Strategies for validating biomarkers using data from a reference set

LU WANG

Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

YING HUANG*

*Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA and Department of Biostatistics,
University of Washington, Seattle, WA 98195, USA*

yhuang@fredhutch.org

ZIDING FENG

Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

SUMMARY

Candidate biomarkers discovered in the laboratory need to be rigorously validated before advancing to clinical application. However, it is often expensive and time-consuming to collect the high quality specimens needed for validation; moreover, such specimens are often limited in volume. The Early Detection Research Network has developed valuable specimen reference sets that can be used by multiple labs for biomarker validation. To optimize the chance of successful validation, it is critical to efficiently utilize the limited specimens in these reference sets on promising candidate biomarkers. Towards this end, we propose a novel two-stage validation strategy that partitions the samples in the reference set into two groups for sequential validation. The proposed strategy adopts the group sequential testing method to control for the type I error rate and rotates group membership to maximize the usage of available samples. We develop analytical formulas for performance parameters of this strategy in terms of the expected numbers of biomarkers that can be evaluated and the truly useful biomarkers that can be successfully validated, which can provide valuable guidance for future study design. The performance of our proposed strategy for validating biomarkers with respect to the points on the receiver operating characteristic curve are evaluated via extensive simulation studies and compared with the default strategy of validating each biomarker using all samples in the reference set. Different types of early stopping rules and boundary shapes in the group sequential testing method are considered. Compared with the default strategy, our proposed strategy makes more efficient use of the limited resources in the reference set by allowing more candidate biomarkers to be evaluated, giving a better chance of having truly useful biomarkers successfully validated.

Keywords: Biomarker validation; Group sequential testing; Reference set; Rotation; Two-stage.

*To whom correspondence should be addressed.

1. INTRODUCTION

The process of biomarker development has been categorized into five consecutive phases (Pepe and others, 2001). In particular, pre-clinical exploratory studies are conducted in phase 1 to identify potentially useful biomarkers based on tumor and control tissues; retrospective biomarker validation is performed in phase 2 using specimens from individuals with and individuals without disease; phase 3 involves retrospective longitudinal studies to evaluate the capacity of biomarkers for detecting disease; phase 4 involves prospective screening studies for assessing performance; and phase 5 studies aim to estimate mortality reduction due to screening.

Usually, phase 1 studies identify a large number of candidate biomarkers from various laboratories and their performance is validated in later phases. The validation of biomarkers in phase 2 studies is typically based on biobanks of stored specimens. In our motivating application, the National Cancer Institute's Early Detection and Research Network (EDRN) has developed highly valuable reference sets of specimens collected under rigorous standards from multiple institutions for the validation of biomarkers (Feng and others, 2013; Haab and others, 2015). Using such high quality specimens has been recommended as a way to increase the chance of successful validation. Moreover, specimens such as blood or urine samples in a reference set can allow the validation of multiple candidate biomarkers from various laboratories. Since obtaining high quality specimens is an expensive and time-consuming process and these specimens are often limited in volume, it is important to utilize these valuable specimens in a reference set wisely and efficiently, in order to optimize the chance of successful validation.

In this article, we focus on biomarker validation study designs using samples from a reference set, assuming equal volumes of specimens are available from each participant. Motivated by the need to more efficiently utilize the specimens, instead of using the common strategy of measuring each biomarker using specimens from all participants, here, we consider a two-stage validation process: for validation of each biomarker, participants in the reference set are partitioned into two groups. Each biomarker is first evaluated using group 1 samples; only those biomarkers satisfying a predefined performance criterion are then tested using group 2 samples. If the split of participants into group 1 and group 2 is the same across biomarkers, samples from group 1 will be quickly exhausted, such that only the group 2 samples will be available for evaluating biomarkers that come in later. To deal with this problem, we also propose a strategy that rotates group membership in order to maximize the usage of all available samples.

To control the error rate in the two-stage validation procedure, we adopt the group sequential testing strategy (Emerson and Fleming, 1989; O'Brien and Fleming, 1979). Group sequential designs allow early termination when the candidate biomarker is evidently superior or inferior, so that the remaining specimens can be conserved for validation of other candidate biomarkers. Previously in the field of biomarker evaluation, Pepe and others (2009) considered conditional estimators of sensitivity and specificity for a dichotomous diagnostic biomarker under a two-stage study that allows early termination for futility. This method was extended in Koopmeiners and others (2012) for the evaluation of continuous biomarkers.

In our research, for validating multiple biomarkers using the limited resources in the reference set, we consider two important criteria to characterize the performance of a validation design. The first is the expected number of biomarkers that can be studied using the available specimens from the reference set; the second is the expected number of truly useful biomarkers that can be successfully validated. We develop analytical formulas for both criteria, which can provide useful guidance for validation study designs in practice.

This article is organized as follows: In Section 2, we propose a new strategy for validating biomarkers in reference sets that combines a two-stage validation process with rotation of participant sets. We also develop analytical formulas for important performance criteria in the proposed validation strategy. In Section 3, we conduct extensive simulation studies to evaluate the performance of our proposed strategy under different types of early stopping rules and stopping boundaries of the group sequential testing, and

compare the performance with the alternative one-stage strategy of testing each biomarker in all samples. We also demonstrate the use of analytical formulas for the design of validation studies. In Section 4, the performance of the proposed strategy is further demonstrated using a numerical setting constructed based on the data from the EDRN pancreatic cancer reference set study. Finally, we end the article with concluding remarks in Section 5.

2. METHODOLOGY

2.1. Group sequential tests of biomarker performance

Suppose we have a reference set that consists of specimens of equal volumes from N participants. The specimen volume (e.g., volume of blood or urine) from each individual is typically determined by the study budget and by practical considerations. Our goal is to have multiple laboratories assess their candidate biomarkers using specimens from this reference set and to validate the classification performance of individual biomarkers.

For the validation of each individual biomarker, we consider a hypothesis test for its classification accuracy to measure whether θ passes a predefined performance threshold. That is, for each biomarker we want to use a one-sided test for the null hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_A : \theta > \theta_0$, with type I error controlled at a level α , where θ_0 is a pre-specified threshold for classification performance. In this article, we consider θ to be the points on the receiver operating characteristic (ROC) curve, although the methods apply in general to other classification measures such as the area under the ROC curve (AUC) or partial AUC (Pepe, 2003; Zhou and others, 2009).

