Published in final edited form as:

Stat Med. 2016 April 15; 35(8): 1267–1280. doi:10.1002/sim.6790.

Group Sequential Testing of the Predictive Accuracy of a Continuous Biomarker with Unknown Prevalence

Joseph S. Koopmeiners^{a,*} and Ziding Feng^b

^aDivision of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A

^bDepartment of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77230, U.S.A

Abstract

Group sequential testing procedures have been proposed as an approach to conserving resources in biomarker validation studies. Previously, Koopmeiners and Feng (2011) derived the asymptotic properties of the sequential empirical positive predictive value (PPV) and negative predictive value curves, which summarize the predictive accuracy of a continuous marker, under case-control sampling. A limitation of their approach is that the prevalence can not be estimated from a case-control study and must be assumed known. In this manuscript, we consider group sequential testing of the predictive accuracy of a continuous biomarker with unknown prevalence. First, we develop asymptotic theory for the sequential empirical PPV and NPV curves when the prevalence must be estimated, rather than assumed known in a case-control study. We then discuss how our results can be combined with standard group sequential methods to develop group sequential testing procedures and bias-adjusted estimators for the PPV and NPV curve. The small sample properties of the proposed group sequential testing procedures and estimators are evaluated by simulation and we illustrate our approach in the context of a study to validate a novel biomarker for prostate cancer.

Keywords

Diagnostic Biomarkers; Group Sequential Methods; PPV Curve; Prostate Cancer

1. Introduction

Diagnostic biomarkers are used to discriminate between diseased (cases) and non-diseased (controls) individuals. One approach to evaluating a diagnostic test is to consider the predictive accuracy, which refers to the test's ability to, conditional on the biomarker value, predict if a subject is truly a case or a control. For a dichotomous test, predictive accuracy is summarized by the positive predictive value (PPV), defined as the probability that a subject is truly a case given that they have a positive test result, and the negative predictive value (NPV), defined as the probability that a subject is truly a control given that they have a

*Correspondence to: Joseph S. Koopmeiners, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455. koopm007@umn.edu.

negative test result. The PPV and NPV curves were proposed as an extension of PPV and NPV to continuous markers [1, 2].

Let D be a Bernoulli random variable indicating disease status with prevalence θ and let X be a biomarker value with conditional distribution $F(x|D = 1) = F_D(x)$ and $F(x|D = 0) = F_{\bar{D}}(x)$. The marginal distribution of X is therefore $F(x) = \theta F_D(x) + (1 - \theta) F_{\bar{D}}(x)$. Assume that larger biomarker values are more indicative of disease. For a given threshold c , the predictive accuracy of the biomarker can be summarized by the positive predictive value, $PPV(c) = P[D = 1|X > c]$, and the negative predictive value, $NPV(c) = P[D = 0|X \leq c]$. The PPV and NPV curves are defined as $PPV(c)$ and $NPV(c)$ for all $c \in (-\infty, \infty)$. The PPV and NPV curves can alternately be indexed by a summary of the marker distribution rather than a generic threshold. In this paper, we consider the PPV and NPV curves indexed by population quantiles.

Diagnostic biomarkers are often evaluated using banks of stored biospecimens. The procurement of high-quality specimens is an expensive and time-consuming process. It is important to use these resources wisely in order to maximize the benefits gained from developing biospecimen banks. Group sequential study designs have been proposed as an approach to conserving specimens when evaluating diagnostic biomarkers for classification or prediction [3]. Group sequential methods allow a study to terminate early if a conclusion has been reached, which conserves the remaining specimens for use in future studies of other markers.

Group sequential methods for evaluating diagnostic tests have been discussed in the literature [4, 5, 6, 7]. These methods focus on evaluating the classification accuracy of the novel test using the receiver operating characteristic (ROC) curve and can not be used for evaluating the predictive accuracy of a novel biomarker. Koopmeiners and Feng [8] investigated the asymptotic properties of the sequential empirical PPV and NPV curves under case-control sampling. They showed that the sequential empirical PPV and NPV curves converge to the sum of independent Kiefer processes and that summaries of the sequential empirical PPV and NPV curves are asymptotically normal with an independent increments covariance structure. This work provides the theoretical foundation for group sequential testing of the PPV and NPV curves but additional work is needed to address the challenges associated with implementing group sequential testing of the PPV and NPV curves in practice.

In this manuscript, we consider group sequential testing of the predictive accuracy of a diagnostic biomarker with unknown prevalence. First, we discuss sequential empirical estimation of the PPV and NPV curves and extend the results of Koopmeiners and Feng [8] to the setting where the prevalence must be estimated, rather than assumed known in a case-control study. This is a non-trivial extension that requires us to account for the correlation between estimates of the prevalence and the distribution of the biomarker. We then evaluate the performance of standard approaches to group sequential testing and estimation when applied to the PPV and NPV curves through simulation. Our simulation results show that the application of group sequential testing procedures results in study designs with the same type-I error and power as the fixed-sample design but with a dramatically smaller sample

size and that standard bias-adjusted estimators can be used to correct for the bias caused by the possibility of early termination. We begin, though, by introducing the PCA3 validation study, which will serve as our illustrative example throughout this manuscript.

