{+C}} + r(1 - r)(1 - \tau)n_*^{(1)}\pi_-^{(1)}\log\frac{h_{-T}}{h_{-C}}, \tag{1}$$

$$E(Q_\phi^{(1)}) = r(1 - r)(1 - \eta)n_\phi^{(1)}\pi_+^{(1)}\log\frac{h_{+T}}{h_{+C}} + r(1 - r)\eta n_\phi^{(1)}\pi_-^{(1)}\log\frac{h_{-T}}{h_{-C}}, \tag{2}$$

With

$$\text{Var}(Q_*^{(1)}) = r(1 - r)\tau n_*^{(1)}\pi_+^{(1)} + r(1 - r)(1 - \tau)n_*^{(1)}\pi_-^{(1)},$$

$$\mathrm{Var}(Q_{\phi}^{(1)}) = r(1-r)(1-\eta)n_{\phi}^{(1)}\pi_{+}^{(1)} + r(1-r)\eta n_{\phi}^{(1)}\pi_{-}^{(1)}.$$

Note that both $E(Q_{*}^{(1)})$ and $E(Q_{\phi}^{(1)})$ are a linear combination of $\log(h_{+T}/h_{+C})$ and $\log(h_{-T}/h_{-C})$. Therefore, we can express both $\log(h_{+T}/h_{+C})$ and $\log(h_{-T}/h_{-C})$ as a function of $E(Q_{*}^{(1)})$ and $E(Q_{\phi}^{(1)})$ in the following,

$$\begin{aligned} r(1-r)n_{+}^{(1)}\pi_{+}^{(1)}\log\frac{h_{+T}}{h_{+C}} &= \frac{\tau n_{*}^{(1)} + (1-\eta)\, n_{\phi}^{(1)}}{n_{*}^{(1)}n_{\phi}^{(1)}} \times \frac{\eta n_{\phi}^{(1)}E\left(Q_{*}^{(1)}\right) - (1-\tau)n_{*}^{(1)}E\left(Q_{\phi}^{(1)}\right)}{\tau+\eta-1} \\ &= \frac{\tau q + (1-\eta)\,(1-q)}{q\,(1-q)} \times \frac{\eta\,(1-q)E\left(Q_{*}^{(1)}\right) - (1-\tau)qE\left(Q_{\phi}^{(1)}\right)}{\tau+\eta-1}, \end{aligned} \tag{3}$$

$$\begin{aligned} r(1-r)n_{-}^{(1)}\pi_{-}^{(1)}\log\frac{h_{-T}}{h_{-C}} &= \frac{(1-\tau)\, n_{*}^{(1)} + \eta n_{\phi}^{(1)}}{n_{*}^{(1)}n_{\phi}^{(1)}} \times \frac{-(1-\eta)n_{\phi}^{(1)}\, E\left(Q_{*}^{(1)}\right) + \tau n_{*}^{(1)}E\left(Q_{\phi}^{(1)}\right)}{\tau+\eta-1} \\ &= \frac{(1-\tau)\, q + \eta\,(1-q)}{q\,(1-q)} \times \frac{-(1-\eta)\,(1-q)\, E\left(Q_{*}^{(1)}\right) + \tau qE\left(Q_{\phi}^{(1)}\right)}{\tau+\eta-1}. \end{aligned} \tag{4}$$

Notice that the left-hand side of Equations (3) and (4) involve the log hazard ratios of treatment vs control arms in biomarker true positive and negative stratum, respectively. Then, we construct the following $Q_{+}^{(1)}$ and $Q_{-}^{(1)}$ statistics to test the treatment effects $\log(h_{+T}/h_{+C})$ and $\log(h_{-T}/h_{-C})$ in the true marker positive and true marker negative cohorts, respectively, by replacing $E(Q_{*}^{(1)})$ and $E(Q_{\phi}^{(1)})$ by $Q_{*}^{(1)}$ and $Q_{\phi}^{(1)}$ based on Equations (3) and (4):

$$Q_{+}^{(1)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \times \frac{\eta(1-q)Q_{*}^{(1)} - (1-\tau)qQ_{\phi}^{(1)}}{\tau+\eta-1}. \tag{5}$$

$$Q_{-}^{(1)} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \times \frac{-(1-\eta)(1-q)Q_{*}^{(1)} + \tau qQ_{\phi}^{(1)}}{\tau+\eta-1}. \tag{6}$$

Alternatively, $Q_{+}^{(1)}$ and $Q_{-}^{(1)}$ can be expressed as

$$Q_{+}^{(1)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \times \frac{\eta(1-q)\sum_{k=1}^{K^{(1)}}\left(d_{*Tk}^{(1)} - \frac{d_{*\cdot k}^{(1)}Y_{*Tk}^{(1)}}{Y_{**k}^{(1)}}\right) - (1-\tau)q\sum_{k=1}^{K^{(1)}}\left(d_{\phi Tk}^{(1)} - \frac{d_{\phi\cdot k}^{(1)}Y_{\phi Tk}^{(1)}}{Y_{\phi\cdot k}^{(1)}}\right)}{\tau+\eta-1},$$

$$Q_{-}^{(1)} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \times \frac{-(1-\eta)(1-q)\sum_{k=1}^{K^{(1)}}\left(d_{*Tk}^{(1)} - \frac{d_{*\cdot k}^{(1)}Y_{*k}^{(1)}}{Y_{*\cdot k}^{(1)}}\right) + \tau q\sum_{k=1}^{K^{(1)}}\left(d_{\phi Tk}^{(1)} - \frac{d_{\phi\cdot k}^{(1)}Y_{\phi Tk}^{(1)}}{Y_{\phi\cdot k}^{(1)}}\right)}{\tau+\eta-1},$$

where $Y_{ijk}^{(1)}$ denotes the number of subjects at risk at the $k$th distinct event time in the $i$th marker appeared stratum and the $j$th treatment group, and $d_{i\cdot k}^{(1)}$ and $Y_{i\cdot k}^{(1)}$ denotes the number of subjects at risk at the $k$th distinct event time in the $i$th marker appeared stratum, and $K^{(1)}$ is the total number of distinct event times in Stage I, for $i = *$ or $\phi$, $j = T$ or $C$.

Applying the expectations to (5) and (6) and use Equations (3) and (4), we have

$$E\left(Q_{+}^{(1)}\right) = r(1-r)n_{+}^{(1)}\pi_{+}^{(1)}\log\frac{h_{+T}}{h_{+C}},$$

$$E\left(Q_{-}^{(1)}\right) = r(1-r)n_{-}^{(1)}\pi_{-}^{(1)}\log\frac{h_{-T}}{h_{-C}}.$$

It can also be derived that

$$\operatorname{Var}\left(Q_{+}^{(1)}\right) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}\eta^{2}\operatorname{Var}\left(Q_{*}^{(1)}\right) + q^{2}(1-\tau)^{2}\operatorname{Var}\left(Q_{\phi}^{(1)}\right)\right\} \overset{\Delta}{=} \left(\sigma_{+}^{(1)}\right)^{2},$$

$$\operatorname{Var}\left(Q_{-}^{(1)}\right) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}(1-\eta)^{2}\operatorname{Var}\left(Q_{*}^{(1)}\right) + q^{2}\tau^{2}\operatorname{Var}\left(Q_{\phi}^{(1)}\right)\right\} \overset{\Delta}{=} \left(\sigma_{-}^{(1)}\right)^{2}.$$

Note that $\widehat{\operatorname{Var}}\left(Q_{*}^{(1)}\right)$ and $\widehat{\operatorname{Var}}\left(Q_{\phi}^{(1)}\right)$ take the same form as $\widehat{\operatorname{Var}}\left(Q_{i}\right)$ given in Section 2.1. Thus, $\operatorname{Var}\left(Q_{+}^{(1)}\right)$ and $\operatorname{Var}\left(Q_{-}^{(1)}\right)$ can be estimated by

$$\begin{aligned}\widehat{\operatorname{Var}}\left(Q_{+}^{(1)}\right) &= \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}\eta^{2}\sum_{k=1}^{K^{(1)}}\frac{Y_{*Tk}^{(1)}Y_{*Ck}^{(1)}d_{*\cdot k}^{(1)}\left(Y_{*\cdot k}^{(1)} - d_{*\cdot k}^{(1)}\right)}{\left(Y_{*\cdot k}^{(1)}\right)^{2}\left(Y_{*\cdot k}^{(1)} - 1\right)} + q^{2}(1-\tau)^{2}\sum_{k=1}^{K^{(1)}}\right.\\ &\left.\frac{Y_{\phi Tk}^{(1)}Y_{\phi Ck}^{(1)}d_{\phi\cdot k}^{(1)}\left(Y_{\phi\cdot k}^{(1)} - d_{\phi\cdot k}^{(1)}\right)}{\left(Y_{\phi\cdot k}^{(1)}\right)^{2}\left(Y_{\phi\cdot k}^{(1)} - 1\right)}\right\}\\ &\triangleq \left(\widehat{\sigma}_{+}^{(1)}\right)^{2},\end{aligned}$$

and

$$\begin{aligned}\widehat{\operatorname{Var}}\left(Q_{-}^{(1)}\right) &= \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}(1-\eta)^{2}\sum_{k=1}^{K^{(1)}}\frac{Y_{*k}^{(1)}Y_{*Ck}^{(1)}d_{*\cdot k}^{(1)}\left(Y_{*\cdot k}^{(1)} - d_{*\cdot k}^{(1)}\right)}{\left(Y_{*\cdot k}^{(1)}\right)^{2}\left(Y_{*\cdot k}^{(1)} - 1\right)} + q^{2}\tau^{2}\sum_{k=1}^{K^{(1)}}\right.\\ &\left.\frac{Y_{\phi Tk}^{(1)}Y_{\phi Ck}^{(1)}d_{\phi\cdot k}^{(1)}\left(Y_{\phi\cdot k}^{(1)} - d_{\phi\cdot k}^{(1)}\right)}{\left(Y_{\phi\cdot k}^{(1)}\right)^{2}\left(Y_{\phi\cdot k}^{(1)} - 1\right)}\right\}\\ &\triangleq \left(\widehat{\sigma}_{-}^{(1)}\right)^{2},\end{aligned}$$

respectively. Note that each of $Q_{+}^{(1)}$ and $Q_{-}^{(1)}$ corresponding to the true biomarker status is a mixture of $Q_{*}^{(1)}$ and $Q_{\phi}^{(1)}$, the log-rank statistics for the observed marker positive and observed marker negative groups. Moreover, $E(Q_j^{(1)}) = 0$ if there is no treatment effect in the true marker group $i$, that is, $H_{0i}\colon \log\frac{h_{iT}}{h_{iC}} = 0$ is true, for $i = +$ or $-$. Therefore we regard $Q_+^{(1)}$ and $Q_-^{(1)}$ as the un-standardized log-rank statistics for testing treatment effect on the true marker positive and negative groups, after adjusting for misclassification at the interim analysis time $\mathcal{T}_1$.