In a study utilizing a reference set to validate biomarkers, all labs with biomarkers of interest can request specimens from the reference set to analyze their biomarkers. Statisticians from the Data Management and Coordinating Center (DMCC) of EDRN are responsible for distributing the appropriate specimens to each lab and later analyzing the biomarker data measured from those specimens to validate their classification performance. For validating individual biomarkers using the reference set, the default strategy is to provide the needed volumes of specimens from all participants in the reference set to each individual lab, and repeat this process until all specimens have been consumed. While this strategy is simple and easy to implement, it almost certainly does not make the most efficient use of the reference set samples. For instance, a number of factors during the biomarker discovery process could lead to over-estimation of a biomarker's performance, meaning that resources could be inefficiently directed towards validating candidate biomarkers with inferior performance. If, however, candidate biomarkers with inferior performance can be identified earlier during the validation process using a small set of samples, the rest of the samples can be spared and used to validate other biomarkers with greater potential. On the other hand, biomarkers with superior performance could potentially also be validated using a smaller number of samples.

We propose a two-stage validation process with rotation of participants into each stage, geared specifically for the reference set study. First, for evaluation of each individual marker, we break the individuals in the reference set into two groups, group 1 and group 2. The biomarker is measured in samples from group 1 to test its classification performance; only the biomarkers that satisfy some predefined performance criteria in group 1 are subsequently measured in group 2 samples, such that their performance is ultimately tested on the whole sample set. To control for type I error rate in this two-stage testing, we adopt the sequential testing strategy used in group sequential designs (Jennison and Turnbull, 1999).

Recall that in a typical group sequential design in which test statistics Z_k are observed at analyses $k = 1, \dots, K$, the one-sided group sequential test that controls the overall type I error at a significance level α has the following form:

At interim analysis $k = 1, \dots, K - 1$,

If $Z_k \geq b_k$, stop, reject H_0

If $Z_k \leq a_k$, stop, accept H_0
 Otherwise, continue to group $k + 1$
 At final analysis K ,
 If $Z_K \geq b_K$, stop, reject H_0
 If $Z_K < a_K$, stop, accept H_0 ,

where constants $a_k < b_k$ for $k = 1, \dots, K - 1$ correspond to the threshold of early termination for futility and efficacy, respectively; $a_K = b_K$ ensures that the test terminates at the last analysis; and $b_k = \infty$ or $a_k = -\infty$ for $k = 1, \dots, K - 1$ corresponds to early termination for futility or efficacy only, respectively. For the design allowing early stopping for both futility and efficacy, we adopt the one-sided symmetric sequential test designs proposed in [Emerson and Fleming \(1989\)](#), which treat the null and alternative hypotheses symmetrically with respect to early termination, i.e., the interchange of null and alternative hypotheses would lead to identical boundaries for stopping. Let r_k be the proportion of sample size enrolled at analysis k relative to the total sample size, for $k = 1, \dots, K$. The boundaries for one-sided symmetric tests are defined as

$$\begin{aligned} a_k &< b_k, \text{ for } k = 1, \dots, K - 1, \\ a_k &= 2r_k b_K - b_k, \text{ and } b_k = (Kr_k)^P c_{r,P}^{(\alpha)}, \text{ for } k = 1, \dots, K, \end{aligned} \quad (2.1)$$

where $c_{r,P}^{(\alpha)}$ is a critical value such that setting $b_k = (Kr_k)^P c_{r,P}^{(\alpha)}$ will result in a test with size α . The parameter P corresponds to the boundary shape; lower values of P lead to more conservative testing at earlier analyses. We consider two different types of boundary: $P = 0$ or 0.5 , corresponding to the O'Brien–Fleming boundary ([O'Brien and Fleming, 1979](#)) and the Pocock boundary ([Pocock, 1977](#)), respectively. Note that while the boundary given by (2.1) defines the test in terms of the partial sum statistic, we can always transform the boundary on a different scale according to a different choice of test statistic.

In this article, we adopt the sequential design with $K = 2$. We consider the construction of a standardized test statistic with respect to the points on the ROC curve. Let $\widehat{ROC}_k(t)$ be the sequential empirical estimator of $ROC(t)$ at analysis $k = 1, 2$ based on all observed data at that analysis point. According to Corollary 3.3 from [Koopmeiners and Feng \(2011\)](#), $(\widehat{ROC}_1(t), \widehat{ROC}_2(t))$ is approximately multivariate normal with

$$\widehat{ROC}_k(t) \sim N\left(ROC(t), \sigma_{\widehat{ROC}(t),k}^2\right), \text{ for } k = 1, 2$$

and

$$\text{cov}\left(\widehat{ROC}_1(t), \widehat{ROC}_2(t)\right) = \sigma_{\widehat{ROC}(t),2}^2,$$

where

$$\sigma_{\widehat{ROC}(t),k}^2 = \frac{ROC(t)(1 - ROC(t))}{N_{D,k}} + \left[\frac{f_D(S_D^{-1}(t))}{f_{\bar{D}}(S_D^{-1}(t))} \right]^2 \frac{t(1-t)}{N_{\bar{D},k}}, \quad (2.2)$$

in which $N_{D,k}$ and $N_{\bar{D},k}$ denote the sample sizes for cases and controls at analysis k , f_D and $f_{\bar{D}}$ are the probability density functions of the biomarker among cases and controls, respectively, and S_D is the

survival function of the biomarker among controls. To perform a one-sided α -level hypothesis test $H_0 : ROC(t) \leq ROC_0(t)$ against $H_A : ROC(t) > ROC_0(t)$, we use the following standardized test statistics:

$$Z_k = \frac{\widehat{ROC}_k(t) - ROC_0(t)}{\sigma_{\widehat{ROC}(t),k}}, \text{ for } k = 1, 2. \quad (2.3)$$

It is immediate from Corollary 3.3 in [Koopmeiners and Feng \(2011\)](#) that Z_k follows a standard normal distribution under the null hypothesis. Under the alternative hypothesis, i.e., $ROC(t) = ROC_A(t) > ROC_0(t)$ holds, the test statistic Z_k is normally distributed with mean μ_k and variance 1, where $\mu_k = [ROC_A(t) - ROC_0(t)] / \sigma_{\widehat{ROC}(t),k}$. Usually the variance $\sigma_{\widehat{ROC}(t),k}^2$ is unknown. In practice, the unknown components in the expression of $\sigma_{\widehat{ROC}(t),k}^2$ need to be estimated. The components $ROC(t)$ and $S_D^{-1}(t)$ in (2.2) can be substituted by their corresponding sequential empirical estimators. As for the ratio of density function $f_D/f_{\bar{D}}$, by Bayes' rule, we have

$$f_D(x) = \frac{P(D=1|x)f(x)}{P(D=1)} \text{ and } f_{\bar{D}}(x) = \frac{P(D=0|x)f(x)}{P(D=0)},$$

where $f(x)$ is the density function of biomarker. Therefore,

$$\frac{f_D(x)}{f_{\bar{D}}(x)} = \frac{1 - P(D=1)}{P(D=1)} \frac{P(D=1|x)}{1 - P(D=1|x)}.$$

Note that $P(D=1|x)$ can be estimated by fitting a logistic regression model of disease status D on biomarker X . Then, the estimate for the ratio of density functions can be obtained.