2. PCA3 Validation Study

PCA3 is a novel urine biomarker for the detection of prostate cancer. Initial studies [9] indicated that PCA3 has better performance than PSA, the most commonly used biomarker for the detection of prostate cancer, for prostate cancer diagnosis (i.e. predicting positive biopsy for prostate cancer) and a larger, validation study was recently completed to confirm these findings [10]. Complicating the evaluation of PCA3 is the presence of two distinct subgroups with different prostate cancer risk. Men in the initial biopsy group, made up of men with no previous prostate biopsies, are at higher risk for prostate cancer, whereas men in the repeat biopsy group, made up of men with a previous negative prostate biopsy, are at lower risk for prostate cancer. PCA3 will be evaluated separately in these two groups to account for the differences in prostate cancer risk. Researchers would like to evaluate the predictive accuracy of PCA3 and are primarily interested in the PPV for the initial biopsy group and the NPV for the repeat biopsy group. We will focus only on the initial biopsy group for the purposes of our illustration.

A secondary aim of the PCA3 validation study was to develop a urinary reference set for future pre-validation studies of urinary biomarkers. To accomplish this goal, the PCA3 study included a single interim analysis with a stopping rule for futility. In the context of this study, the potential for early termination refers to the decision to assay the samples, rather than the decision to enroll more subjects, as in a sequential clinical trial. Specifically, a fixed number of men were enrolled regardless of the decision to terminate the study for futility. At the interim analysis, half of the samples were used to determine if the study should terminate for futility. If the study terminates for futility, the remaining samples are not assayed and can be stored for future evaluation. Otherwise, the remaining samples are assayed and used to evaluate the performance of PCA3. This allows the remaining samples to be used in future studies if PCA3 does not appear promising at the interim analysis.

The PCA3 validation study was designed to stop for futility at the interim analysis if the upper limit of the 95% confidence interval for the parameter of interest was less than the alternative hypothesis. The option of early termination for futility will not inflate the type-I error rate but this stopping rule is somewhat ad hoc in that the overall type-I error rate will be less than the nominal rate, which will also reduce the power of the study. In contrast, a formal group sequential stopping rule for futility would result in an overall type-I error rate equal to the nominal rate and a more powerful design. Replacing the stopping rule used in the PCA3 validation study with a more rigorously justified stopping rules would be preferable and would also allow researchers to implement stopping rules for superiority and futility, which would allow additional samples to be stored if the predictive accuracy of PCA3 clearly exceeds the desired level of performance.

The remainder of this manuscript proceeds as follows. In Section 3 we discuss sequential empirical estimation of the PPV and NPV curves and extend the theoretical results of

Koopmeiners and Feng [8] to the case where the prevalence is estimated in a cohort study. In Section 4, we evaluate the performance of standard approaches to group sequential testing and estimation when applied to the PPV and NPV curves and, in Section 5, we illustrate group sequential testing of the predictive accuracy of a diagnostic biomarker using the PCA3 validation study. We conclude with a brief discussion in Section 6.

3. Sequential Empirical Estimation of the PPV and NPV Curves

3.1. Notation and Definitions

In Section 1, we presented the PPV and NPV curves indexed by a generic threshold, c . In practice, the PPV and NPV curves are indexed by a clinically meaningful quantity and the threshold, c , is a function of the biomarker distribution. The most common approach is to index the PPV and NPV curves by the proportion of the population that are classified as negative, u , and positive, $1 - u$ [1, 11, 12]. The advantage of this indexing is that it allows multiple markers to be compared on a common scale by transforming any marker value to its corresponding population quantile. In this case, the PPV and NPV curves are defined as $PPV(u) = P[D = 1 | X > F^{-1}(u)]$ and $NPV(u) = P[D = 0 | X \leq F^{-1}(u)]$ for all $u \in (0, 1)$. Under this indexing, the PPV curve can be written as

$$PPV(u) = \frac{S_D(F^{-1}(u)) \theta}{1 - u}, \quad (1)$$

where $S_D(x) = 1 - F_D(x)$. Similarly, the NPV curve can be written as

$$NPV(u) = \frac{F_D(F^{-1}(u)) (1 - \theta)}{u}.$$

It should also be noted that the NPV curve can be expressed as a function of the PPV curve

$$NPV(u) = \frac{u - \theta}{u} + \frac{1 - u}{u} PPV(u), \quad (2)$$

and, therefore, it suffices to study the PPV curve when considering estimation of the PPV and NPV curves.

Koopmeiners and Feng [8] considered the sequential empirical PPV and NPV curves under case-control sampling, which requires that the prevalence be known. In most cases, including the PCA3 validation study, the prevalence must be estimated from a cohort study. The sequential empirical estimate of $PPV(u)$ under cohort sampling is defined as,

$$\widehat{PPV}_r(u) = \frac{\hat{S}_{D,r}(\hat{F}_r^{-1}(u)) \hat{\theta}_r}{1 - u}, \quad (3)$$

where $\hat{\theta}_r$, $\hat{S}_{D,r}(x)$ and $\hat{F}_r^{-1}(u)$ are the sequential empirical estimates of θ , $S_D(x)$ and $F^{-1}(u)$, respectively, and $r \in (0, 1]$ is the proportion of observations observed at a given time point. Here,

$$\hat{\theta}_r = \begin{cases} 0, & 0 \leq r < \frac{1}{n}, \\ \frac{1}{[rn]} \sum_{i=1}^{[rn]} D_i, & -\infty < x < \infty, \quad \frac{1}{n} \leq r \leq 1, \end{cases}$$

and $\hat{S}_{D,r}(x) = 1 - \hat{F}_{D,r}(x)$, where,

$$\hat{F}_{D,r}(x) = \begin{cases} 0, & \hat{\theta}_r = 0, \\ \frac{\hat{P}_r(X \leq x, D=1)}{\hat{\theta}_r}, & -\infty < x < \infty, \quad \hat{\theta}_r > 0, \end{cases}$$

and

$$\hat{P}_r(X \leq x, D=1) = \begin{cases} 0, & 0 \leq r < \frac{1}{n}, \\ \frac{1}{[rn]} \sum_{i=1}^{[rn]} 1\{X_i \leq x, D_i=1\}, & -\infty < x < \infty, \quad \frac{1}{n} \leq r \leq 1. \end{cases}$$