This further implies that the log-rank statistic for testing the treatment effect on the overall population $H_{0a}$ at time $\mathcal{T}_1$, after adjusting for misclassification, can be constructed as

$$Q^{(1)} = p\frac{Q_+^{(1)}}{\sigma_+^{(1)}} + (1-p)\frac{Q_-^{(1)}}{\sigma_-^{(1)}}.$$

with

$$\mathrm{Var}(Q^{(1)}) = p^2 + (1-p)^2 - 2p(1-p)\frac{\eta(1-\eta)(1-q)^2\mathrm{Var}\left(Q_*^{(1)}\right) + \tau(1-\tau)q^2\mathrm{Var}\left(Q_\phi^{(1)}\right)}{\sigma_+^{(1)}\sigma_-^{(1)}} \overset{\Delta}{=} \left(\sigma^{(1)}\right)^2,$$

which can be estimated by

$$\widehat{\mathrm{Var}}(Q^{(1)}) = p^2 + (1-p)^2 - 2p(1-p)\frac{\eta(1-\eta)(1-q)^2\widehat{\mathrm{Var}}(Q_*^{(1)}) + \tau(1-\tau)q^2\widehat{\mathrm{Var}}\left(Q_\phi^{(1)}\right)}{\sigma_+^{(1)}\sigma_-^{(1)}} \overset{\Delta}{=} (\widehat{\sigma}^{(1)})^2$$

Under $H_{0a}$, $E[Q^{(1)}] = 0$. Thus the standardized log-rank statistic for testing the treatment effect on the overall population $H_{0a}$ at interim analysis $\mathcal{T}_1$, after adjusting for misclassification, can be defined as

$$Z^{(1)} = \frac{Q^{(1)}}{\sqrt{\mathrm{Var}(Q^{(1)})}} = \frac{Q^{(1)}}{\sigma^{(1)}} = \frac{pZ_+^{(1)} + (1-p)Z_-^{(1)}}{\sigma^{(1)}},$$

where

$$Z_+^{(1)} = \frac{Q_+^{(1)}}{\sqrt{\mathrm{Var}(Q_+^{(1)})}} = \frac{Q_+^{(1)}}{\sigma_+^{(1)}}, Z_-^{(1)} = \frac{Q_-^{(1)}}{\sqrt{\mathrm{Var}(Q_-^{(1)})}} = \frac{Q_-^{(1)}}{\sigma_-^{(1)}},$$

and $Z_+^{(1)}$ and $Z_-^{(1)}$ are the standardized log-rank statistics for testing treatment effect on the true marker positive and negative groups at $\mathcal{T}_1$ after adjusting for misclassification, which can be estimated by $\widehat{Z}^{(1)} = \frac{Q^{(1)}}{\hat{\sigma}^{(1)}}$, $\widehat{Z}_+^{(1)} = \frac{Q_+^{(1)}}{\hat{\sigma}_+^{(1)}}$, and $\widehat{Z}_-^{(1)} = \frac{Q_-^{(1)}}{\hat{\sigma}_-^{(1)}}$, respectively.

### 3.2 | Adjusted log-rank statistics at final analysis

If $H_0$ is not rejected at the interim analysis time $\mathcal{T}_1$, we will enroll new subjects for Stage II while continue following up Stage I subjects to final analysis time. Thus, in the final data analysis, the test statistics would involve both Stage I subjects with prolonged follow-up until $\mathscr{T}_2$ and Stage II subjects that are recruited after $\mathscr{T}_1$ and followed up until $\mathscr{T}_2$. In the remainder of this paper, the same method to construct the corresponding statistics and variance estimates used in Section 3.1 are applied without repeating the form of statistics to ease notation.

#### 3.2.1 | Log-rank statistics at final analysis for subjects recruited at Stage I—

For subjects recruited at Stage I and continued being followed up until $\mathscr{T}_2$, let $Q_*^{(1A)}$ and $Q_\phi^{(1A)}$ denote the log-rank statistics corresponding to observed marker positive and observed marker negative groups at final analysis time $\mathscr{T}_2$, and let $\pi_i^{(1A)}$ denote the probability to observe an event in true marker status $i = +$ or $-$ group (including both active and control treatments) at final analysis time $\mathscr{T}_2$. Using the same arguments in Section 3.1, we can construct $Q_+^{(1A)}$ and $Q_-^{(1A)}$ as

$$Q_+^{(1A)} = \frac{\tau q + (1 - \eta)(1 - q)}{q(1 - q)} \times \frac{\eta(1 - q)Q_*^{(1A)} - (1 - \tau)qQ_\phi^{(1A)}}{\tau + \eta - 1},$$
$$Q_-^{(1A)} = \frac{(1 - \tau)q + \eta(1 - q)}{q(1 - q)} \times \frac{-(1 - \eta)(1 - q)Q_*^{(1A)} + \tau q Q_\phi^{(1A)}}{\tau + \eta - 1}.$$

Asymptotically, we have

$$E(Q_+^{(1A)}) = r(1-r)n_+^{(1)}\pi_+^{(1A)}\log\frac{h_{+T}}{h_{+C}},$$
$$E(Q_-^{(1A)}) = r(1-r)n_-^{(1)}\pi_-^{(1A)}\log\frac{h_{-T}}{h_{-C}},$$

with variances

$$(\sigma_+^{(1A)})^2 = \mathrm{Var}(Q_+^{(1A)}) = \left(\frac{\tau q + (1 - \eta)(1 - q)}{q(1 - q)(\tau + \eta - 1)}\right)^2 \left\{(1-q)^2\eta^2\mathrm{Var}(Q_*^{(1A)}) + q^2(1-\tau)^2\mathrm{Var}(Q_\phi^{(1A)})\right\},$$
$$\left(\sigma_-^{(1A)}\right)^2 = \mathrm{Var}(Q_-^{(1A)}) = \left(\frac{(1 - \tau)q + \eta(1 - q)}{q(1 - q)(\tau + \eta - 1)}\right)^2 \left\{(1-q)^2(1-\eta)^2\mathrm{Var}(Q_*^{(1A)}) + q^2\tau^2\mathrm{Var}(Q_\phi^{(1A)})\right\}.$$

#### 3.2.2 | Log-rank statistics for subjects recruited at Stage II—

Similarly, for subjects recruited at Stage II, we use the same arguments in Section 3.1, the log-rank statistics corresponding to true marker status $i = +$ or $-$:

$$Q_+^{(2)} = \frac{\tau q + (1 - \eta)(1-q)}{q(1 - q)} \times \frac{\eta(1 - q)Q_*^{(2)} - (1 - \tau)qQ_\phi^{(2)}}{\tau + \eta - 1},$$
$$Q_-^{(2)} = \frac{(1 - \tau)q + \eta(1 - q)}{q(1 - q)} \times \frac{-(1 - \eta)(1 - q)Q_*^{(2)} + \tau q Q_\phi^{(2)}}{\tau + \eta - 1},$$

where $Q_*^{(2)}$ and $Q_\phi^{(2)}$ are the log-rank statistics at final analysis time $\mathscr{T}_2$ for subjects recruited at Stage II according to their observed marker status. For Stage II enrolled subjects, let $\tilde{\pi}_i$ denotes the probability to observe an event in true marker group $i = +$ or $-$ (including both active and control treatment). Asymptotically, we have

$$E(Q_{+}^{(2)}) = r(1-r)n_{+}^{(2)}\widetilde{\pi}_{+}\log\frac{h_{+T}}{h_{+C}},$$
$$E(Q_{-}^{(2)}) = r(1-r)n_{-}^{(2)}\widetilde{\pi}_{-}\log\frac{h_{-T}}{h_{-C}},$$

and variances

$$\mathrm{Var}(Q_{+}^{(2)}) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}\eta^{2}\mathrm{Var}(Q_{*}^{(2)}) + q^{2}(1-\tau)^{2}\mathrm{Var}(Q_{\phi}^{(2)})\right\},$$
$$\mathrm{Var}(Q_{-}^{(2)}) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}(1-\eta)^{2}\mathrm{Var}(Q_{*}^{(2)}) + (q\tau)^{2}\mathrm{Var}(Q_{\phi}^{(2)})\right\}.$$

Again, each of $Q_{+}^{(1A)}$ and $Q_{-}^{(1A)}$ is a function of observed statistics $Q_{*}^{(1A)}$ and $Q_{\phi}^{(1A)}$, and each of $Q_{+}^{(2)}$ and $Q_{-}^{(2)}$ is a function of observed statistics $Q_{*}^{(2)}$ and $Q_{\phi}^{(2)}$. Moreover, $E(Q_{i}^{(1A)}) = 0$ and $E(Q_{i}^{(2)}) = 0$ if $H_{0i}\colon\log\frac{h_{iT}}{h_{iC}} = 0$ is true, for $i = +$ or $-$.

### 3.2.3 | Test statistics at final analysis time $\mathcal{T}_2$—To test the treatment effect in the final analysis $\mathcal{T}_2$, we construct $Q_{+}$ and $Q_{-}$ for true marker status as

$$Q_{+} = Q_{+}^{(1A)} + Q_{+}^{(2)} \text{ and } Q_{-} = Q_{-}^{(1A)} + Q_{-}^{(2)},$$

with variances $\mathrm{Var}(Q_i) = \sigma_i^2 = \mathrm{Var}(Q_i^{(1A)}) + \mathrm{Var}(Q_i^{(2)})$, for $i = +$ or $-$. We can also use weighted combination based on the amount of information from the two stages.

Therefore, to test $H_{0i}$, the treatment effect in true marker status $i$ at the final analysis time $\mathcal{T}_2$, for $i = +$ or $-$, the standardized log-rank statistic is

$$Z_i = \frac{Q_i}{\sqrt{\mathrm{Var}(Q_i)}} = \frac{Q_i^{(1A)} + Q_i^{(2)}}{\sigma_i},$$

where

$$\sigma_{+}^{2} = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}\eta^{2}\mathrm{Var}(Q_{*}^{(1A)}) + q^{2}(1-\tau)^{2}\mathrm{Var}(Q_{\phi}^{(1A)})\right\}$$
$$+ \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}\eta^{2}\mathrm{Var}(Q_{*}^{(2)}) + (q(1-\tau))^{2}\mathrm{Var}(Q_{\phi}^{(2)})\right\},$$

and

$$\sigma_{-}^{2} = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}(1-\eta)^{2}\mathrm{Var}(Q_{*}^{(1A)}) + q^{2}\tau^{2}\mathrm{Var}(Q_{\phi}^{(1A)})\right\}$$
$$+ \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^{2}\left\{(1-q)^{2}(1-\eta)^{2}\mathrm{Var}(Q_{*}^{(2)}) + (q\tau)^{2}\mathrm{Var}(Q_{\phi}^{(2)})\right\}.$$

To test the treatment in the overall population $H_{0a}$, we construct

$$Q = pQ_+/\sigma_+ + (1-p)Q\_/\sigma_-,$$

with

$$\sigma^2 = \mathrm{Var}(Q) = p^2 + (1-p)^2 - 2p(1-p)\frac{\eta(1-\eta)(1-q)^2\mathrm{Var}(Q_*) + \tau(1-\tau)q^2\mathrm{Var}(Q_\phi)}{\sigma_+\sigma_-}.$$

where

$$Q_* = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\left(Q_*^{(1A)} + Q_*^{(2)}\right) \text{ and } Q_\phi = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\left(Q_\phi^{(1A)} + Q_\phi^{(2)}\right).$$