2.2. Rotation of group membership

Consider the two-stage testing strategy described in Section 2.1 and let $r \in (0, 1)$ denote the proportion of samples in group 1 relative to the total sample size in the reference set. If $Z_1 \geq b_1$ or $Z_1 \leq a_1$, then the validation of a biomarker stops at the first stage and only $100 \times r\%$ samples are used for measuring this marker. Thus, the remaining $100 \times (1 - r)\%$ samples are spared and can be used to validate other biomarkers. If instead we have $a_1 < Z_1 < b_1$, the biomarker passes the first stage of testing and samples from all individuals in the reference set would be used for validating this biomarker. Note that if the split of individuals into group 1 and group 2 is the same across biomarkers, then specimens from the individuals in group 1 will be quickly exhausted and only specimens from individuals in group 2 will be left for evaluation of biomarkers that are evaluated later in the queue. We thus propose a procedure that rotates group membership.

We consider a design where the total number of participants in the reference set is Q -fold of the number included in the first stage, for some positive number Q . That is, the biomarker is first assessed in specimens from $r = 1/Q$ participants. To begin with, for the first biomarker entering the validation process, we randomly partition the participants in the reference set into Q blocks. One block is randomly selected to be group 1 and the rest of the blocks are treated as group 2. A one-sided group sequential test with $K = 2$ is then performed. Next, each time a new biomarker enters the validation process, we first check how many specimen volumes remain within each block. Among the blocks with the maximum remaining amount of specimens, we randomly draw one block as group 1 and then perform this sequential test again. This procedure is repeated for evaluating incoming biomarkers until the specimens have been fully consumed in at least one of the blocks. Note that at this point it is possible that specimens could remain in other blocks. One could potentially use those remaining available specimens to validate another biomarker

with a fix-sample test. We explored adding this extra step in our numerical studies. The difference with respect to the performance criteria is negligible when the number of biomarkers that can be evaluated in a reference set is not small. Therefore, we present results without this extra step.

2.3. Theoretical results

In this section, we develop theoretical results for several performance criteria of the proposed validation design. These results will be useful in guiding the choice of validation design in practice.

2.3.1. Performance criteria under the proposed design. Without loss of generality, here, we consider a simple setting where measurement of each biomarker requires equal specimen volume (denoted as one unit). We also set the total volume of specimens from each participant in the reference set as M units. The validation strategy described in Section 2.2, however, is applicable to general settings where biomarkers may require different specimen volumes for measurement.

We consider the following two performance criteria, which are important for biomarker validation using a reference set. The first is the expected number of biomarkers that can be evaluated using the available specimens from the reference set. The second is the expected number of truly useful biomarkers that can be successfully validated. The analytical formulas of the two criteria for the proposed design are presented in Results 1 and 3, respectively.

RESULT 1 Let m be the number of biomarkers evaluated using the proposed two-stage validation design. For any integer $M \geq 1$ and $Q \geq 2$, the expected number of biomarkers evaluated equals

$$E(m) = M + \sum_{R=0}^M (Q-1)R \binom{M+(Q-1)R}{QR} p^{QR} (1-p)^{M-R} \\ + \sum_{R=0}^{M-1} \sum_{T=0}^{Q-2} [(Q-1)R+T] \binom{M+(Q-1)R+T}{QR+T+1} p^{QR+T+1} (1-p)^{M-R}$$

where $p = P(Z_1 \geq b_1) + P(Z_1 \leq a_1)$.

Proof. For a group sequential design with $K = 2$, the proportion of samples used for evaluating one biomarker equals

$$q = \begin{cases} 1/Q & \text{with probability } p = P(Z_1 \geq b_1) + P(Z_1 \leq a_1) \\ 1 & \text{with probability } 1-p = P(a_1 < Z_1 < b_1) \end{cases}.$$

Obviously, $m \in [M, QM]$, and we further notice that

$$P(m = M + (Q-1)R) = \binom{M+(Q-1)R}{QR} p^{QR} (1-p)^{M-R} + \binom{M+(Q-1)R}{QR+1} p^{QR+1} (1-p)^{M-R}, \\ P(m = M + (Q-1)R + T) = \binom{M+(Q-1)R+T}{QR+T+1} p^{QR+T+1} (1-p)^{M-R},$$

for $R = 0, 1, \dots, M-1$, $T = 1, 2, \dots, Q-1$ and $P(m = QM) = p^{QM}$. Consequently, the expected value of m equals

$$\begin{aligned}
 E(m) &= \sum_{R=0}^{M-1} \{[M + (Q-1)R] P(m = M + (Q-1)R)\} \\
 &\quad + \sum_{R=0}^{M-1} \sum_{T=1}^{Q-1} \{[M + (Q-1)R + T] P(m = M + (Q-1)R + T)\} + QM \times P(m = QM) \\
 &= M + \sum_{R=0}^M (Q-1)R \binom{M + (Q-1)R}{QR} p^{QR} (1-p)^{M-R} \\
 &\quad + \sum_{R=0}^{M-1} \sum_{T=0}^{Q-2} [(Q-1)R + T] \binom{M + (Q-1)R + T}{QR + T + 1} p^{QR+T+1} (1-p)^{M-R}. \quad \square
 \end{aligned}$$

Result 1 provides a general analytical formula for the expected number of biomarkers evaluated using our proposed two-stage validation design. In our numerical studies, we adopt the validation design with $Q = 2$ or 3 , i.e., $r = 1/2$ or $1/3$.

The second performance criterion, we are interested in is the expected number of truly useful biomarkers that are successfully validated. Before presenting the corresponding analytical formula, we first present the following result regarding the expected number of tested biomarkers found to be useful using the proposed design.

RESULT 2 Let n_m^* denote the number of biomarkers evaluated with the null hypothesis rejected, the expected number of biomarkers declared to have adequate performance equals

$$E(n_m^*) = p_{rej} E(m),$$

where $p_{rej} = [P(Z_1 \geq b_1) + P(a_1 < Z_1 < b_1, Z_2 \geq b_2)]$ is the rejection rate, and $E(m)$ is the expected number of biomarkers evaluated which is given in Result 1.