Finally, the sequential empirical estimate of $F^{-1}(u)$, $\hat{F}_r^{-1}(u)$, is defined as in [13],

$$\hat{F}_r^{-1}(t) = \begin{cases} X_{1,[rn]} & \text{if } t=0, \quad 0 \leq r \leq 1, \\ X_{k,[rn]} & \text{if } \frac{k-1}{[rn]} < t \leq \frac{k}{[rn]}, \\ & 1 \leq k \leq [rn], \quad 0 \leq t \leq 1, \end{cases}$$

where $X_{1,[rn]}, X_{2,[rn]}, \dots, X_{[rn],[rn]}$ are the sequential order statistics of the biomarker values in the entire population. The sequential empirical estimate of $NPV(u)$ is defined by substituting the sequential empirical estimates of $PPV(u)$ and θ into (2),

$$\widehat{NPV}_r(u) = \frac{u - \hat{\theta}_r}{u} + \frac{1-u}{u} \widehat{PPV}_r(u).$$

Finally, we define the following,

$$P_r(u) = n^{-1/2} [nr] \left(\widehat{PPV}_r(u) - PPV(u) \right),$$

and

$$N_r(u) = n^{-1/2} [nr] \left(\widehat{NPV}_r(u) - NPV(u) \right),$$

for ease of notation when stating the asymptotic results presented in Section 3.2.

3.2. Asymptotic Results and Group Sequential Testing

We next present asymptotic results for $\widehat{PPV}_r(u)$ and $\widehat{NPV}_r(u)$, which extend the results of Koopmeiners and Feng [8] to the case where the prevalence must be estimated from the data. Under cohort sampling, the sequential empirical PPV curve converges to the sum of independent Kiefer processes and an independent Brownian motion, where the Brownian motion captures the variability due to estimating the prevalence. Using this result, we can derive a similar result for the sequential empirical NPV curve.

Theorem 3.1—Assume A1 – A3 from the Supplementary Materials hold, $\theta \in (0, 1)$ and let

$\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))}$ be bounded on $[a, b]$. As $n \rightarrow \infty$

$$\begin{aligned} P_r(u) \rightarrow_D & -\frac{\theta(1-\theta)}{1-u} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} \frac{1}{\sqrt{\theta}} K_1(F_D(F^{-1}(u)), r) \\ & + \frac{\theta(1-\theta)}{1-u} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} \frac{1}{\sqrt{1-\theta}} K_2(F_D(F^{-1}(u)), r) \\ & + \left(1 - \frac{(1-\theta)f_D(F^{-1}(u))}{f(F^{-1}(u))} F_D(F^{-1}(u)) - \frac{\theta f_D(F^{-1}(u))}{f(F^{-1}(u))} F_D(F^{-1}(u)) \right) \frac{\sqrt{\theta(1-\theta)}}{1-u} W(r) \end{aligned}$$

uniformly for $u \in [a, b]$ and $r \in [c, 1]$ where K_1 and K_2 are independent Kiefer Processes and W is Brownian motion independent of K_1 and K_2 .

A proof of Theorem 3.1 can be found in the Supplementary Materials. The proof closely follows the arguments found in [14]. The primary difficulty to the proof of Theorem 3.1 is dealing with correlation between $\hat{S}_{D,r}(x)$ and $\hat{F}_r^{-1}(t)$. By noting that $F_r(\hat{x})$ can be written as a function of $F_{D,r}(\hat{x})$, $F_{D,r}(\hat{x})$ and $\hat{\theta}_r$, we are able to decompose $P_r(u)$ into three terms that are asymptotically independent and we arrive at the above result. Corollary 3.2, which provides a similar result for $N_r(u)$, is a direct result of Theorem 3.1 and equation (2).

Corollary 3.2—Assume A1 – A3 from the Supplementary Materials hold, $\theta \in (0, 1)$ and let

$\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))}$ be bounded on $[a, b]$. As $n \rightarrow \infty$

$$\begin{aligned} N_r(u) \rightarrow_D & -\frac{\theta(1-\theta)}{u} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} \frac{1}{\sqrt{\theta}} K_1(F_D(F^{-1}(u)), r) \\ & + \frac{\theta(1-\theta)}{u} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} \frac{1}{\sqrt{1-\theta}} K_2(F_D(F^{-1}(u)), r) \\ & - \left(\frac{(1-\theta)f_D(F^{-1}(u))}{f(F^{-1}(u))} F_D(F^{-1}(u)) + \frac{\theta f_D(F^{-1}(u))}{f(F^{-1}(u))} F_D(F^{-1}(u)) \right) \frac{\sqrt{\theta(1-\theta)}}{u} W(r) \end{aligned}$$

uniformly for $u \in [a, b]$, $r \in [c, 1]$ where K_1 and K_2 are independent Kiefer Processes and W is Brownian motion independent of K_1 and K_2 .