Then we define the standardized log-rank statistic for testing $H_{0a}$ as

$$Z = \frac{Q}{\sqrt{\mathrm{Var}(Q)}} = \frac{Q}{\sigma} = \frac{pZ_+ + (1-p)Z_-}{\sigma}.$$

Note that, $E[Z_i] = 0$ under $H_{0i}$, and $E[Z] = 0$ under $H_{0a}$, for $i = +$ or $-$. Similar to Section 3.1, we can use the form of the $Q_i$ statistic and the variance estimate $\widehat{\mathrm{Var}(Q_i)}$ in Section 2.1 to re-express $Q_+^{(1A)}, Q_-^{(1A)}, Q_+^{(2)}, Q_-^{(2)}, Z, Z_+, Z_-$ and their corresponding variance estimates $\hat{\sigma}$ ‘s. To ease notation, the expressions of these $Q$- and $Z$-statistics and the corresponding variance estimates are omitted. $Z$, $Z_+^{(1)}$ and $Z_-^{(1)}$ are the standardized log-rank statistics for testing treatment effect on the overall, true marker positive and negative groups at $\mathscr{T}_2$ after adjusting for misclassification, which can be estimated by $\hat{Z} = \frac{Q}{\hat{\sigma}}$, $\hat{Z}_+ = \frac{Q_+}{\hat{\sigma}_+}$, and $\hat{Z}_- = \frac{Q_-}{\hat{\sigma}_-}$, respectively. Moreover, due to the possible marker misclassification and that Stage I subjects are followed through the end of Stage II, all the Z statistics constructed previously are correlated. Correlations between the standardized log-rank statistics $Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+$, and $Z_-$ are derived in Appendix C, and their asymptotic distributions are given in Appendix D. These results are crucial to determine critical values to control the type I error rate, calculate the power of the study, and determine the sample size, as discussed in the next a few sections.

### 3.3 | Type I error $\alpha$ allocation and critical values

To test the global hypothesis $H_0$, we can split the overall alpha (eg, $\alpha = 0.025$) between the Stage I and Stage II, following a traditional sequential design. Specifically, at the interim analysis, a fraction of the overall alpha, $\alpha_1$, can be allocated to test the global composite hypothesis $H_0 : \delta = \delta_+ = 0 (H_{0a} \cap H_{0+})$ for Stage I:

$$\begin{aligned}\alpha_1 &= P(\text{Reject } H_0 \mid H_0) = P_0(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2) \\ &= P_0(Z^{(1)} < -c_1) + P_0(Z^{(1)} \geq -c_1 \text{ and } Z_+^{(1)} < -c_2) = \alpha_{1a} + \alpha_{1b}\end{aligned} \quad (7)$$

The critical value $c_1$, defined as critical value to test overall population at interim analysis, can be obtained by allocating $\alpha_{1a}$, a portion of $\alpha_1$, for testing $H_{0a}$ first, that is, by solving $P_0(Z^{(1)} < -c_1) = \alpha_{1a}$. Once $c_1$ is determined, $c_2$, defined as critical value to test true positive population at interim analysis, can be solved for testing $H_{0+}$ in the above equation, or simply $P_0(Z^{(1)} \geq -c_1 \text{ or } Z_+^{(1)} < -c_2) = \alpha_{1b}$.

For Stage II, the overall alpha is left with $\alpha_2 = \alpha - \alpha_1$, where

$$\begin{aligned}\alpha_2 = \alpha - \alpha_1 &= P(\text{Not reject } H_0 \text{ at Stage I, Reject } H_0 \text{ at Stage II} \mid H_0)\\ &= P_0(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z < -b_1)\\ &\quad + P_0(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2) = \alpha_{2a} + (\alpha_2 - \alpha_{2a})\,.\end{aligned} \tag{8}$$

Then, given $c_1$ and $c_2$ that are determined previously, the critical value $b_1$, defined as critical value to test overall population at final analysis, is obtained by allocating $\alpha_{2a}$, a portion of $\alpha_2$, for testing $H_{0a}$. Then $b_2$, defined as critical value to test true marker positive population at final analysis, can be solved for testing $H_{0+}$ in the above equation again. Table 1 shows the critical values for $\alpha = 0.025, \alpha_1 = 0.004, \alpha_{1a} = \alpha_1/2, \alpha_2 = 0.021, \alpha_{2a} = \alpha_2/2, r = 0.5$, with interim analysis at information fraction $I_n$ at 0.3 or 0.5. Given these $\alpha$'s, we can see that the critical value $c_1$ for testing the treatment effect in the overall population at the interim analysis is not affected by the prevalence rate $p$, sensitivity $\lambda_{\text{sen}}$, specificity $\lambda_{\text{spec}}$ or the information time $I_n$; the critical value $c_2$ for testing marker-positive population at interim analysis is affected by $p$, $\lambda_{\text{sen}}$, $\lambda_{\text{spec}}$, but not by $I_n$. However, the critical values $b_1$ and $b_2$ for testing the overall population and true marker-positive population at final analysis time are affected by all the $p$, $\lambda_{\text{sen}}$, $\lambda_{\text{spec}}$ and $I_n$.

### 3.4 | Global and marginal power

Power for testing the treatment effect in the overall population is

$$\begin{aligned}1 - \beta_a &= P(\text{Reject } H_{0a} \mid H_1) = P(\text{Reject } H_{0a} \text{ at Stage } I \mid H_1) + \text{P}(\text{Reject } H_{0a} \text{ at Stage } II \mid H_1)\\ &= P_1(Z^{(1)} < -c_1) + p_{2a} = p_{1a} + p_{2a}\,.\end{aligned} \tag{9}$$

With the critical values used to control the Type I error of $\alpha$ in Section 3.3, the global and marginal powers can be found as follows. The global power is

$$\begin{aligned}1 - \beta = P(\text{Reject } H_0 \mid H_1) &= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+} \mid H_1)\\ &= P_1(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2) + P_1(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z < -b_1)\\ &\quad + P_1(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2)\\ &= p_1 + p_{2a} + p_{2+},\end{aligned} \tag{10}$$

where

$$\begin{aligned} p_1 &= P(\text{ Reject } H_{0a} \text{ or Reject } H_{0+} \text{ at Stage } I \mid H_1) \\ &= P_1(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2), \end{aligned}$$

and

$$\begin{aligned} p_{2a} &= P(\text{ Not reject } H_0 \text{ at Stage } I \text{ and Reject } H_{0a} \text{ at Stage } II \mid H_1) \\ &= P_1(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z < -b_1)\,. \end{aligned}$$

Power for testing the treatment effect in the true marker-positive population is

$$\begin{aligned} 1-\beta_+ &= P(\text{Reject } H_{0+} \mid H_1) \\ &= P(\text{Reject } H_{0+} \text{ at Stage } I \mid H_1) + P(\text{Reject } H_{0+} \text{ at Stage } II \mid H_1) \\ &= P_1(Z_+^{(1)} < -c_2) + P_1(Z^{(1)} \geq -c_1, Z_+^{(1)} \geq -c_2, Z_+ < -b_2)\,. \end{aligned} \tag{11}$$

Figures 1 to 3 show the contour plots of power surfaces for global (testing $H_1$), overall population (testing $H_{1a}$) and true marker-positive cohort (testing $H_{1+}$) hypotheses, respectively, across $-0.10 \geq \delta \geq -0.40$ and $-0.10 \geq \delta_+ \geq -0.40$ by $N$ (total enrolled subjects) and $p$ (true prevalence), assuming $\alpha = 0.025, \alpha_1 = 0.004, w_+ = p, w_- = 1-p, \delta = p\delta_+ + (1-p)\delta_{-,}$ $r = 0.5$.

The power increases as $N$, $p$, $\lambda_{\text{sen}}$, or $\lambda_{\text{spec}}$ increases. For example, with $p = 0.5$, and $N = 500, \delta = -0.15$, and $\delta_+ = -0.4$, Figure 1A with $\lambda_{\text{sen}} = \lambda_{\text{spec}} = 1$ shows power 70%. However, Figure 1D shows power only about 45%. We also see the $\lambda_{\text{sen}}$ has less impact than $\lambda_{\text{spec}}$ in terms of power.

Figure 1 shows that the global power increases with increasing treatment effect for overall population and positive population (decreasing $\delta$ and/or decreasing $\delta_+$). Figure 2 shows that the power for overall population increases with increasing treatment effect for overall population when treatment effect for positive population is fixed (decreasing $\delta$). Figure 3 shows that the power for positive subgroup increases with increasing treatment effect for positive population (decreasing $\delta_+$) but decreases with increasing treatment effect for overall population (decreasing $\delta$).

### 3.5 | Sample size calculations

Given the global type I error $\alpha$ and power $1-\beta$, assuming that we have the estimated prevalence rate $p$, sensitivity $\lambda_{\text{sen}}$ and specificity $\lambda_{\text{spec}}$ from previous studies, the sample size needed to detect the treatment effect can be found based on the formulas in Sections 3.3 and 3.4, after we specify the design parameters in these sections. Due to the complexity in the test statistics, there is no closed form of sample size calculations but needs to rely on numerical root finding methods. Specifically, given $\alpha, \alpha_1, \alpha_{1a}, \alpha_{2a}$, the randomization ratio $r$ and information time $I$, one may determine the critical values according to Equations (7) and (8) in Sections 3.3. Then, once the critical values are determined and given the global power for testing $H_0$, one may determine the total sample size based on prespecified effect sizes $\delta$ and $\delta_+$, using Equation (10) via numerical root finding methods. Based on the total sample size, one may determine the power for testing $H_{0a}$ and $H_{0+}$ by Equations (9) and (11), respectively.

Appendix E shows the total sample size based on testing the marker-positive $H_{1+}$ with 80% and 90% power, and $\alpha = 0.025, \alpha_1 = 0.004, r = 0.5$, and $\delta = p\ \log\theta_+ + (1-p)\log\theta_- = -0.15, \delta_+ = \log\theta_+ = -0.3, -0.4$, or −0.5. Table E1 is useful when the study plan sets a sufficient power (eg, 80% or 90%) in testing $H_{1+}$ for the marker-positive cohort. It shows the required total sample size $N$ and the resulting global power and the power for testing $H_{1a}$ for the overall population. In this case, the global power is always higher since it is to test the composite hypothesis $H_1$. The power for the overall population would be low when the emphasis is on the much hopeful marker-positive cohort. This is reflected by displaying $\delta = -0.15$ while $\delta_+ = -0.3$ to −0.5 in this table. In these situations, though the power is low for testing the overall treatment effect, the design has sufficient power to detect the treatment effect on the marker-positive subset. This implies that fewer patients are needed when the main objective of the study is to show treatment effect on the marker-positive cohort.