Proof. We observe that $n_m^*|m \sim \text{Binomial}(m, p_{rej})$, thus one has $E(n_m^*) = E(E(n_m^*|m)) = E(mp_{rej}) = p_{rej} E(m)$. $\square$

To study the second performance criterion of the proposal design, we make the assumption that the true performance of the biomarkers to be validated in the reference set follows a mixture distribution, i.e., $\theta = \theta_0^* \leq \theta_0$ with probability λ and $\theta = \theta_A > \theta_0$ with probability $1 - \lambda$, where $\lambda \in [0, 1]$. Notice that when $\lambda = 0$, all n_m^* biomarkers testing positive using the reference set samples are truly useful. When $\lambda > 0$, some biomarkers with null performance can incorrectly test as positive. We are interested in finding out how many truly useful biomarkers are successfully validated as positive. Let n_m denote the number of truly useful biomarkers that are successfully validated. We then have the following result about its expected value.

RESULT 3 Let n_m denote the number of truly useful biomarkers that are successfully validated, the expected number of truly useful biomarkers that are successfully validated equals

$$E(n_m) = p_{rej}^{H_A} (1 - \lambda) E(m),$$

where $p_{rej}^{H_A}$ is the value of $[P(Z_1 > b_1) + P(a_1 < Z_1 < b_1, Z_2 \geq b_2)]$ under the alternative hypothesis, $\lambda = P(\theta \leq \theta_0)$, $E(m)$ is the expected number of biomarkers evaluated which is given in Result 1.

Proof. Let m^* denote the number of truly useful biomarkers evaluated among the m evaluated biomarkers. Since $n_m|m^* \sim \text{Binomial}(m^*, p_{rej}^{H_A})$, we have $E(n_m) = E(E(n_m|m^*)) = E(m^* p_{rej}^{H_A}) = p_{rej}^{H_A} E(m^*) = p_{rej}^{H_A} E(E(m^*|m)) = p_{rej}^{H_A} E((1 - \lambda)m) = p_{rej}^{H_A} (1 - \lambda) E(m)$. $\square$

Suppose the standardized test statistic used in the sequential testing method follows the standard normal distribution under the null hypothesis and is normally distributed with mean μ_k and variance σ_k^2 at analysis k . Under the assumption that the true performance of the biomarkers to be validated in the reference set follows a mixture distribution, the probability $p = P(Z_1 \geq b_1) + P(Z_1 \leq a_1)$ defined in Result 1 has the form

$$p = 1 - \lambda \Phi(b_1) - (1 - \lambda) \Phi\left(\frac{b_1 - \mu_1}{\sigma_1}\right) + \lambda \Phi(a_1) + (1 - \lambda) \Phi\left(\frac{a_1 - \mu_1}{\sigma_1}\right),$$

where Φ denotes the cumulative distribution function of a standard normal distribution. The corresponding value of $E(m)$ can then be computed based on Result 1. Furthermore, the rejection rate p_{rej} defined in Result 2 can be written as

$$p_{rej}(\lambda) = 1 - \lambda \Phi(b_1) - (1 - \lambda) \Phi\left(\frac{b_1 - \mu_1}{\sigma_1}\right) + \lambda \int_{b_2}^{+\infty} \int_{a_1}^{b_1} f_{\theta_0}(Z_1, Z_2) dZ_1 dZ_2 + (1 - \lambda) \int_{b_2}^{+\infty} \int_{a_1}^{b_1} f_{\theta_A}(Z_1, Z_2) dZ_1 dZ_2,$$

where $f_{\theta}(Z_1, Z_2) = \left(2\pi\sigma_1\sigma_2\sqrt{1 - \rho_{\theta}^2}\right)^{-1} \exp\left\{-\frac{1}{2(1 - \rho_{\theta}^2)}\left[\frac{(Z_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho_{\theta}(Z_1 - \mu_1)(Z_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(Z_2 - \mu_2)^2}{\sigma_2^2}\right]\right\}$ is the joint probability density function of (Z_1, Z_2) , in which $\rho_{\theta} = \text{cor}(Z_1, Z_2|\theta)$, $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$ when the null hypothesis holds. In addition, $p_{rej}^{H_A}$ defined in Result 3 equals $p_{rej}(0)$. We can then compute $E(n_m)$ and $E(n_m^*)$ based on Results 2 and 3.

3. SIMULATION STUDY

In this section, we conduct simulation studies to evaluate the performance of the proposed validation design as described in Section 2 for validating candidate biomarkers using samples from a reference set.

We consider a reference set that includes specimens with $M = 50$ volume units from each participant. We further assume that there are many candidate biomarkers waiting to be evaluated using the specimens in the reference set, and that the validation of a single biomarker requires one unit volume of specimen. Let D be a binary disease outcome. Biomarker X for the controls is generated from a standard normal distribution, and for the cases X follows a normal distribution with mean μ_D and variance 1. We consider case/control samples with 1:1 ratio and total sample size $N = 120, 240, 540$, or 1200. Suppose the performance of interest is $ROC(0.1)$. For the validation of each individual biomarker, we conduct a one-sided test of the null hypothesis $H_0 : ROC(0.1) \leq ROC_0(0.1)$ with type I error probability controlled at $\alpha = 0.05$, where the threshold for the null hypothesis is set at $ROC_0(0.1) = 0.25$. Suppose the biomarkers to be validated comprise a mixture of null biomarkers and useful biomarkers, such that the true $ROC(0.1)$ follows a mixture distribution: $ROC(0.1) = 0.25$ with probability λ , and $ROC(0.1) = 0.35$ with probability $1 - \lambda$. In our simulations, we achieve this by simulating $\lambda \times 100\%$ of biomarkers with $\mu_D = 0.6$ such that

those biomarkers have null performance with $ROC(0.1) = 0.25$, and for the other $(1 - \lambda) \times 100\%$ of biomarkers, we set $\mu_D = 0.9$ such that $ROC(0.1) = 0.35$. A series of λ varying within $\{0, 0.2, 0.5, 0.8\}$ is considered in the simulation studies.

For validating the performance of individual biomarkers, we apply our proposed two-stage testing strategy with rotation of group membership. To complete the test procedure, we use the standardized test statistic defined in (2.3). We consider the design for which a biomarker is first validated using specimens from half or one-third of the participants, i.e., $r = 1/2$ or $r = 1/3$. With respect to the group sequential designs, two different types of stopping rule and boundary shape are considered. Specifically, the designs allowing early stopping for both futility and efficacy, i.e., $a_1 \neq -\infty$ and $b_1 \neq \infty$, or allowing early stopping for futility only, i.e., $a_1 \neq -\infty$ and $b_1 = \infty$, are considered. We hereafter refer to these two types of stopping rules as Stop = “Both” and Stop = “Null”, respectively, where the one-sided symmetric boundaries proposed in Emerson and Fleming (1989) are used for the design with Stop = “Both”. Also, two different types of boundary shape, the O’Brien–Fleming boundary ($P = 0$) and the Pocock boundary ($P = 0.5$), are considered. In our simulations, the boundaries are calculated using the RCTdesign package in R (<http://www.rctdesign.org/Welcome.html>). For comparison purposes, we also consider the default design ($K = 1$) in which the test statistic Z_2 is compared with a standard normal distribution. To evaluate the performance of different designs, we examine the number of biomarkers evaluated, m , and the number of truly useful biomarkers that are successfully validated, n_m . Average values of m and n_m based on 5000 Monte Carlo simulations are calculated, denoted by $E(m)$ and $E(n_m)$, respectively.