Theorem 3.1 and Corollary 3.2 provide important insight into the asymptotic behavior of $P_r(u)$ and $N_r(u)$. First, asymptotic results for summaries of the PPV and NPV curve (i.e. a point or points on the curve, etc.) can be derived from Theorem 3.1 and Corollary 3.2 as a special case. In the remainder of this manuscript, we discuss how standard approaches to group sequential testing can be applied to testing a point on the PPV curve due to these results. Second, Theorem 3.1 and Corollary 3.2 illustrate that $P_r(u)$ and $N_r(u)$ converge to the sum of the same Kiefer processes and Brownian motion. This allows us to evaluate the joint behavior of the PPV and NPV curves. We do not exploit this result in this manuscript but this information could provide useful if a researcher were interested in simultaneously evaluating the PPV and NPV curves at different cut-points.

We now focus our attention on group sequential testing of $PPV(u)$ and $NPV(u)$, a point on the PPV and NPV curve, respectively. We can derive the asymptotic distribution of the sequential empirical estimators of $PPV(u)$ and $NPV(u)$, $\widehat{PPV}_r(u)$ and $\widehat{NPV}_r(u)$, as a special case of Theorem 3.1 and Corollary 3.2. Corollary 3.3 establishes that the sequential empirical estimates of $PPV(u)$ and $NPV(u)$, a point on the PPV and NPV curve, respectively, are asymptotically normal with an independent increments covariance structure.

Corollary 3.3—Assume A1 – A3 from the Supplementary Material hold, $\theta \in (0, 1)$ and let

$\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))}$ be bounded on $[a, b]$. For $u \in (0, 1)$ and J stopping times,

A. $(\widehat{PPV}_{r_1}(u), \widehat{PPV}_{r_2}(u), \dots, \widehat{PPV}_{r_J}(u))$ is approximately multivariate normal with,

$$\widehat{PPV}_{r_i}(u) \sim N\left(PPV(u), \sigma_{PPV(u), r_i}^2\right) \quad i=1, 2, \dots, J$$

and

$$Cov\left[\widehat{PPV}_{r_i}(u), \widehat{PPV}_{r_j}(u)\right] = Var\left[\widehat{PPV}_{r_j}(u)\right] = \sigma_{PPV(u), r_j}^2, \quad r_i \leq r_j$$

where

$$\begin{aligned} \sigma_{PPV(u), r_i}^2 = & \frac{\left(\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))}(1-\theta)\right)^2 PPV(u) \left(\frac{\theta}{1-u} - PPV(u)\right)}{\theta n r_i} \\ & + \frac{\left(\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))}\theta\right)^2 (1-PPV(u)) \left(\frac{u-\theta}{1-u} + PPV(u)\right)}{(1-\theta) n r_i} \\ & + \frac{\left(\frac{(1-\theta)}{\theta} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} PPV(u) + \frac{\theta}{(1-\theta)} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} (1-PPV(u))\right)^2}{n r_i}. \end{aligned}$$

- B.** $(\widehat{NPV}_{r_1}(u), \widehat{NPV}_{r_2}(u), \dots, \widehat{NPV}_{r_J}(u))$, is approximately multivariate normal with,

$$\widehat{NPV}_{r_i}(u) \sim N(NPV(u), \sigma_{NPV(u), r_i}^2) \quad i=1, 2, \dots, J$$

and

$$Cov[\widehat{NPV}_{r_i}(u), \widehat{NPV}_{r_j}(u)] = Var[\widehat{NPV}_{r_j}(u)] = \sigma_{NPV(u), r_j}^2, \quad r_i \leq r_j$$

where

$$\begin{aligned} \sigma_{NPV(u), r_i}^2 = & \frac{\left(\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} (1-\theta) \right)^2 (NPV(u) + \frac{\theta-u}{u}) (1-NPV(u))}{\theta n r_i} \\ & + \frac{\left(\frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} \theta \right)^2 NPV(u) \left(\frac{1-\theta}{u} - NPV(u) \right)}{(1-\theta) n r_i} \\ & + \frac{\left(\frac{(1-\theta)}{\theta} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} (1-NPV(u)) + \frac{\theta}{(1-\theta)} \frac{f_D(F^{-1}(u))}{f(F^{-1}(u))} NPV(u) \right)^2}{n r_i}. \end{aligned}$$

A proof of Corollary 3.3 can be found in the Supplementary Materials. Group sequential testing of $PPV(u)$ and $NPV(u)$ can be completed using the following standardized test statistics:

$$Z_{PPV(u), r} = \frac{\widehat{PPV}_r(u) - PPV(u)_0}{\sigma_{PPV(u)_0, r}} \quad \text{and} \quad Z_{NPV(u), r} = \frac{\widehat{NPV}_r(u) - NPV(u)_0}{\sigma_{NPV(u)_0, r}} \quad (4)$$

It is immediate from Corollary 3.3 that, for any r , $Z_{PPV(u), r}$ and $Z_{NPV(u), r}$ follow standard normal distributions under the null hypothesis and that

$$Cov[Z_{PPV(u), r_i}, Z_{PPV(u), r_j}] = \frac{\sigma_{PPV(u)_0, r_j}}{\sigma_{PPV(u)_0, r_i}} \quad \text{and} \quad Cov[Z_{NPV(u), r_i}, Z_{NPV(u), r_j}] = \frac{\sigma_{NPV(u)_0, r_j}}{\sigma_{NPV(u)_0, r_i}}$$

for $r_i \leq r_j$. This result, along with Corollary 3.3, are important results that allow standard approaches to group sequential testing and estimation to be applied to $PPV(u)$ and $NPV(u)$. We will evaluate the small sample properties of these methods in Section 4 and illustrate the application of these methods in the context of the PCA3 validation study in Section 5.