## 4 | NUMERIC EXAMPLES

Immunotherapy is a new paradigm for the treatment of non-small-cell lung cancer (NSCLC), and targeting the PD-1/PD-L1 pathway is a promising therapeutic option.[20] Pembrolizumab is a new immunotherapy that blocks the PD-1 pathway and restores the body's immune response against cancer cells and allows the immune system to recognize and kill cancer cells. As an illustration, we used our method to design a two-stage stratified trial with the KEYNOTE-10 trial.[21] as the prototype. KEYNOTE-10 is a phase II/III randomized trial to study Pembrolizumab vs Docetaxel for previously treated, PD-L1-positive, advanced NSCLC patients[21] in which the qualified patients were randomized by a biomarker's tumor proportion score (TPS, 1%–49% vs ≥50%) that measures the extent of PD-L1 expression. The companion diagnostic assay for PD-L1 expression was the Dako EnVision FLEX+HRP-Polymer kit using the 22C3 antibody clone, which was validated in the phase 1KEYNOTE-001trial.[22] Here for illustration, we use the phase 1 KEYNOTE-001 data as the basis to "redesign" the KEYNOTE-10 as an "imaginary" two-stage group sequential trial to assess the treatment effect of Pembrolizumab 2 mg (Pem), compared to Docetaxel (Dox) as the control (standard-of-care), in PD-L1 positive (TPS > 1%) NSCLC patients.

From the phase I KEYNOTE-001 trial, the prevalence was about 0.39 for TPS <1%, 0.38 for TPS = 1% to 49%, and 0.23 for TPS $\geq 50\%$; $\lambda_{\text{sen}} = \lambda_{\text{spec}} = 0.8$.[22] Thus, we estimated the prevalence rate of the PD-L1 true "strongly positive" (TPS $\geq 50\%$) to be $p \approx 0.40$, among the PD-L1 positive (TPS >1%) NSCLC patients; the observed PD-L1 "strongly positive" prevalence was $q = P(M = *) = p\lambda_{\text{sen}} + (1-p)(1-\lambda_{\text{spec}}) = 0.40 \times 0.80 + (1-0.40)(1-0.80) = 0.44$. Hence $\text{PPV} = \tau = \frac{p\lambda_{\text{sen}}}{p\lambda_{\text{sen}} + (1-p)(1-\lambda_{\text{spec}})} = \frac{p\lambda_{\text{sen}}}{q} = \frac{0.4 \times 0.8}{0.44} = 0.73$, and

$\mathrm{NPV} = \eta = \frac{(1-p)\lambda_{\mathrm{spec}}}{1-P(M=*)} = \frac{(1-p)\lambda_{\mathrm{spec}}}{1-q} = \frac{(1-0.4)\times 0.8}{1-0.44} = 0.86$. The real Phase II/III KEYNOTE-10 trial had both overall survival and progression-free survival (PFS) as primary end-points. Here we only used PFS as the primary endpoint for illustration purpose. Suppose that the overall (one-sided) $\alpha = 0.025$ and an interim analysis is planned at $I_n = 0.5$, with $\alpha_1 = 0.004$ allocated for Stage I. As stated in Herbst et al,[21] the study aimed to show a benefit of Pem over Dox in PFS in patients with TPS ≥ 50% as well as in the whole TPS ≥ 1% cohort, so we allocated $\alpha_{1a} = \alpha_{1b} = \frac{\alpha_1}{2} = 0.002$ which leaves $\alpha_2 = 0.021$ for Stage II; then we equally split $\alpha_2$ so that it was 0.0105 for testing the overall population and equally 0.0105 for the PD-L1 strongly positive subpopulation.

From Herbst et al,[21] the total enrollment time was 19 months for $N$ of 688 subjects. For this illustrative trial, we expect a total of 688 × 0.92 = 632 PFS events at the end of the trial. For Stage I, we plan to enroll 70% subjects (13.3 months from the start of the study). A decision is made at interim analysis $\mathscr{T}_1$ (information: $I_n = 0.5$, when 316 = 632×0.5 PFS events are expected. If the global null hypothesis $H_0$ is not rejected at interim analysis, additional 30% subjects will be enrolled in the second stage. After the recruitment is completed, the study will follow up to calendar time $\mathscr{T}_2$ for final analysis: the total number of 632 PFS are observed, with 461 PFS events from Stage I enrolled subjects and 171 PFS events from Stage II enrolled subjects. The number of PFS events from Stage I and Stage II enrolled subjects were determined, to make sure the number of events arrives approximately at the same calendar time.

To illustrate power calculation, we take the treatment effect information from Herbst et al.[21] Assume PFS times are exponentially distributed: For marker positive group, the hazard rate for treated patients: $h_{+T} = 1/9.90$, and hazard rate for placebo treated patients: $h_{+C} = 1/5.85$. For marker negative group, the hazard rate for treated patients: $h_{-T} = 1/5.14$ and for placebo-treated patients: $h_{-C} = 1/5.85$. These design parameters lead to the critical values $(c_1,\ c_2,\ b_1,\ b_2) = (2.878,\ 2.848,\ 2.269,\ 2.211)$ when $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 1$; critical values $(c_1, c_2, b_1, b_2) = (2.878,\ 2.874,\ 2.273,\ 2.260)$ when $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 0.8$. With a total of $N = 688$ patients and prevalence rate $p = 0.4$, we have 97% power to test global hypothesis $H_1$,97% power to test $H_{1+}$, and 5.7% power to test $H_{1a}$ when $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 1$. However, when $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 0.8$, with a total of $N = 688$ patients and prevalence rate $p = 0.4$, we have 72% power to test global hypothesis $H_1$,70% power to test $H_{1+}$, and 8.0% power to test $H_{1a}$. If $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}}$ is close to 0.8, say $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 0.775$, or 0.825, then we have 66% and 78% power to test global hypothesis $H_1$,63% and 76% power to test $H_{1+}$, and 8.1% and 8.0% power to test $H_{1a}$, respectively.

To illustrate the sample size calculation, with the same $\alpha$ allocation and the same treatment effects shown, assuming $\lambda_{\mathrm{sen}} = \lambda_{\mathrm{spec}} = 0.8$: the log-hazard ratio (Dox vs Pem) $\delta = -0.128$ for the overall population, and the log-hazard ratio (Dox vs Pem) $\delta_+ = -0.528$ for the marker positive cohort. Trials usually aim a power of either 80% or 90%. Assuming a target of 90% power for testing $H_{0+}$, then a total sample size $N$ of 1125 is needed. The global power for testing the composite hypothesis $H_0$ is 91 %, and the power for testing the overall population is 10%. If the target power for testing $H_{1+}$ is 80%, a total sample size $N$ of 855 is needed. The global power for testing the composite hypothesis $H_1$ is 82%, and the power for testing the overall population $H_{1a}$ is 9.2%.

## 5 | DISCUSSION

Misclassification of biomarker is common and needs to be properly adjusted in the clinical study designs. It is shown in Lin et al[18] that biomarker misclassification may have a big impact on type I error and power in the design of clinical studies. As shown in Shih and Lin,[23] stratified design is more efficient than precision medicine design, which is a family of complex biomarker-based-strategy designs discussed in, for instance, Mandrekar and Sargent[24–26] and Young et al.[27] However, to the best of our knowledge, no stratified study design for survival outcomes, based on predictive biomarkers with an adjustment of misclassification, has been proposed in the statistical literature.

A two-stage stratified study design with survival outcomes is proposed in this paper to adjust for the misclassification. We use the conventional log-rank statistics based on observed marker positive and observed marker negative strata to construct the adjusted log-rank statistics for true marker positive and true marker negative strata. Type I error control is achieved by utilizing correlation of the adjusted log-rank statistics from the same and/or different design stages with appropriate allocation of alpha’s.

At the design stage, we assume that the prevalence rate $p$, sensitivity $\lambda_{\text{sen}}$ and specificity $\lambda_{\text{spec}}$ are known and are not estimated in the study. As in many clinical study designs, these parameters can be externally estimated from the related literature and prior studies. In the numerical calculations (eg, Figures 1–3 and Table E1), we have observed that power (global, overall population, and marker positive cohort) increases with the prevalence rate, sensitivity, and specificity, while the sample size decreases with the prevalence rate, sensitivity, and specificity. Therefore, if these design parameters are over-estimated based on external resources, it may result in over-estimated power and under-estimated sample size. However, if the deviations of these design parameters are not too far off from the true values, the impact of power and sample size estimates may not be serious, as suggested by our numerical calculations.

Finally, to facilitate the implementation of our method, we have developed R programs to calculate power and sample size, in addition to the tables and graphs provided in this paper as illustrations. The R programs are available upon request.

Moreover, we plan to extend this misclassification adjustment method to an adaptive enrichment design with survival outcome, where we have a futility rule for biomarker negative stratum at the interim analysis, and possibly only enroll the biomarker positive patients after the interim analysis.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

The authors would like to thank an associate editor and two referees for their comments and suggestions that greatly improved the manuscript. The research of YL, SL and WJS was partially supported by NIH/NCI CCSG Grant P30CA072720.

### Funding information

NIH/NCI CCSG Grant, Grant/Award Number: P30CA072720

## DATA AVAILABILITY STATEMENT

Data used in the numerical examples are based on published papers.[20–22] R functions to implement our methods and simulations are included in Data S1.

## APPENDIX A.: GLOSSARY OF TERMS

See Table A1.

### TABLE A1

Glossary of terms.

| Notation | Definition |
|---|---|
| $p$ | Prevalence rate of true biomarker positive |
| $q$ | Observed biomarker positive rate |
| $\lambda_{\mathrm{sen}}$, $\lambda_{\mathrm{spec}}$ | Sensitivity, specificity of biomarker test, respectively |
| $I_n$ | Information time for interim analysis |
| $\tau$ | Positive predictive value |
| $\eta$ | The negative predictive value |
| $h_{ij}(t)$ | Hazard at time $t$: $i$ = + (positive) or − (negative) for true marker status, and $j = T$ for active treatment or $C$ for control treatment |
| N | Total number of subjects to be enrolled |
| $\delta_+$ and $\delta_-$ | Treatment effect for true biomarker positive: $\log\frac{h_{+T}(t)}{h_{+C}(t)}$ and true biomarker negative: $\log\frac{h_{-T}(t)}{h_{-C}(t)}$, respectively |
| $Q_*^{(1)}$ and $Q_\phi^{(1)}$ | Log-rank statistics at the interim analysis for observed marker positive and observed marker negative stratum, respectively |
| $Q_+^{(1)}$, $Q_-^{(1)}$ and $Q^{(1)}$ | Log-rank statistics for testing treatment effect on the true marker positive, negative groups, and overall population at the interim analysis time, respectively |
| $Q_+^{(1A)}$ and $Q_-^{(1A)}$ | Log-rank statistics at final analysis time $\mathcal{T}_2$ for subjects recruited at Stage I according to their observed marker status |
| $Q_+^{(2)}$ and $Q_\phi^{(2)}$ | Log-rank statistics at final analysis time $\mathcal{T}_2$ for subjects recruited at Stage II according to their observed marker status |
| $Q_+$, $Q_-$ and $Q$ | Log-rank statistics for testing treatment effect on the true marker positive, negative groups, and overall population at the final analysis time, respectively |
| $n_{mi}^{(1)}$ | Expected number of patients in the observed marker group $m$ =* or $\phi$ and $i$ = + or For example: $n_{*+}^{(1)} = \tau n_*^{(1)}$ |
| $n_{mi}^{(2)}$ | Expected number of patients in the observed marker group $m$ =* or $\phi$ and $i$ = + or −, for subjects enrolled after interim analysis |
| r | Randomization ratio: r to active treatment and $1 - r$ to placebo treatment |

| Notation | Definition |
|---|---|
| $t$ | Proportion of subjects to be enrolled before interim analysis |
| $\pi_i^{(1)}, \pi_i^{(1A)}$ | The probability to observe an event in true marker group $i$ : $i$ = + or − for at interim, final analysis time for Stage I enrolled subjects |
| $\tilde{\pi}_i$ | The probability to observe an event in true marker group i (including both active and control treatment) at final analysis time for Stage II enrolled subjects |
| $c_1$, $c_2$, $b_1$, and $b_2$ | Critical value for overall population, marker-positive population at interim analysis, overall population and marker-positive population at final analysis time, respectively |