3.1. Results comparing the proposed design with the default design

Table 1 compares the simulation results for testing $ROC(0.1)$ using the proposed validation strategy with $r = 1/2$ to those obtained using the default strategy. The results clearly show that, compared with the default strategy, our proposed validation strategy can evaluate more candidate biomarkers. Moreover, our proposed validation strategy can correctly validate more biomarkers that are truly useful. Among the different types of stopping rules and boundary shapes used for the proposed strategy, the design that allows early stopping under both null and alternative hypotheses with the Pocock boundary (Stop = “Both”, $P = 0.5$) generally has the best performance in terms of largest $E(m)$ and $E(n_m)$. The design that allows early stopping under the null hypothesis alone using the O’Brien–Fleming boundary (Stop = “Null”, $P = 0$) has smaller $E(m)$ and $E(n_m)$, but still yields improved results compared with the default strategy. Among both designs with Stop = “Both” and designs with Stop = “Null”, using the Pocock boundary ($P = 0.5$) leads to better performance compared with the O’Brien–Fleming boundary, in terms of both $E(m)$ and $E(n_m)$. This is due to the fact that using the O’Brien–Fleming boundary leads to more conservative testing in the first stage compared with the Pocock boundary. Thus, more candidate biomarkers would enter the second stage of testing, fewer biomarkers would be evaluated, and consequently fewer truly useful biomarkers would be successfully validated. Table 1 shows that for a moderate sample size of $N = 540$, compared with the default design, the relative increases of $E(m)$ and $E(n_m)$ using the proposed design range from 42.4% to 61.7% and from 25.2% to 44.9%, respectively, with Stop = “Both” and $P = 0.5$, and from 2.4% to 25.0% and from 1.2% to 27.3%, respectively, with Stop = “Null” and $P = 0$, for λ varying from 0 to 0.8. In reality, the candidate biomarkers are not truly useful in most cases. Therefore, λ is typically large, with its value close to 1.

In addition, when examining the absolute change of $E(m)$, we notice that for the proposed design with Stop = “Both”, $E(m)$ tends to increase as sample size increases. It also increases as the value of λ increases for small or moderate sample size, while the magnitude of its change as λ varies is small for large sample size. As for the proposed design with Stop = “Null”, $E(m)$ decreases slightly as sample size increases when $\lambda \leq 0.5$, and increases as λ increases. This is because early termination is more likely when there are more biomarkers with null performance. Although the pattern of the change in $E(m)$ as sample size

Table 1. Comparison of the performance of the default design ($K = 1$) and the proposed two-stage validation design with rotation of group membership ($K = 2, r = 1/2$) using different types of stopping rules (Stop = “Both” or “Null”) and boundary shapes ($P = 0$ or 0.5). The table shows the average number of biomarkers evaluated ($E(m)$), the average number of truly useful biomarkers that are successfully validated ($E(n_m)$), and the corresponding standard error of m and n_m (in parentheses) for testing $\text{ROC}(0.1) \leq 0.25$. Simulations were performed using case/control samples with a 1:1 ratio and total sample size $N = 120, 240, 540$, or 1200 . λ is the proportion of null markers in the population

λ	N	Performance	$K = 2, r = \frac{1}{2}$				$K = 1$
			Stop = “Both”		Stop = “Null”		
			$P = 0$	$P = 0.5$	$P = 0$	$P = 0.5$	
0	120	$E(m)$	57.84 (0.03)	73.06 (0.04)	53.94 (0.02)	59.42 (0.03)	50
		$E(n_m)$	20.99 (0.05)	25.01 (0.06)	18.25 (0.05)	19.85 (0.05)	17.16 (0.05)
	240	$E(m)$	56.61 (0.03)	71.58 (0.04)	52.19 (0.02)	57.15 (0.03)	50
		$E(n_m)$	27.46 (0.06)	31.56 (0.06)	24.67 (0.05)	26.36 (0.05)	23.79 (0.05)
	540	$E(m)$	57.50 (0.03)	71.20 (0.04)	51.20 (0.01)	54.14 (0.02)	50
		$E(n_m)$	41.32 (0.05)	45.53 (0.06)	36.82 (0.04)	37.58 (0.04)	36.37 (0.04)
0.2	1200	$E(m)$	66.93 (0.04)	80.02 (0.04)	50.04 (0.003)	50.59 (0.01)	50
		$E(n_m)$	63.44 (0.05)	73.04 (0.06)	47.49 (0.02)	47.40 (0.02)	47.52 (0.02)
	120	$E(m)$	58.86 (0.03)	74.40 (0.04)	55.41 (0.02)	61.47 (0.03)	50
		$E(n_m)$	17.28 (0.05)	20.55 (0.05)	15.22 (0.04)	16.66 (0.04)	13.85 (0.04)
	240	$E(m)$	57.78 (0.03)	73.24 (0.04)	53.85 (0.02)	59.71 (0.03)	50
		$E(n_m)$	22.62 (0.05)	26.11 (0.06)	20.51 (0.05)	22.13 (0.05)	18.99 (0.05)
0.5	540	$E(m)$	59.11 (0.03)	73.40 (0.04)	53.57 (0.02)	57.48 (0.02)	50
		$E(n_m)$	34.20 (0.05)	37.72 (0.06)	31.15 (0.04)	32.16 (0.04)	29.11 (0.04)
	1200	$E(m)$	66.67 (0.04)	80.54 (0.04)	52.08 (0.01)	54.12 (0.02)	50
		$E(n_m)$	50.93 (0.04)	59.15 (0.05)	39.94 (0.02)	40.96 (0.02)	38.02 (0.02)
	120	$E(m)$	60.61 (0.03)	76.78 (0.04)	57.91 (0.03)	65.04 (0.04)	50
		$E(n_m)$	11.12 (0.04)	13.24 (0.04)	9.89 (0.04)	11.00 (0.04)	8.58 (0.03)
0.8	240	$E(m)$	59.67 (0.03)	75.96 (0.04)	56.76 (0.02)	64.18 (0.03)	50
		$E(n_m)$	14.58 (0.04)	16.89 (0.05)	13.51 (0.04)	14.91 (0.04)	11.87 (0.04)
	540	$E(m)$	61.79 (0.03)	76.97 (0.04)	57.72 (0.02)	63.67 (0.03)	50
		$E(n_m)$	22.42 (0.04)	24.73 (0.04)	20.97 (0.03)	22.24 (0.03)	18.22 (0.03)
	1200	$E(m)$	66.40 (0.04)	81.35 (0.04)	55.97 (0.02)	60.70 (0.02)	50
		$E(n_m)$	31.70 (0.03)	37.35 (0.03)	26.77 (0.02)	28.62 (0.02)	23.76 (0.02)
0.8	120	$E(m)$	62.28 (0.03)	79.09 (0.04)	60.47 (0.03)	69.04 (0.04)	50
		$E(n_m)$	4.65 (0.02)	5.53 (0.03)	4.21 (0.02)	4.71 (0.02)	3.43 (0.02)
	240	$E(m)$	61.60 (0.03)	78.97 (0.04)	59.94 (0.03)	69.53 (0.04)	50
		$E(n_m)$	6.18 (0.03)	7.14 (0.03)	5.84 (0.02)	6.58 (0.03)	4.77 (0.02)
	540	$E(m)$	64.62 (0.04)	80.83 (0.04)	62.51 (0.03)	71.11 (0.04)	50
		$E(n_m)$	9.57 (0.02)	10.55 (0.03)	9.27 (0.02)	10.15 (0.02)	7.28 (0.02)
0.8	1200	$E(m)$	66.02 (0.04)	82.10 (0.04)	60.39 (0.03)	68.92 (0.03)	50
		$E(n_m)$	12.92 (0.01)	15.36 (0.02)	11.83 (0.01)	13.28 (0.01)	9.52 (0.01)