4. Finite Sample Properties

We completed a simulation study to evaluate the type-I error rate and power of a group sequential test of $PPV(u)$ in finite samples. Disease status was simulated from a Bernoulli distribution with a prevalence of θ . Biomarker values for the controls were drawn from a

standard normal distribution and biomarker values for the cases were simulated from a normal distribution with a standard deviation equal to 1 and mean chosen to achieve the desired value of $PPV(u)$. For each simulated study, we complete a one-sided α -level test of the null hypothesis $H_0 : PPV(u) = PPV(u)_0$ using the test statistic defined in (4). For a fixed-sample test, we compare $Z_{PPV(u),1}$ to the standard normal distribution and reject the null hypothesis if $Z_{PPV(u),1} > Z_{1-\alpha}$. For a group sequential test with K stopping times, we compare the sequence of test statistics $(Z_{PPV(u),r1}, \dots, Z_{PPV(u),1})$ to group sequential stopping boundaries derived to achieve the desired type-I error rate. For our simulations, we consider the one-sided, symmetric boundaries proposed by [15]. We will consider symmetric boundaries with shape parameters equal to 0 and 0.5. A shape parameter of 0 will result in boundaries that are similar to O'Brien-Fleming boundaries in that more extreme critical values are used early in the trial, while less extreme critical values are used to declare a significant difference at the end of the trial. In contrast, a shape parameter of 0.5 will result in boundaries that are similar to Pocock boundaries in that a constant upper boundary is used to declare superiority. The symmetric boundaries used in our simulations were calculated using the RCTdesign package in R [16].

We considered scenarios with θ equal to 0.3 or 0.4, u equal to 0.7 or 0.8 and $PPV(u)_0$ equal to 0.6 or 0.7. Setting u equal to 0.7 and $PPV(u)_0$ equal to 0.6, for example, implies that we are completing a one-sided, $\alpha = 0.05$ hypothesis test of $PPV(0.7) = 0.6$. That is, we are testing whether or not the positive predictive value is less than 0.6 when the top 30% of the population are considered positive and the bottom 70% are considered negative. We ran two sets of simulations for all scenarios. In the first set of simulations, we set n equal to 250 or 500 and evaluated the type-I error rate when group sequential testing is applied to $PPV(u)$. In the second set of simulations, we fixed the power at 0.80 or 0.90 for detecting a difference of 0.1 in $PPV(u)$ and evaluated the power to determine if the results in Section 3 can be used to prospectively design a group sequential biomarker validation study with the appropriate power. In both sets of simulations, we evaluated the operating characteristics of a fixed-sample test (i.e. a test with no interim analysis) and group sequential tests with 2, 3 and 4 interim analyses. 10,000 simulated studies were completed for all scenarios.

Table 1 provides the simulated type-I error rate for a group sequential test of $PPV(u) = PPV(u)_0$ using one-sided symmetric stopping boundaries with a shape parameter equal to 0. The empirical type-I error rate is close to 0.05 in all cases regardless if a fixed-sample or group sequential design is considered. In addition, we see that the group sequential designs results in a substantial reduction in the expected sample size ($E(n)$) compared to the fixed-sample design, which demonstrates the potential for group sequential designs to conserve specimens through early termination. Results when the shape parameter is set equal to 0.5 can be found in the Supplementary Materials and are similar to those found in Table 1.

Table 2 provides simulated power assuming an alternative hypothesis of $PPV(u)_a = PPV(u)_0 + 0.10$ for a fixed-sample test and a one-sided group sequential test using symmetric boundaries with 2, 3 and 4 interim analyses and a shape parameter equal to 0. Sample size for the fixed-sample test can be calculated using the standard sample size formula for an asymptotically normal test statistic:

$$n = \frac{(Z_{1-\alpha}\sigma_{PPV(u)_0} - Z_{1-\beta}\sigma_{PPV(u)_a})^2}{(PPV(u)_a - PPV(u)_0)^2},$$

where α is the type-I error rate, β is the type-II error rate, $Z_{1-\alpha}$ represents the $1 - \alpha$ quantile of the standard normal distribution and $Z_{1-\beta}$ is defined analogously. Sample size for the group sequential design is determined by multiplying the fixed-sample sample size by an inflation factor that depends on the power and number of interim analyses. We considered designs with 80% and 90% power. We see that the observed power is slightly below the nominal power in most cases. More importantly, though, we see that within a scenario, the power is constant across the fixed-sample and group sequential designs. This suggests that the modest decrease in power is likely due to variability in estimating the variance of $\widehat{PPV}_r(u)$ and not due to utilizing a group sequential design. In addition, we see that the expected sample size for the three group sequential designs is substantially smaller than the corresponding fixed-sample design. Additional simulation results with shape parameter equal to 0.5 can be found in the Supplementary Materials and are similar. The results in Table 2, along with the results in Table 1, illustrate that group sequential testing of the non-parametric PPV curve results in study designs with the same type-I error rate and power as the corresponding fixed-sample design but a substantial reduction in the expected sample size.

In addition, we completed a set of simulations to compare the type-I error and power using the methods described in Section 3, to the case where the prevalence is assumed to be known and $PPV(u)$ is estimated using the method described in [17]. Simulations were completed with u equal to 0.8 and $PPV(u)$ equal to 0.6 or 0.7. For each scenario, we specify a design prevalence that is fixed in advance and used to estimate $PPV(u)$. Data are then simulated assuming a true prevalence that is either equal to the design prevalence or ± 0.05 from the design prevalence. This will allow us to evaluate the impact of incorrectly specifying the design prevalence, as compared to estimating the prevalence from the data.