## APPENDIX B.: ASYMPTOTIC DISTRIBUTIONS OF THE LOG-RANK TEST STATISTIC

Let $H_{ij}(t)$ denote the distribution function of the censoring, and $f_{ij}(t)$, and $S_{ij}(t)$ denote the density and distribution function of the survival, where $i = +$ or − and $j = T$ or $C$. Consider the local alternatives (Pitman alternatives)[28–30] $H_a: h_{iT}(t) \neq h_{iC}(t)$, with $H_{1i}: \log(h_{iT}/h_{iC}) = d_i/\sqrt{n}$, for some constant $d_i$, for $i = +$ or − and $n$ = total sample size. Then Schoenfeld[19] showed that under alternative $H_a: h_{iT}(t) \neq h_{iC}(t)$, satisfying $\log(h_{iT}(t)/h_{iC}(t)) = O\left(n^{-1/2}\right)$,

$$Z_i = \frac{Q_i}{\sqrt{\widehat{\mathrm{Var}}(Q_i)}} \sim \mathrm{AN}(\psi_i, 1), \tag{B1}$$

with

$$\psi_i = \frac{\sqrt{n_i}\int_0^\infty \log(h_{iT}(t)/h_{iC}(t))r_i(t)(1 - r_i(t))V_i(t)dt}{\left(\int_0^\infty r_i(t)(1 - r_i(t))V_i(t)dt\right)^{1/2}},$$

where $V_i(t) = (1 - r)f_{iC}(t)(1 - H_{iC}(t)) + rf_{iT}(t)(1 - H_{iT}(t))$ is the process of observing events,

$$r_i(t) = \frac{rS_{iT}(t)(1 - H_{iT}(t))}{(1 - r)S_{iC}(t)(1 - H_{iC}(t)) + rS_{iT}(t)(1 - H_{iT}(t))}.$$

In the proof of (B1) in Schoenfeld,[19] using the law of large numbers, it is easy to see that,

$$\frac{1}{n}Q_i \xrightarrow{P} \int_0^\infty \log(h_{iT}(t)/h_{iC}(t))r_i(t)(1 - r_i(t))V_i(t)dt,$$

and

$$\frac{1}{n}\widehat{Var}(Q_i) \xrightarrow{P} \int_0^\infty r_i(t)(1 - r_i(t))V_i(t)dt,$$

hence

$$\frac{1}{n}Var(Q_i) \to \int_0^\infty r_i(t)(1 - r_i(t))V_i(t)dt\,.$$

Suppose that $h_{iT}(t)/h_{iC}(t)$ is a constant and $H_{iC}(t) = H_{iT}(t)$ as in log-rank test. Since $S_{iT}(t) \to S_{iC}(t), r_i(t) \to r$, we have

$$\begin{aligned}\int_0^\infty \log(h_{iT}(t)/h_{iC}(t))r_i(t)(1 - r_i(t))V_i(t)dt &\to \log(h_{iT}/h_{iC})r(1 - r)\int_0^\infty V_i(t)dt \\ &= \log(h_{iT}/h_{iC})r(1 - r)\pi_i,\end{aligned}$$

and

$$\int_0^\infty r_i(t)(1 - r_i(t))V_i(t)dt \to r(1 - r)\pi_i,$$

where $\pi_i = \int_0^\infty V_i(t)dt$ is the combined probability of event. Hence, asymptotically,

$$E(Q_i) = n_i \log(h_{iT}/h_{iC})r(1 - r)\pi_i = \log(h_{iT}/h_{iC})r(1 - r)D_i,$$

and

$$Var(Q_i) = n_i r(1 - r)\pi_i = r(1 - r)D_i,$$

where $D_i = n_i\pi_i$ is the expected total number of events in $i$th group, and

$$\psi_i = \frac{\sqrt{n_i}\log(h_{iT}/h_{iC})r(1 - r)\pi_i}{(r(1 - r)\pi_i)^{1/2}} = \sqrt{r(1 - r)D_i}\log\left(\frac{h_{iT}}{h_{iC}}\right),$$

## APPENDIX C.: CORRELATIONS BETWEEN TEST STATISTICS

Because all the $Z$ statistics constructed previously are correlated. Correlations between the standardized log-rank statistics $Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+$, and $Z_-$ are shown below. These results are crucial to determine critical values to control the type I error rate, calculate the power of the study, and determine the sample size. Let

$$
\begin{aligned}
A &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}, & B &= \frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)},\\
F_1 &= (1-q)^2\eta^2\mathrm{Var}(Q_*^{(1)}) + q^2(1-\tau)^2\mathrm{Var}(Q_\phi^{(1)}), & F_2 &= (1-q)^2\eta^2\mathrm{Var}(Q_*) + q^2(1-\tau)^2\mathrm{Var}(Q_\phi),\\
G_1 &= \eta(1-\eta)(1-q)^2\mathrm{Var}(Q_*^{(1)}) + \tau(1-\tau)q^2\mathrm{Var}(Q_\phi^{(1)}), & G_2 &= \eta(1-\eta)(1-q)^2\mathrm{Var}(Q_*) + \tau(1-\tau)q^2\mathrm{Var}(Q_\phi),\\
H_1 &= (1-q)^2(1-\eta)^2\mathrm{Var}(Q_*^{(1)}) + q^2\tau^2\mathrm{Var}(Q_\phi^{(1)}), & H_2 &= (1-q)^2(1-\eta)^2\mathrm{Var}(Q_*) + q^2\tau^2\mathrm{Var}(Q_\phi),\\
C_1 &= \frac{p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_1}}, & D_1 &= \frac{1-p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_1}},\\
C_2 &= \frac{p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}}, & D_2 &= \frac{1-p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}},\\
E &= C_1C_2\frac{\sigma_+^{(1)}}{\sigma_+} + D_1D_2\frac{\sigma_-^{(1)}}{\sigma_-} - ABG_1\left(\frac{C_1D_2}{\sigma_-\sigma_1^{(1)}} + \frac{C_2D_1}{\sigma_+\sigma_-^{(1)}}\right).
\end{aligned}
$$

The correlation matrix between the standardized log-rank statistics is:

$$
Cov((Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+, Z_-)^T)
$$
$$
= \begin{bmatrix}
1 & C_1 - \frac{ABD_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & D_1 - \frac{ABC_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & E & C_1\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_1G_1}{\sigma_+\sigma_-^{(1)}} & \frac{-ABC_1G_1}{\sigma_-\sigma_+^{(1)}} + D_1\frac{\sigma_-^{(1)}}{\sigma_-}\\
 & 1 & \frac{-ABG_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & C_2\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_2G_1}{\sigma_-\sigma_+^{(1)}} & \sigma_+^{(1)}/\sigma_+ & \frac{-ABG_1}{\sigma_-\sigma_+^{(1)}}\\
 & & 1 & -\frac{ABC_2G_1^{+}}{\sigma_+\sigma_-^{(1)}} + D_2\sigma_0^{(1)}/\sigma_- & \frac{-ABG_1}{\sigma_+\sigma_-^{(1)}} & \sigma_-^{(1)}/\sigma_-\\
 & & & 1 & C_2 - \frac{ABD_2G_2}{\sigma_+\sigma_-} & D_2 - \frac{ABC_2G_2}{\sigma_+\sigma_-}\\
 & & & & 1 & -\frac{ABG_2}{\sigma_+\sigma_-}\\
 & & & & & 1
\end{bmatrix}
$$

## APPENDIX D.: ASYMPTOTIC DISTRIBUTION OF TEST STATISTICS

Given $t, N, p, \lambda_{\text{sen}}, \lambda_{\text{spec}}$, then $q, \tau, \eta$ are fixed. Under the alternative, given $\pi_+^{(1)}, \pi_+^{(2)}, \widetilde{\pi}_+, \pi_-^{(1)}, \pi_-^{(2)}, \widetilde{\pi}_-$, and assumption of proportional hazard $\log\theta_+ = \log\frac{h_{+T}}{h_{+C}}$, and $\log\theta_- = \log\frac{h_{-T}}{h_{-C}}$. Asymptotically,[19,31] we have

$$Z_{+}^{(1)} \sim \mathrm{AN}\left(\sqrt{tNr(1-r)}\frac{\pi_{+}^{(1)}\log\theta_{+}}{\sqrt{m_1}}, 1\right), Z_{-}^{(1)} \sim \mathrm{AN}\left(\sqrt{tNr(1-r)}\frac{\pi_{-}^{(1)}\log\theta_{-}}{\sqrt{m_0}}, 1\right),$$

$$Z^{(1)} \sim \mathrm{AN}\left(\frac{\sqrt{tNr(1-r)}\left\{p\pi_{+}^{(1)}\log\theta_{+}/\sqrt{m_1} + (1-p)\pi_{-}^{(1)}\log\theta_{-}/\sqrt{m_0}\right\}}{\sqrt{p^2 + (1-p)^2 - 2p(1-p)m_2/\sqrt{m_1m_0}, 1}}\right),$$

$$Z_{+} \sim \mathrm{AN}\left(\sqrt{Nr(1-r)}\frac{t\pi_{+}^{(2)} + (1-t)\widetilde{\pi}_{+}}{\sqrt{m_3}}\log\theta_{+}, 1\right),$$

$$Z_{-} \sim \mathrm{AN}\left(\sqrt{Nr(1-r)}\frac{t\pi_{-}^{(2)} + (1-t)\widetilde{\pi}_{-}}{\sqrt{m_4}}\log\theta_{-}, 1\right),$$

$$Z \sim \mathrm{AN}\left(\frac{\sqrt{Nr(1-r)}\left\{p\left[t\pi_{+}^{(2)} + (1-t)\widetilde{\pi}_{+}\right]\frac{\log\theta_{+}}{\sqrt{m_3}} + (1-p)\left[t\pi_{-}^{(2)} + (1-t)\widetilde{\pi}_{-}\right]\frac{\log\theta_{-}}{\sqrt{m_4}}\right\}}{\sqrt{p^2 + (1-p)^2 - 2p(1-p)m_5/\sqrt{m_3m_4}}}, 1\right),$$