or λ varies between designs with different types of stopping rules, the pattern of the change in $E(n_m)$ is consistent across different types of designs. We observe that $E(n_m)$ increases as sample size increases, and decreases as the value of λ increases. For example, for the proposed design with Stop = “Both” and $P = 0.5$, when $\lambda = 0.5$, $E(n_m)$ increases from 13.24 to 37.35 as the sample size increases from 120 to

1200 and $E(n_m)$ decreases from 73.04 to 15.36 as λ increases from 0 to 0.8 for sample size $N = 1200$. This finding is expected, since a larger sample size provides greater power, while a larger value of λ indicates fewer truly useful biomarkers to be validated. In fact, it is more relevant to see if $E(n_m)$ increases relative to the default study design. Let $E(n_0)$ indicate the average number of truly useful biomarkers validated in the default design. When compared with the default design, we observed that for the proposed design with Stop = “Both”, the increase in truly useful biomarkers identified (i.e., $E(n_m) - E(n_0)$) decreases as λ increases; for the proposed design with Stop = “Null”, $E(n_m) - E(n_0)$ increases as λ increases from 0 to 0.5 and then falls back when $\lambda = 0.8$. The relative improvement in truly useful biomarkers identified, i.e., $[E(n_m) - E(n_0)]/E(n_0)$, clearly increases as λ increases. For example, for the proposed design with Stop = “Both” and $P = 0.5$, for sample size $N = 540$, $[E(n_m) - E(n_0)]/E(n_0)$ increases from 25.2% to 44.9% when λ increases from 0 to 0.8. To summarize, our proposed design can have appreciable practical advantage compared with the default design, especially when the proportion of null markers is large in the population (λ is large). For example, when $\lambda = 0.8$ and $N = 1200$, the relative improvements of our design compared with the default design with respect to $[E(m) - E(m_0)]/E(m_0)$ and $[E(n_m) - E(n_0)]/E(n_0)$ are 64.2% and 61.3%, respectively, using the Pocock boundary with Stop = “Both”; corresponding relative improvements are 37.8% and 39.4% with Stop = “Null”.

For the proposed strategy with $r = 1/3$, simulation results are presented in Table 1 in Appendix A of the [supplementary material](#) available at *Biostatistics* online. As in the case of $r = 1/2$, in both designs with Stop = “Both” and Stop = “Null”, using the Pocock boundary ($P = 0.5$) has better performance with respect to $E(m)$ and $E(n_m)$ compared with using the O’Brien–Fleming boundary ($P = 0$). The performance of the proposed design is similar between $r = 1/2$ and $r = 1/3$ with the Pocock boundary and better in $r = 1/2$ compared with $r = 1/3$ with the O’Brien–Fleming boundary. The pattern of the change in $E(m)$, $E(n_m)$, $E(n_m) - E(n_0)$, and $[E(n_m) - E(n_0)]/E(n_0)$ as λ and sample size vary is similar between the designs using $r = 1/2$ and $r = 1/3$. In practice, if a more conservative test (corresponding to the lower value of the shape parameter, P) is preferred in the first stage, then using a larger value of r , i.e., $r = 1/2$, is recommended. As the shape parameter increases to 0.5, the results suggest that the performance of designs using smaller values of r (such as $r = 1/3$) would be greatly improved and approach that of a design using $r = 1/2$. When $P = 0.5$, designs using $r = 1/2$ and $r = 1/3$ have comparable performance.

Table 2 in Appendix B of the [supplementary material](#) available at *Biostatistics* online presents the power of the group sequential strategy and the default strategy (i.e., for scenarios with $\lambda = 0$). A small loss of power can be observed using the group sequential design compared with the default strategy when sample size is large. However, such minor loss of power is negligible, especially considering its great advantage in allowing the evaluation of more candidate biomarkers and the successful validation of more truly useful biomarkers.

3.2. Use of analytical formulas for design recommendation

In section 2.3, we developed analytical formulas for the expected number of biomarkers that can be evaluated and the expected number of truly useful biomarkers that are successfully validated, using a given amount of specimens from a reference set. These analytical results can provide valuable guidance for the choices of validation design in practice.