Simulation results using symmetric stopping boundaries with a shape parameter of 0 can be found in Table 4. First, we present data with $n = 250$ to evaluate the type-I error rate. The simulated type-I error rate is close to 0.05 for both methods when the design prevalence equals the true prevalence but fixing the prevalence results in an inflated type-I error rate when the true prevalence is less than the design prevalence and a type-I error rate that is much lower than 0.05 when the true prevalence is larger than the design prevalence. In contrast, estimating the prevalence allows us to control the type-I error rate when the true prevalence is not equal to the design prevalence. Next, we present data when the study was designed to achieve 80% power for detecting a difference of 0.10 in $PPV(u)$. The results for the power are the opposite of what was observed for the type-I error rate. The power is close to 1 when the true prevalence is less than the design prevalence (when the type-I error rate is inflated), while the power is much lower than 80% when the true prevalence is less than the design prevalence. As was the case for the type-I error rate, estimating the prevalence from the data allows us to achieve the correct power regardless of the relationship between the

true and design prevalence. Results using symmetric stopping boundaries with a shape parameter of 0.5 can be found in the Supplementary Materials and exhibit a similar pattern.

We conclude this section by considering estimation after a group sequential biomarker validation study. It has been shown that standard, unadjusted estimators will be biased at the completion of a group sequential study [18]. The bias arises because the sample size is not an ancillary statistic. As a result, conditioning on the sample size will not yield a consistent estimator and adjusted estimators that correct for the bias due to the possibility for early termination are needed. Similarly, the standard 95% confidence interval will not provide correct coverage and adjusted confidence intervals are needed. Therefore, we completed a simulation study to evaluate the performance of several bias-adjusted estimators and adjusted confidence intervals that account for the possibility of early termination. Estimators considered include the Rao-Blackwell estimator proposed in [19], the mean-adjusted estimator proposed in [18] and the median-adjusted estimator proposed in [20]. Briefly, the Rao-Blackwell estimator is the expected value of $\widehat{PPV}(u)_{r_1}$ (i.e. the estimate of $PPV(u)$ at the first interim analysis) conditional on the stopping time and the unadjusted estimate at study termination, the mean-adjusted estimator is the true value of $PPV(u)$ that results in an expected value of the unadjusted estimator equal to what was observed at termination and the median-adjusted estimator is the true value of $PPV(u)$ that results in a median value of the unadjusted estimator equal to what was observed at termination. Formal definitions of the three estimators can be found in the Supplementary Materials. The median-adjusted estimator and confidence intervals utilize the analysis-time ordering proposed in [21]. All three estimators are functions of the unadjusted estimator and can be derived using the sequential asymptotic distribution presented in Corollary 3.3.

Simulation results with θ equal to 0.4, u equal to 0.7 or 0.8 and one-sided symmetric boundaries with shape parameter equal to 0 can be found in Table 4. We see that the unadjusted estimator ($\widehat{PPV}(u)$) is biased with the bias increasing as the number of interim analyses increases. In contrast, the bias-adjusted estimators, while not unbiased, exhibit less bias than the unadjusted estimator with the Rao-Blackwell estimator ($\widehat{PPV}(u)_{rb}$) exhibiting the best performance. Similarly, the unadjusted confidence interval fails to provide the correct coverage in many cases, with the coverage probability decreasing as the number of interim looks increases. In contrast, the adjusted confidence intervals provide at least 95% coverage in all cases. Additional simulations with similar results can be found in the Supplementary Materials.

5. Illustration Using the PCA3 Validation Study

We now return to the PCA3 validation study to illustrate how the theoretical results found in Section 3 can be applied to group sequential testing of the PPV curve. Recall that the PCA3 validation study was evaluated separately in the initial biopsy and repeat biopsy cohorts but we will focus only on the initial biopsy cohort for the purposes of this illustration.

The PCA3 validation study enrolled 562 men scheduled for prostate biopsy in the initial biopsy cohort. PCA3 is a continuous biomarker but, for the purposes of this study,

researchers dichotomized PCA3 using a pre-determined cut-off for labeling men as positive and negative for prostate cancer. The primary aim of the PCA3 validation study was to test the PPV corresponding to this cut-off. Alternately, researchers could have allowed PCA3 to remain a continuous biomarker and evaluated a point on the PPV curve with a pre-defined proportion of the population defined as positive and negative that corresponds to the cut-off used in the study. This would allow the performance of PCA3 to be compared to other markers on a common scale. We will consider PCA3 as a continuous biomarker for the purposes of this illustration.

Men in the initial biopsy population were considered positive if their PCA3 value was greater than 60, which corresponds to the top 25% of the population. This is equivalent to $PPV(0.75)$, the positive predictive value if the top 25% of the population are considered positive, and we will test the one-sided null hypothesis $H_0 : PPV(0.75) = 0.55$. The PCA3 validation study used a one-sided type-I error rate of 0.0125 in each of the two study populations to achieve an overall one-sided type-I error rate of 0.025. Assuming a prevalence of 0.45, we have approximately 92% power to detect an alternative hypothesis of $H_a : PPV(0.75) = 0.70$ with a one-sided type-I error rate of 0.0125 and a sample size of 562 men in the initial biopsy group.