where

$$\begin{aligned}
&m_1 = \eta^2(1-q)^2\{\tau q\pi_{+}^{(1)} + (1-\tau)q\pi_{-}^{(1)}\}A^2 + (1-\tau)^2q^2\{(1-\eta)(1-q)\pi_{+}^{(1)} + \eta(1-q)\pi_{-}^{(1)}\}A^2,\\
&m_0 = (1-\eta)^2(1-q)^2\{\tau q\pi_{+}^{(1)} + (1-\tau)q\pi_{-}^{(1)}\}B^2 + \tau^2q^2\{(1-\eta)(1-q)\pi_{+}^{(1)} + \eta(1-q)\pi_{-}^{(1)}\}B^2,\\
&m_2 = AB\eta(1-\eta)(1-q)^2\{r(1-r)tN[\tau q\pi_{+}^{(1)} + AB(1-\tau)q\pi_{-}^{(1)}]\} + \tau(1-\tau)q^2\\
&\{r(1-r)tN[(1-\eta)(1-q)\pi_{+}^{(1)} + \eta(1-q)\pi_{-}^{(1)}]\},\\
&m_3 = A^2\eta^2(1-q)^2\{t[\tau q\pi_{+}^{(2)} + (1-\tau)q\pi_{-}^{(2)}] + (1-t)[\tau q\widetilde{\pi}_{+} + (1-\tau)q\widetilde{\pi}_{-}]\}\\
&\quad + A^2(1-\tau)^2q^2\{t[(1-\eta)(1-q)\pi_{+}^{(2)} + \eta(1-q)\pi_{-}^{(2)}] + (1-t)[(1-\eta)(1-q)\widetilde{\pi}_{+} + \eta(1-q)\widetilde{\pi}_{-}]\},\\
&m_4 = B^2(1-\eta)^2(1-q)^2\{t[\tau q\pi_{+}^{(2)} + (1-\tau)q\pi_{-}^{(2)}] + (1-t)[\tau q\widetilde{\pi}_{+} + (1-\tau)q\widetilde{\pi}_{-}]\}\\
&\quad + B^2\tau^2q^2\{t[(1-\eta)(1-q)\pi_{+}^{(2)} + \eta(1-q)\pi_{-}^{(2)}] + (1-t)[(1-\eta)(1-q)\widetilde{\pi}_{+} + \eta(1-q)\widetilde{\pi}_{-}]\},\\
&m_5 = AB\eta(1-\eta)(1-q)^2r(1-r)\{tN[\tau q\pi_{+}^{(2)} + (1-\tau)q\pi_{-}^{(2)}] + (1-t)N[\tau q\widetilde{\pi}_{+} + (1-\tau)q\widetilde{\pi}_{-}]\}\\
&\quad + AB\tau(1-\tau)q^2r(1-r)\{tN[(1-\eta)(1-q)\pi_{+}^{(2)} + \eta(1-q)\pi_{-}^{(2)}] + (1-t)N[(1-\eta)(1-q)\widetilde{\pi}_{+} + \eta(1-q)\widetilde{\pi}_{-}]\}\ .
\end{aligned}$$

## APPENDIX E.: TOTAL SAMPLE SIZE TABLES FOR SELECTED SCENARIOS

### TABLE E1

Total sample size when $\alpha = 0.025$, $\alpha_1 = 0.004$, $r = 0.5$, and $\delta = p \log \theta_{+} + (1 - p) \log \theta_{-} = -0.15, \delta_{+} = \log \delta_{+} = -0.3, -0.4, or - 0.5$.

| | | | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** |
|---|---|---|---|---|---|---|---|---|---|---|
| **($\lambda_{sen}$, $\lambda_{spen}$)** | | | **80% power for testing $H_{1+}$** | | | | | | | |
| **$p$** | **$\delta_+$** | | **Information $I_n$ = 0.3** | | | | **Information $I_n$ = 0.5** | | | |
| 0.3 | −0.3 | $N$ | 2162 | 11 538 | 2892 | 23 898 | 2292 | 6993 | 3347 | 14136 |
| | | $H_1$ | 0.904 | 1.000 | 0.945 | 1.000 | 0.923 | 1.000 | 0.971 | 1.000 |
| | | $H_{1a}$ | 0.557 | 0.768 | 0.666 | 0.920 | 0.518 | 0.760 | 0.628 | 0.917 |
| | −0.4 | $N$ | 1162 | 1983 | 1482 | 3258 | 1162 | 2037 | 1495 | 3561 |
| | | $H_1$ | 0.836 | 0.883 | 0.856 | 0.934 | 0.840 | 0.895 | 0.862 | 0.956 |
| | | $H_{1a}$ | 0.252 | 0.365 | 0.300 | 0.489 | 0.218 | 0.330 | 0.264 | 0.457 |
| | −0.5 | $N$ | 749 | 1272 | 949 | 2031 | 743 | 1271 | 943 | 2046 |
| | | $H_1$ | 0.815 | 0.836 | 0.824 | 0.863 | 0.816 | 0.840 | 0.825 | 0.870 |
| | | $H_{1a}$ | 0.125 | 0.167 | 0.142 | 0.212 | 0.104 | 0.147 | 0.122 | 0.194 |
| 0.4 | −0.3 | $N$ | 1516 | 2295 | 1990 | 12 127 | 1538 | 2474 | 2085 | 7285 |
| | | $H_1$ | 0.856 | 0.914 | 0.893 | 1.000 | 0.864 | 0.935 | 0.909 | 0.999 |

| | | | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $H_{1a}$ | 0.443 | 0.599 | 0.544 | 0.807 | 0.398 | 0.564 | 0.504 | 0.804 |
| | −0.4 | $N$ | 842 | 1233 | 1077 | 1982 | 836 | 1236 | 1075 | 2043 |
| | | $H_1$ | 0.817 | 0.840 | 0.830 | 0.884 | 0.818 | 0.845 | 0.833 | 0.897 |
| | | $H_{1a}$ | 0.204 | 0.277 | 0.248 | 0.399 | 0.170 | 0.241 | 0.213 | 0.361 |
| | −0.5 | $N$ | 547 | 802 | 695 | 1226 | 541 | 796 | 689 | 1264 |
| | | $H_1$ | 0.807 | 0.816 | 0.812 | 0.835 | 0.807 | 0.818 | 0.813 | 0.839 |
| | | $H_{1a}$ | 0.107 | 0.136 | 0.124 | 0.183 | 0.085 | 0.114 | 0.101 | 0.159 |
| 0.5 | −0.3 | $N$ | 1148 | 1527 | 1500 | 2663 | 1147 | 1558 | 1527 | 3172 |
| | | $H_1$ | 0.827 | 0.860 | 0.857 | 0.945 | 0.831 | 0.870 | 0.867 | 0.974 |
| | | $H_{1a}$ | 0.396 | 0.517 | 0.508 | 0.754 | 0.348 | 0.473 | 0.464 | 0.747 |
| | −0.4 | $N$ | 647 | 852 | 830 | 1356 | 641 | 847 | 824 | 1374 |
| | | $H_1$ | 0.807 | 0.818 | 0.817 | 0.854 | 0.808 | 0.820 | 0.818 | 0.863 |
| | | $H_{1a}$ | 0.201 | 0.267 | 0.259 | 0.420 | 0.162 | 0.225 | 0.217 | 0.375 |
| | −0.5 | $N$ | 422 | 557 | 537 | 872 | 417 | 552 | 532 | 868 |
| | | $H_1$ | 0.804 | 0.807 | 0.806 | 0.822 | 0.803 | 0.808 | 0.806 | 0.824 |
| | | $H_{1a}$ | 0.116 | 0.151 | 0.145 | 0.231 | 0.088 | 0.120 | 0.115 | 0.195 |
| | | | **90% power for testing $H_{1+}$** | | | | | | | |
| 0.3 | −0.3 | $N$ | 3519 | 18 317 | 12 027 | 30 726 | 4005 | 10 829 | 7159 | 18195 |
| | | $H_1$ | 0.988 | 1.000 | 1.000 | 1.000 | 0.995 | 1.000 | 1.000 | 1.000 |
| | | $H_{1a}$ | 0.684 | 0.877 | 0.756 | 0.970 | 0.616 | 0.872 | 0.750 | 0.969 |
| | −0.4 | $N$ | 1561 | 2820 | 2027 | 14 594 | 1567 | 3055 | 2072 | 8538 |
| | | $H_1$ | 0.930 | 0.967 | 0.946 | 1.000 | 0.933 | 0.978 | 0.951 | 1.000 |
| | | $H_{1a}$ | 0.316 | 0.464 | 0.377 | 0.555 | 0.256 | 0.394 | 0.313 | 0.541 |
| | −0.5 | $N$ | 988 | 1697 | 1256 | 2751 | 980 | 1705 | 1250 | 2837 |
| | | $H_1$ | 0.912 | 0.929 | 0.919 | 0.949 | 0.912 | 0.933 | 0.920 | 0.957 |
| | | $H_{1a}$ | 0.153 | 0.209 | 0.175 | 0.270 | 0.118 | 0.171 | 0.139 | 0.230 |
| 0.4 | −0.3 | $N$ | 2106 | 7276 | 3025 | 18 426 | 2159 | 4727 | 3354 | 10 901 |
| | | $H_1$ | 0.947 | 1.000 | 0.978 | 1.000 | 0.953 | 0.998 | 0.988 | 1.000 |
| | | $H_{1a}$ | 0.549 | 0.698 | 0.666 | 0.905 | 0.474 | 0.670 | 0.598 | 0.903 |
| | −0.4 | $N$ | 1116 | 1662 | 1440 | 2828 | 1107 | 1676 | 1440 | 3082 |
| | | $H_1$ | 0.914 | 0.933 | 0.925 | 0.968 | 0.914 | 0.937 | 0.927 | 0.980 |
| | | $H_{1a}$ | 0.254 | 0.349 | 0.311 | 0.503 | 0.197 | 0.285 | 0.250 | 0.429 |
| | −0.5 | $N$ | 719 | 1059 | 915 | 1685 | 710 | 1051 | 907 | 1693 |
| | | $H_1$ | 0.906 | 0.913 | 0.909 | 0.927 | 0.904 | 0.913 | 0.909 | 0.931 |
| | | $H_{1a}$ | 0.128 | 0.167 | 0.150 | 0.228 | 0.093 | 0.129 | 0.114 | 0.186 |
| 0.5 | −0.3 | $N$ | 1546 | 2147 | 2096 | 11083 | 1543 | 2224 | 2161 | 6592 |
| | | $H_1$ | 0.923 | 0.952 | 0.949 | 1.000 | 0.924 | 0.959 | 0.956 | 1.000 |
| | | $H_{1a}$ | 0.491 | 0.631 | 0.620 | 0.862 | 0.414 | 0.563 | 0.551 | 0.862 |
| | −0.4 | $N$ | 855 | 1133 | 1101 | 1865 | 845 | 1125 | 1093 | 1914 |
| | | $H_1$ | 0.906 | 0.915 | 0.914 | 0.945 | 0.905 | 0.916 | 0.914 | 0.951 |
| | | $H_{1a}$ | 0.249 | 0.334 | 0.323 | 0.522 | 0.187 | 0.265 | 0.255 | 0.445 |
| | −0.5 | $N$ | 555 | 734 | 707 | 1157 | 549 | 726 | 700 | 1152 |

| | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** | **(1, 1)** | **(1, 0.8)** | **(0.8, 1)** | **(0.8, 0.8)** |
|---|---|---|---|---|---|---|---|---|
| $H_1$ | 0.903 | 0.906 | 0.905 | 0.917 | 0.902 | 0.905 | 0.905 | 0.918 |
| $H_{1a}$ | 0.139 | 0.185 | 0.177 | 0.289 | 0.097 | 0.136 | 0.129 | 0.227 |