As an example, we compute the expected numbers of m and n_m given $M = 50$ using the analytical formulas presented in Results 1 and 3, and plot in Figure 1 the values of $E(m)$, $E(n_m)$, $E(n_m) - E(n_0)$ and $[E(n_m) - E(n_0)]/E(n_0)$ against λ using the proposed designs and the default design. The plots show that the proposed two-stage strategy can evaluate more candidate biomarkers and successfully validate more truly useful markers than the default design. The proposed design with Stop = “Both” and the shape parameter $P = 0.5$ gives the maximum values of $E(m)$ and $E(n_m)$ among all designs considered. Its advantage is apparent especially for small λ . For both designs with stopping rule Stop = “Both” and Stop = “Null”,

Fig. 1. Comparison of the expected numbers of biomarkers evaluated ($E(m)$) and truly useful biomarkers that are successfully validated ($E(n_m)$) using the two-stage validation strategy with rotation of group membership and the default strategy. Results were calculated using case/control samples with a 1:1 ratio and total sample size $N = 120, 240, 540$, or 1200. Different types of stopping rule (Stop = "Both" or "Null") and boundary shape ($P = 0$ or 0.5) of group sequential testing are represented by the dashed (Stop = "Both", $P = 0$), dotted (Stop = "Both", $P = 0.5$), dot-dash (Stop = "Null", $P = 0$), and long-dash (Stop = "Null", $P = 0.5$) lines. Solid lines represent the results from the default design. $E(n_0)$ denotes the expected number of truly useful biomarkers successfully validated based on the default design.

the designs using the Pocock boundary have the best performance in terms of $E(m)$ and $E(n_m)$. These findings are consistent with what we observed in our simulation results. Furthermore, we notice that as λ increases, $E(n_m)$ decreases because the total number of truly useful biomarkers decreases, but its relative improvement over the default design $[E(n_m) - E(n_0)]/E(n_0)$ increases as λ increases to approach 1. Note that $[E(n_m) - E(n_0)]/E(n_0)$ is undefined when $\lambda = 1$, since $E(n_0) = 0$ when there are no truly useful biomarkers.

To further compare the theoretical results with the numerical results, we next plot the values of $E(m)$ and $E(n_m)$ using the analytical formulas along with the corresponding values based on Monte-Carlo studies. Figure 2 shows that as the sample size increases, the numerical results based on Monte-Carlo simulations are closer to the corresponding theoretical values. For smaller sample sizes ($N = 120, 240$), the finite sample performances of $E(m)$ ($E(n_m)$) tend to be slightly smaller (larger) than the corresponding theoretical results. However, the pattern is similar for both the proposed design and the default design, so the relative order in performance of the two designs is preserved. When sample size increases to $N = 540$, the finite

Fig. 2. Comparison of the expected numbers of biomarkers evaluated and truly useful biomarkers that are successfully validated, as obtained from analytical formulas, with those obtained from the simulation results, for case/control samples with a 1:1 ratio and a total sample size of $N = 120, 240$, or 540 . Results are shown for the default design and proposed two-stage validation design using different types of stopping rule (Stop = "Both" or "Null") and shape parameter ($P = 0$ or 1). Solid black and dashed gray lines represent the corresponding values of $E(m)$ or $E(n_m)$ using the proposed and default strategies, respectively. Black circles and gray rectangles represent the corresponding values of $E(m)$ or $E(n_m)$ using the proposed and default strategies, respectively, based on 5000 Monte Carlo simulations.

sample performance of each design is very close to the corresponding theoretical result. Therefore, with moderate or large sample size ($N > 500$), the developed analytical formula can provide an accurate estimation of the design performance. When sample size is small, the analytical formula can again be used to provide preliminary results for investigating performance of the proposed designs under various parameters, such as percent of truly useful markers and sample size, followed by Monte Carlo studies for more precise evaluation under a few scenarios of particular interest.

4. PANCREATIC CANCER REFERENCE SET STUDY

In this section, we demonstrate the application of the proposed two-stage validation strategy using a numerical setting constructed based on data from a multicenter EDNR pancreatic cancer reference study (Haab and others, 2015). This pancreatic cancer reference set study was developed to facilitate the validation of biomarkers for pancreatic cancer early detection. Blood samples were collected at five different institutions under a standard operating procedure, from 98 patients with pancreatic cancer, 62 patients with chronic pancreatitis, 31 patients with benign biliary obstruction, and 61 healthy individuals. For this analysis, we consider validation of biomarkers for separating patients with pancreatic cancer from healthy controls, i.e., $n_D = 98$ and $n_{\bar{D}} = 61$.

The reference set dataset includes measurements of a set of biomarkers from every study participant, including ATQ, TNC, FIILC, AT, TFPI, CA199:CSLEX1, CSLEX1:CSLEX1, GDF15, MUC5AC, and ATATQ. We create a numerical setting that mimics the biomarker performance observed in this dataset by bootstrapping participant data with replacement stratified on case/control status and then randomly selecting one marker among the set as the incoming biomarker for validation each time. We assume there are $M = 50$ volume units of specimen from each participant, and the evaluation of a single biomarker requires one unit volume of specimen. For validation of each single biomarker, we perform a one-sided test for specificity at sensitivity = 0.9. That is, we consider the null hypothesis $H_0 : SPE_{SEN=0.9} \leq SPE_{SEN=0.9}^0$ with type I error rate controlled at $\alpha = 0.05$. Note that the methodology we developed for testing $ROC(t)$ can be simply applied here by reversing case/control status and reverting the sign of the biomarker. The value of $SPE_{SEN=0.9}$ for the set of markers varies within $\{0.016, 0.039, 0.048, 0.110, 0.143, 0.146, 0.164, 0.292, 0.325, 0.454\}$. We set the value of $SPE_{SEN=0.9}^0 = 0.148$ based on the performance of a standard CA19-9 assay. We consider the proposed design where a biomarker is first evaluated using specimens from half or one-third of participants, i.e., $r = 1/2$ or $1/3$. Again, two different types of stopping rule (Stop = “Both” and Stop = “Null”) and boundary shape (O’Brien–Fleming boundary and Pocock boundary) are adopted. Performance of designs is evaluated by examining the average number of biomarkers evaluated, $E(m)$, and average number of truly useful biomarkers that are successfully validated, $E(n_m)$.

Table 2 shows the results for the validation of biomarkers using the proposed two-stage strategy and the default strategy. Obviously, adopting the proposed validation strategy allows us to evaluate more candidate biomarkers and eventually be able to successfully validate more truly useful biomarkers. Furthermore, it can be seen that using the Pocock boundary gives the best performance for either the design allowing early stopping for both futility and efficacy or the design allowing early stopping for futility only. These findings are consistent with that from the simulation studies.