The PCA3 validation study utilized a stopping rule for futility (described in Section 2), which would allow additional samples to be stored for use in future studies if the trial terminated early for futility. Using the results from Section 3, specifically, Corollary 3.3, we are able to apply standard methods from the group sequential clinical trials literature to develop formal stopping boundaries for the PCA3 validation study. A key advantage of this approach, as compared to the approach used when designing the PCA3 validation study, is that we are able to consider group sequential designs that allow early termination for both futility and superiority. This would allow investigators to conserve specimens for future studies in both the case where the performance of the candidate biomarker dramatically exceeds expectations, as well as the case where the biomarker does not appear promising at the interim analyses.

We will consider one-sided, symmetric group sequential stopping boundaries with shape parameters of 0 and 0.5 [15]. Critical values can be obtained using the `RCTdesign` package in R [16]. Typically, the maximum sample size must be increased when using group sequential stopping boundaries to offset the loss in power due to the possibility of early termination for futility and because of the increase in the final critical value that is required to control the overall type-I error rate for the study. Having already been completed, the maximum sample size for the PCA3 validation study is fixed and, instead, we note the power of the group sequential design with a maximum sample size of 562 and the sample size required to achieve the power obtained from a fixed-sample design with a sample size of 562 (i.e. 92% power).

Group sequential stopping boundaries and sequentially monitored test statistics for the PCA3 validation study can be found in Figure 1. We see that the symmetric boundaries with shape parameter equal to 0 have more extreme critical values early in the study but critical values close to those used in the fixed-sample design at the end of the study, whereas

symmetric boundaries with shape parameter equal to 0.5 have a constant upper boundary for rejecting the null hypothesis. As a result, early termination (for both superiority and futility) is much more likely using a shape parameter equal to 0.5. A drawback to using a shape parameter equal to 0.5 is that the power with a maximum sample size of 562 is only 83% (as compared to 92% for the fixed-sample design) and the maximum sample size would have to increase to 737 to retain the power of the fixed-sample design. In contrast, a shape parameter equal to 0 would have 91% power and the sample size would have to increase to 579 to retain the power of the fixed-sample design. In the case of the PCA3 validation study, the sequentially monitored test statistic would have exceeded the critical value for superiority at the first interim analysis ($N = 141$) regardless of the shape of the stopping boundaries. As a result, the PCA3 validation study could have stopped and rejected the null hypothesis after using only 141 samples, while conserving 421 samples for future studies of other candidate markers.

Estimates and confidence intervals for $PPV(0.75)$ after a group sequential test that terminates at the first interim analysis can be found in Table 5. Presented are the unadjusted estimator, the three adjusted estimators discussed in Section 4 (the Rao-Blackwell estimator, the mean-adjusted estimator and the median adjusted estimator), the unadjusted confidence interval and the adjusted confidence interval. The median-adjusted estimator and adjusted confidence interval again utilize the analysis-time ordering discussed in [21]. We observe a modest difference between the unadjusted and adjusted estimates and confidence intervals. Specifically, we observe no difference between the unadjusted estimator and the Rao-Blackwell and median-adjusted estimators (the Rao-Blackwell estimator will always equal the unadjusted estimator when the trial stops at the first interim analysis) and only a small difference between the mean-adjusted estimator when the shape parameter equals 0. The adjusted confidence intervals have the same upper-bound as the unadjusted confidence interval but the lower confidence limit of the adjusted interval is lower than the lower confidence limit of the unadjusted estimator for both designs.

6. Discussion

We investigated group sequential testing of the predictive accuracy of a continuous biomarker. Koopmeiners and Feng [8] previously developed asymptotic theory for the sequential empirical PPV and NPV curves but additional work was needed to translate their theoretical results into practice. First, we extended their asymptotic results to the important case where the prevalence must be estimated in the context of a cohort study. We then completed a simulation study to evaluate the performance of standard group sequential approaches to testing and estimation when applied to $PPV(u)$. Our simulation results illustrate that the application of group sequential approaches to testing $PPV(u)$ results in study designs with the correct type-I error rate and power but substantially smaller expected sample size than the corresponding fixed-sample design. In addition, our simulation results also illustrate that standard group sequential estimators can be used to obtain bias-adjusted estimates of $PPV(u)$. We concluded by illustrating the advantages of the group sequential approach in the context of a biomarker validation study for PCA3, a novel biomarker for prostate cancer.

There are many important, practical implications of this research. In a prospective cohort study to evaluate the predictive accuracy of a novel biomarker, group sequential methods allow us to potentially terminate the study early, which results in a substantial savings in both time and money compared to a longer, fixed-sample study. We might also evaluate the predictive accuracy of a novel biomarker using banked samples from a completed cohort study. One could apply a group sequential procedure to processing the samples where a subset of the samples are randomly selected and processed, the predictive accuracy of the biomarker is estimated and the decision as to whether or not more samples should be processed is determined by a group sequential stopping rule. This could potentially allow us to evaluate the novel biomarker using only a subset of the samples, leaving the remaining samples for evaluation of other candidate markers. The Early Detection Research Network has developed multiple specimen reference sets for the pre-validation of new biomarkers under these circumstances [22]. The methods proposed in this manuscript would allow these reference sets to be utilized in a more rigorous and efficient manner.

The asymptotic results presented in Section 3 were derived on the process level, rather than focusing only on a point on the PPV or NPV curves. An advantage to this approach is that we are easily able to derive asymptotic results for the joint distribution of multiple points on the PPV curve, corresponding to multiple possible marker cut-offs, as well as the joint behavior of a point on the PPV curve and a point on the NPV curve. The later is particularly important as researchers are often interested in the trade-off between PPV and NPV and may be interested in evaluating separate cut-offs for identifying subjects as negative and positive. We did not discuss this application in the current manuscript but additional discussion of this idea can be found in [8].