## REFERENCES

1. Bailey AM, Mao Y, Zeng J, et al. Implementation of biomarker-driven cancer therapy: existing tools and remaining gaps. Discov Med. 2014;17(92):101–114. [PubMed: 24534473]
2. Kurnit K, Bailey AM, Zeng J, et al. “Personalized cancer therapy”: a publicly available precision oncology resource. Cancer Res. 2017;77:e123–e126. [PubMed: 29092956]
3. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med. 2005;24:329–339. [PubMed: 15551403]
4. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res. 2005;10:6759–6763.
5. Simon R The use of genomics in clinical trial design. Clin Cancer Res. 2008;14(19):5984–5993. [PubMed: 18829477]
6. Renfro L, Mallick H, An M-W, Sargent D, Mandrekar S. Clinical trial designs incorporating predictive biomarkers. Cancer Treat Rev. 2016;43:74–82. [PubMed: 26827695]
7. Kim ES, Hirsh V, Mok T, et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (interest): a randomised phase iii trial. Lancet. 2008;372(9652):1809–1818. [PubMed: 19027483]
8. Wakelee H, Kernstine K, Vokes E, et al. Cooperative group research efforts in lung cancer 2008: focus on advanced-stage non-small-cell lung cancer. Clin Lung Cancer. 2008;9(6):346–351. [PubMed: 19073517]
9. Shih W, Lin Y. On study designs and hypotheses for clinical trials with predictive biomarkers. Contemp Clin Trials. 2017;62:140–145. [PubMed: 28838813]
10. Krisam J, Kieser M. Decision rules for subgroup selection based on a predictive biomarker. J Biopharm Stat. 2014;24:188–202. [PubMed: 24392985]
11. Freidlin B, Korn EL, Gray R. Marker sequential test (MaST) design. Clin Trials. 2014;11:19–27. [PubMed: 24085774]
12. Wang S-J, O’Neill R, Hung H. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat. 2007;6:227–244. [PubMed: 17688238]
13. Wang S-J, Hung H, O’Neill R. Genomic classifier for patient enrichment:misclassification and type i error issues in pharmacogenomics noninferiority trial. Stat Biopharm Res. 2011;3:310–319.
14. Wang S-J, Lim M Impacts of predictive genomic classifier performance on subpopulation-specific treatment effects assessment. StatBiosci. 2016;8:129–158.
15. Liu C, Liu A, Hu J, Yuan V, Halabi S. Adjusting for misclassification in a stratified biomarker clinical trial. Stat Biosci. 2014;33: 3100–3113.
16. Zang Y, Lee J, Yuan Y. Two-stage marker-stratified clinical trial design in the presence of biomarker misclassification. J R Stat Soc Appl Stat. 2016;65:585–601.
17. Zang Y, Guo B. Optimal two-stage enrichment design correcting for biomarker misclassification. Stat Methods Med Res. 2018;27:35–47. [PubMed: 26614756]
18. Lin Y, Shih W, Lu S-E. Two-stage enrichment clinical trial design with adjustment for misclassification in predictive biomarkers. Stat Med. 2019;38:5445–5469. [PubMed: 31621944]
19. Schoenfeld D The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika. 1981;68:316–319.
20. Iwai Y, Ishida M, Tanaka K, Okazaki T, Honjo T, Minato N. Involvement of PD-L1 on tumor cells in the escape from host immune system and tumor immunotherapy by PD-L1 blockade. Proc Natl Acad Sci U S A. 2002;99:12293–12297. [PubMed: 12218188]

21. Herbst R, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, pd-l1-positive, advanced non-small-cell lung cancer (keynote-010): a randomised controlled trial. Lancet. 2016;387:1540–1550. [PubMed: 26712084]

22. Garon E, Rizvi N, Hui R, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Eng J Med. 2015;372:2018–2028.

23. Shih W, Lin Y. Relative efficiency of precision medicine designs for clinical trials with predictive biomarkers. Stat Med. 2018; 54(3):411–424.

24. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat. 2009;19:530–542. [PubMed: 19384694]

25. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol. 2009;27:4027–4034. [PubMed: 19597023]

26. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. Clin Trials. 2010;7:567–573. [PubMed: 20392785]

27. Young KY, Laird A, Zhou XH. The efficiency of clinical trial designs for predictive biomarker validation. Clin Trials. 2010;7:557–566. [PubMed: 20571132]

28. Gill RD. Censoring and Stochastic Integrals. Mathematical Centre Tract 124. Amsterdam: Mathematisch Centrum; 1980.

29. Boos DD, Stefanski LA. Essential Statistical Inference: Theory and Methods. Vol 591. New York: Springer; 2013.

30. van der Vaart AW. Asymptotic Statistics. New York: Cambridge University Press; 1998 (Chapter 7).

31. Fleming T, Harrington D. Counting Process and Survival Analysis. Hoboken, NJ: Wiley-Interscience Publication; 1991.

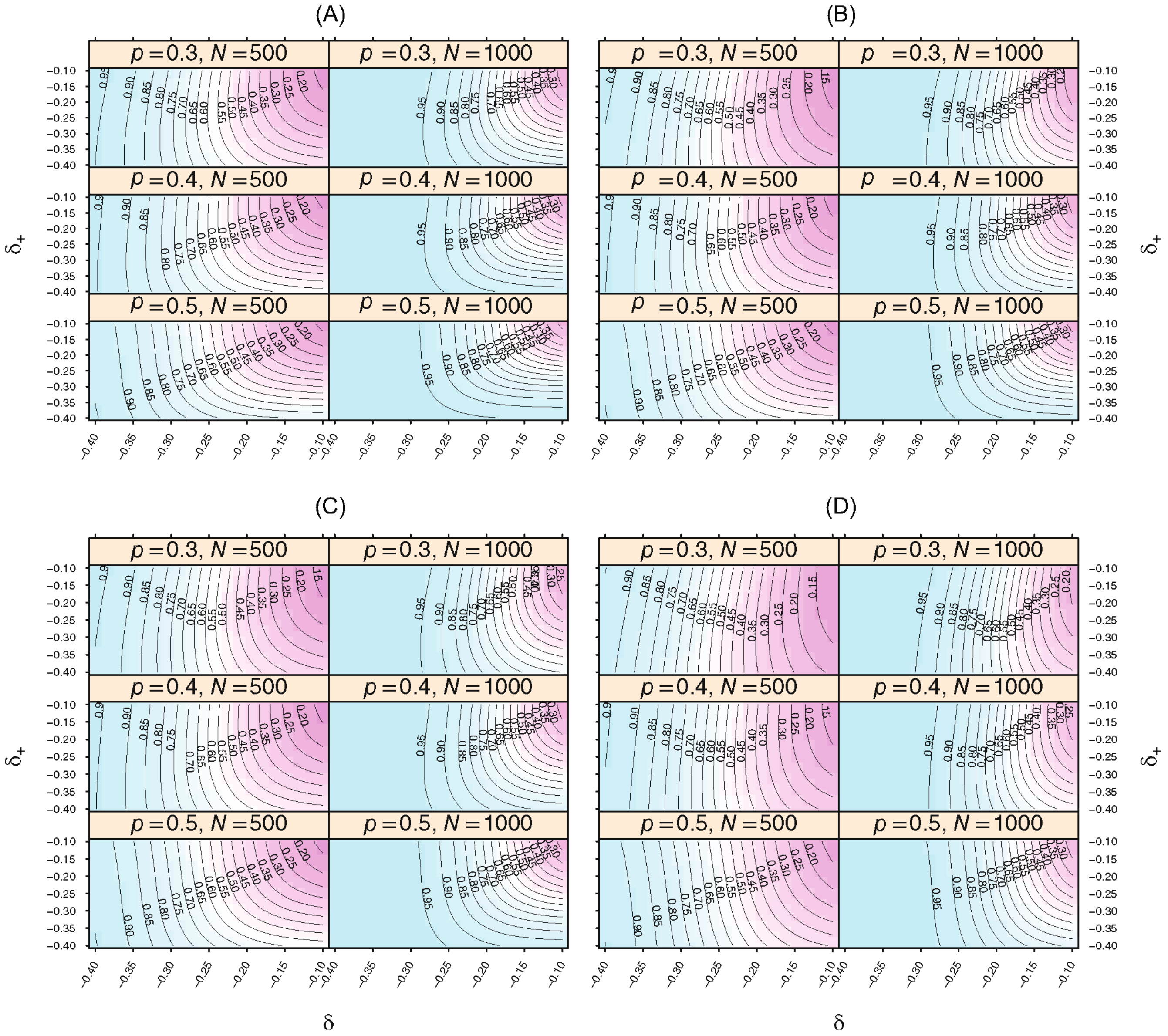


**FIGURE 1.**

Contour plot of global power surface. (A) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (B) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$; (C) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (D) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$.