5. CONCLUDING REMARKS

In this article, we considered design strategies for validating multiple candidate biomarkers using a reference set with limited specimen volume. In order to efficiently utilize the valuable specimens from the reference set, we proposed a new strategy for validating candidate biomarkers that incorporates a two-stage validation process with rotation of group membership, which allows the validation of some biomarkers

Table 2. Performance of the default design ($K = 1$) and the proposed two-stage validation design with rotation of group membership ($K = 2$) using different types of stopping rules (Stop = “Both” or “Null”) and boundary shapes ($P = 0$ or 0.5), for the numerical setting constructed based on data from the pancreatic cancer reference set study. The table shows the average number of biomarkers evaluated ($E(m)$), the average number of truly useful biomarkers that are successfully validated ($E(n_m)$), and the corresponding standard error of m and n_m (in parentheses)

r	Performance	$K = 2$				$K = 1$
		Stop = “Both”		Stop = “Null”		
		$P = 0$	$P = 0.5$	$P = 0$	$P = 0.5$	
$\frac{1}{2}$	$E(m)$	77.77 (0.04)	89.17 (0.04)	73.61 (0.03)	77.69 (0.04)	50
	$E(n_m)$	7.19 (0.02)	7.38 (0.03)	6.40 (0.02)	6.67 (0.02)	5.04 (0.02)
$\frac{1}{3}$	$E(m)$	74.92 (0.05)	100.71 (0.07)	71.33 (0.04)	86.69 (0.06)	50
	$E(n_m)$	7.42 (0.02)	9.76 (0.03)	6.66 (0.02)	7.60 (0.02)	5.04 (0.02)

using samples from only a subset of the study participants. Compared with the default strategy that validates every biomarker using samples from all study participants, the advantage of the proposed validation strategy is clearly demonstrated in our numerical studies. In particular, it allows the evaluation of more candidate biomarkers with the reference set samples, and eventually allows more truly useful biomarkers to be successfully validated. We adopted the sequential testing technique in the two-stage design and recommend the use of the Pocock boundary, which appears to perform best for either the design allowing early termination for futility and efficacy or the design allowing early termination for futility only. Moreover, we developed analytical formulas for calculating the expected numbers of candidate biomarkers that can be evaluated and the truly useful biomarkers that can be successfully validated. These analytical results can provide valuable guidance to researchers in practice for selecting validation strategies.

We, here, assume a simple setting where equal volumes of specimens are available from each participant. When this assumption does not hold, i.e., when volume of specimens varies among participants, the basic idea of our proposed design is the same while some technical details need to be modified. Instead of randomly partitioning the participants in the reference set into Q blocks, the participants should be grouped according to the available specimen volume. To begin with, we list the participants in descending order by the volume of specimens, and then partition the ordered participants into Q blocks to ensure rough balance in specimen volumes across blocks. One of blocks with the maximum average amount of specimens is randomly selected and treated as group 1, then the rest of blocks are treated as group 2. The sequential testing procedure is the same as that for the simple setting. Each time when a new biomarker enters the validation process, we randomly draw one block with the maximum average amount of specimens as group 1 and then perform the test again. This procedure can be repeated until the number of participants who have adequate specimen remaining is low.

For the validation of biomarker, in this article, we considered a hypothesis test with respect to the points on the ROC curve. However, the proposed method is generally applicable to other classification measures, such as the AUC or partial AUC. Moreover, while we here focused on the group sequential design with two analyses for ease of implementation and given the popularity of two-stage designs in sequential trials, our proposed strategy can be extended to designs with multiple interim analyses.

In this article, our main interest is to provide a sample source (a common reference set) to allow different labs to have an early phase validation of their candidate biomarker performance. Multiplicity adjustment is not deemed essential here and is not always feasible: candidate biomarkers are coming in one by one and validation results are usually reported to individual labs during the process before all samples are

used up. There are scenarios however when multiplicity adjustment could be of interest. For example, at the end of the study when all samples have been used up, there might be plans for follow-up studies and interests to have a simultaneous review of all biomarkers, in order to identify a small subset to be validated in the next step. Multiplicity adjustment can be easily accommodated in our proposed design to control for criteria such as the false discovery rate or family-wise error rate. Instead of declaring significance of individual biomarkers based on their nominal p -values, one can apply p -value adjustment methods to control for the desired error rate among all biomarkers that have been evaluated and make decisions based on the adjusted p -value. To illustrate how this works, we have applied the false discovery rate adjustment method proposed by Benjamini and Hochberg (1995) to adjust the nominal p -value of individual tests to control the false discovery rate at 0.05 level. The results are reported in Table 3 in Appendix C of the [supplementary material](#) available at *Biostatistics* online. The proposed two-stage strategy still performs better than the default design; it successfully validated more truly useful biomarkers, especially under the scenario where majority of biomarkers are not actually useful.

Finally, in practice one might have some established biomarker(s) for detecting a certain disease, and the goal is to evaluate individual markers for combining with the established biomarker(s) to develop a new classification panel. For this problem, one can first measure the established biomarkers from samples in all participants in the reference set and then apply the proposed design to the rest of the specimens. This is a topic for future research.

6. SOFTWARE

The R code is available at <https://github.com/WangLu88/ValidationDesign.git>.

SUPPLEMENTARY MATERIAL

[Supplementary material](http://biostatistics.oxfordjournals.org) is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This work was supported by the U.S. National Institutes of Health under awards R01GM106177 and U24CA086368.

REFERENCES

- BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300.
- EMERSON, S. S. AND FLEMING, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905–923.
- FENG, Z., KAGAN, J., PEPE, M., THORNQUIST, M., RINAUDO, J. A., DAHLGREN, J., KRUEGER, K., ZHENG, Y., PATRIOTIS, C., HUANG, Y. and others. (2013). The early detection research network's specimen reference sets: paving the way for rapid evaluation of potential biomarkers. *Clinical Chemistry* **59**, 68–74.
- HAAB, B. B., HUANG, Y., BALASENTHIL, S., PARTYKA, K., TANG, H., ANDERSON, M., ALLEN, P., SASSON, A., ZEH, H., KAUL, K. and others. (2015). Definitive characterization of ca 19-9 in resectable pancreatic cancer using a reference set of serum and plasma specimens. *PLoS One* **10**, e0139049.
- JENNISON, C. AND TURNBULL, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC.

- KOOPMEINERS, J. S. AND FENG, Z. (2011). Asymptotic properties of the sequential empirical roc, ppv and npv curves under case-control sampling. *Annals of Statistics* **39**, 3234.
- KOOPMEINERS, J. S., FENG, Z. AND PEPE, M. S. (2012). Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics in Medicine* **31**, 420–435.
- O'BRIEN, P. C. AND FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.
- PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press.
- PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THORNQUIST, M., WINGET, M. AND YASUI, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 1054–1061.
- PEPE, M. S., FENG, Z., LONGTON, G. AND KOOPMEINERS, J. (2009). Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Statistics in Medicine* **28**, 762–779.
- POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.
- ZHOU, X. H., MCCLISH, D. K. AND OBUCHOWSKI, N. A. (2009). *Statistical Methods in Diagnostic Medicine*, Volume 569. John Wiley & Sons.

[Received January 7, 2019; revised June 1, 2019; accepted for publication July 15, 2019]