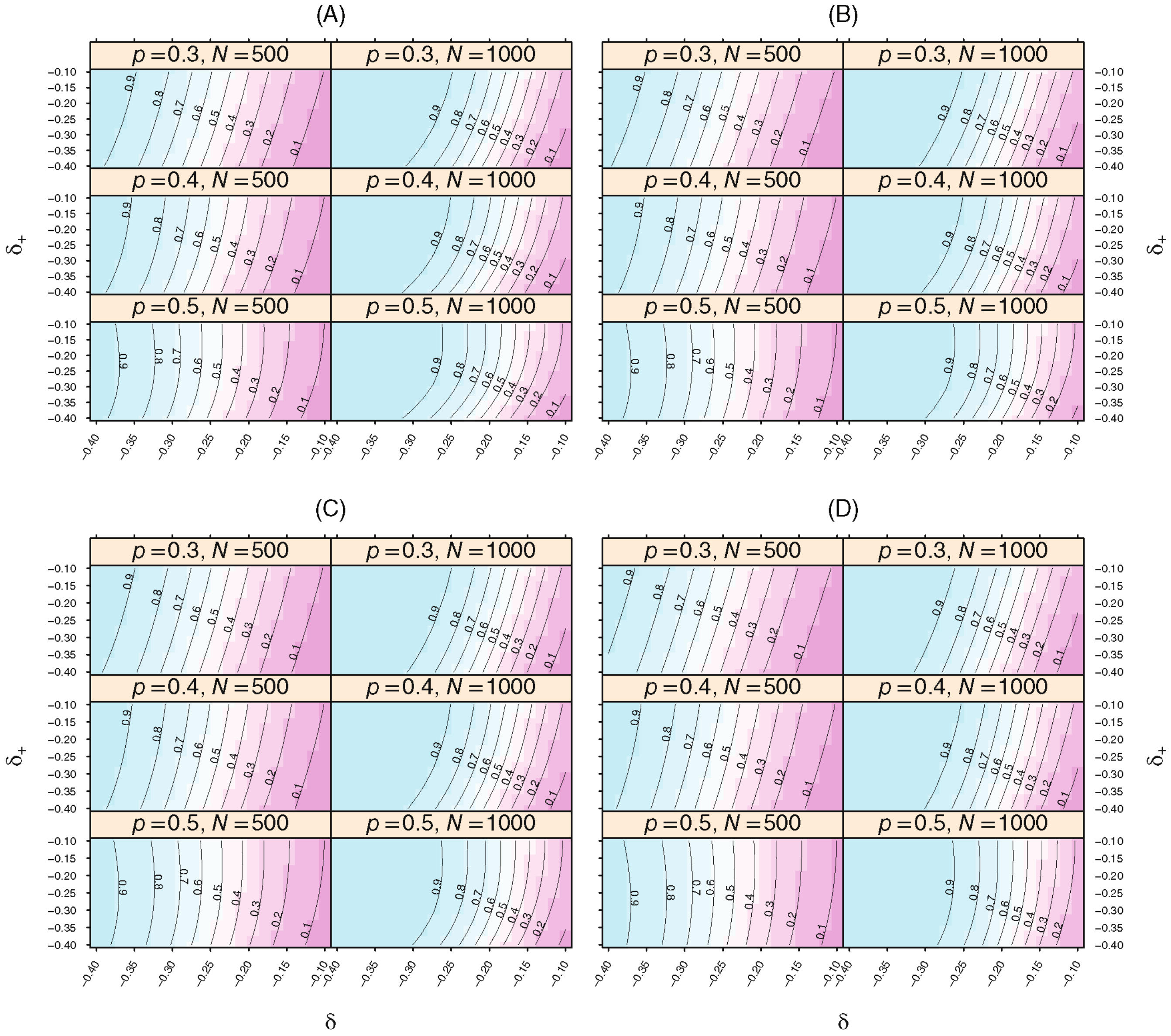


**FIGURE 2.**

Contour plot of power surface for overall population. (A) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (B) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$; (C) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (D) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$.

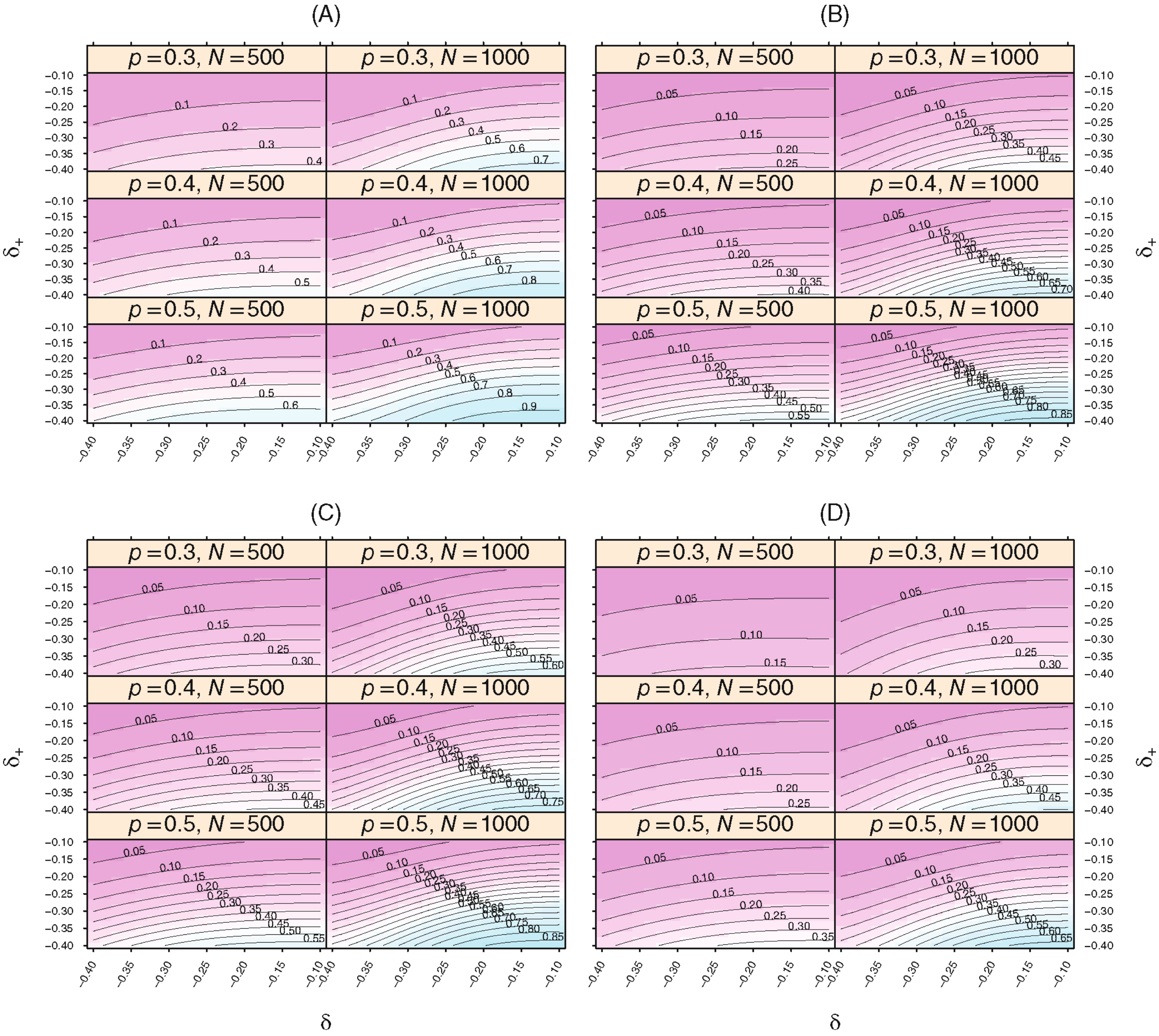


**FIGURE 3.**

Contour plot of power surface for marker positive subgroup. (A) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (B) For $\lambda_{\text{sen}} = 1$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$; (C) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 1$ by $p$ and $N$; (D) For $\lambda_{\text{sen}} = 0.8$ and $\lambda_{\text{spec}} = 0.8$ by $p$ and $N$.

## TABLE 1

Critical values when $\alpha = 0.025$, $\alpha_1 = 0.004$, $\alpha_{1a} = \alpha_1/2$, $\alpha_2 = 0.021$, $\alpha_{2a} = \alpha_2/2$, $r = 0.5$.

| | | | **Information $I_n$ = 0.3** | | | | **Information $I_n$ = 0.5** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ |
| 1.0 | 1.0 | 0.3 | 2.878 | 2.866 | 2.287 | 2.255 | 2.878 | 2.866 | 2.271 | 2.240 |
| | | 0.4 | 2.878 | 2.848 | 2.286 | 2.224 | 2.878 | 2.848 | 2.269 | 2.210 |
| | | 0.5 | 2.878 | 2.816 | 2.284 | 2.178 | 2.878 | 2.816 | 2.266 | 2.164 |
| 0.95 | 0.95 | 0.3 | 2.878 | 2.871 | 2.288 | 2.267 | 2.878 | 2.871 | 2.272 | 2.252 |
| | | 0.4 | 2.878 | 2.857 | 2.286 | 2.238 | 2.878 | 2.857 | 2.270 | 2.224 |
| | | 0.5 | 2.878 | 2.826 | 2.285 | 2.192 | 2.878 | 2.826 | 2.266 | 2.177 |
| 0.9 | 0.9 | 0.3 | 2.878 | 2.875 | 2.288 | 2.276 | 2.878 | 2.875 | 2.273 | 2.261 |
| | | 0.4 | 2.878 | 2.864 | 2.287 | 2.252 | 2.878 | 2.864 | 2.271 | 2.237 |
| | | 0.5 | 2.878 | 2.836 | 2.285 | 2.205 | 2.878 | 2.836 | 2.267 | 2.191 |
| 0.85 | 0.85 | 0.3 | 2.878 | 2.877 | 2.289 | 2.283 | 2.878 | 2.877 | 2.274 | 2.268 |
| | | 0.4 | 2.878 | 2.870 | 2.288 | 2.264 | 2.878 | 2.870 | 2.272 | 2.249 |
| | | 0.5 | 2.878 | 2.845 | 2.286 | 2.219 | 2.878 | 2.845 | 2.268 | 2.205 |
| 0.8 | 0.8 | 0.3 | 2.878 | 2.878 | 2.289 | 2.287 | 2.878 | 2.878 | 2.274 | 2.272 |
| | | 0.4 | 2.878 | 2.874 | 2.288 | 2.274 | 2.878 | 2.874 | 2.273 | 2.259 |
| | | 0.5 | 2.878 | 2.854 | 2.286 | 2.233 | 2.878 | 2.854 | 2.269 | 2.219 |
| 0.75 | 0.75 | 0.3 | 2.878 | 2.878 | 2.289 | 2.289 | 2.878 | 2.878 | 2.274 | 2.274 |
| | | 0.4 | 2.878 | 2.877 | 2.288 | 2.282 | 2.878 | 2.877 | 2.274 | 2.267 |
| | | 0.5 | 2.878 | 2.861 | 2.270 | 2.231 | 2.878 | 2.861 | 2.270 | 2.231 |
| 0.7 | 0.7 | 0.3 | 2.878 | 2.878 | 2.290 | 2.290 | 2.878 | 2.878 | 2.275 | 2.275 |
| | | 0.4 | 2.878 | 2.878 | 2.274 | 2.272 | 2.878 | 2.878 | 2.274 | 2.272 |
| | | 0.5 | 2.878 | 2.867 | 2.287 | 2.258 | 2.878 | 2.867 | 2.271 | 2.243 |
| 1 | 0.8 | 0.3 | 2.878 | 2.876 | 2.288 | 2.279 | 2.878 | 2.876 | 2.273 | 2.264 |
| | | 0.4 | 2.878 | 2.865 | 2.287 | 2.253 | 2.878 | 2.865 | 2.271 | 2.239 |
| | | 0.5 | 2.878 | 2.835 | 2.285 | 2.204 | 2.878 | 2.835 | 2.267 | 2.189 |
| 0.8 | 1 | 0.3 | 2.878 | 2.872 | 2.288 | 2.268 | 2.878 | 2.872 | 2.272 | 2.254 |
| | | 0.4 | 2.878 | 2.860 | 2.287 | 2.245 | 2.878 | 2.860 | 2.270 | 2.231 |
| | | 0.5 | 2.878 | 2.834 | 2.285 | 2.202 | 2.878 | 2.834 | 2.267 | 2.188 |

*Notes:* $c_1$, $c_2$, $b_1$, and $b_2$: critical value for overall population, true marker-positive population at interim analysis, overall population, and true marker-positive population at final analysis time, respectively.