A limitation of the group sequential approach is that standard estimators will be biased at study completion. Several bias-adjusted estimators exist but there is no unique optimal solution to bias correction after a group sequential clinical trial. In our simulation results, the Rao-Blackwell estimator exhibited the best performance of the three estimators considered but bias-correction also resulted in increased variance compared to the standard estimator, in some cases. In practice, researchers must consider the trade-off between a potential increase in variance and the potential decrease in the sample size as a result of the group sequential approach. A formal evaluation of this trade-off and the resulting implications for the design of group sequential biomarker validation studies is the subject of continuing work by the first author.

Finally, our results are applicable to the case where biomarker values are available for all subjects in a cohort study. In many cases, candidate biomarkers are evaluated using stored specimens collected on a subset of patients in the context of a larger cohort study. In this case, case-controls status is available for all subjects but biomarker values are only available for a subset of subjects in the study. We are actively working on extending the results of this manuscript to group sequential testing of the PPV and NPV curve when biomarker values are collected on only a subset of subjects in a larger cohort study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by NIH grants P01-CA053996 and U24-CA086368.

References

1. Moskowitz CS, Pepe MS. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*. 2004; 5(1):113–127. [PubMed: 14744831]
2. Zheng Y, Cai T, Pepe MS, Levy WC. Time-Dependent Predictive Values of Prognostic Biomarkers With Failure Time Outcome. *Journal of the American Statistical Association*. 2008; 103(481):362–368. [PubMed: 19655041]
3. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *Journal of the National Cancer Institute*. 2008; 100(20):1432–1438. [PubMed: 18840817]
4. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. *Statistics in Medicine*. 2003; 22(5):727–739. [PubMed: 12587102]
5. Mazumdar M. Group sequential design for comparative diagnostic accuracy studies: Implications and guidelines for practitioners. *Medical Decision Making*. Sep-Oct;2004 24(5):525–533.10.1177/0272898X04269240 [PubMed: 15359002]
6. Tang L, Emerson SS, Zhou XH. Nonparametric and Semiparametric Group Sequential Methods for Comparing Accuracy of Diagnostic Tests. *Biometrics*. 2008; 64(4):1137–1145. [PubMed: 18371124]
7. Tang LL, Liu A. Sample size recalculation in sequential diagnostic trials. *Biostatistics*. 2010; 11(1): 151–163. [PubMed: 19822693]
8. Koopmeiners JS, Feng Z. Asymptotic properties of the sequential empirical ROC, PPV and NPV curves under case-control sampling. *Annals of Statistics*. 2011; 39(6):3234–3261. [PubMed: 24039313]
9. Deras IL, Aubin SM, Blase A, Day JR, Koo S, Partin AW, Ellis WJ, Marks LS, Fradet Y, Rittenhouse H, et al. Pca3: A molecular urine assay for predicting prostate biopsy outcome. *The Journal of Urology*. 2008; 179(4):1587–1592. [PubMed: 18295257]
10. Wei JT, Feng Z, Partin AW, Brown E, Thompson I, Sokoll L, Chan DW, Lotan Y, Kibel AS, Busby JE, et al. Can urinary pca3 supplement psa in the early detection of prostate cancer? *Journal of Clinical Oncology*. 2014; 32(36):4066–4072.10.1200/JCO.2013.52.8505 [PubMed: 25385735]
11. Zheng Y, Cai T, Stanford JL, Feng Z. Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics*. 2010; 66(1):50–60. [PubMed: 19397579]
12. Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*. 2012; 68(2):388–396. [PubMed: 22150576]
13. Csörgő, M.; Szyszkowicz, B. Order statistics: theory & methods, *Handbook of Statist.* Vol. 16. North-Holland; Amsterdam: 1998. Sequential quantile and Bahadur-Kiefer processes; p. 631–688.
14. Pyke R, Shorack G. Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems (Corr: 75AnlsProb 3 p1068). *The Annals of Mathematical Statistics*. 1968; 39:755–771.
15. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics*. 1989; 45(3):905–923. [PubMed: 2675998]
16. Emerson, SS.; Gillen, DL.; Kittelson, JK.; Emerson, SC.; Levin, GP. RCTdesign: Group Sequential Trial Design. 2012. R package version 1.0
17. Koopmeiners, JS.; Feng, Z. Working Paper 345. May 11. 2010 Asymptotic properties of the sequential empirical ROC and PPV curves.

18. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*. 1986; 73(3):573–581.10.1093/biomet/73.3.573
19. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika*. 1990; 77(4):875–892.10.1093/biomet/77.4.875
20. Whitehead, J. The design and analysis of sequential clinical trials. Ellis Horwood Ltd; 1983.
21. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics*. 1984; 40:797–803. [PubMed: 6518248]
22. Feng Z, Kagan J, Pepe M, Thornquist M, Ann Rinaudo J, Dahlgren J, Krueger K, Zheng Y, Patriotis C, Huang Y, et al. The early detection research network’s specimen reference sets: Paving the way for rapid evaluation of potential biomarkers. *Clinical Chemistry*. 2013; 59(1):68–74.10.1373/clinchem.2012.185140 [PubMed: 23193062]

Figure 1. Group sequential stopping boundaries (solid lines) and sequentially monitored test statistics (dashed lines) for the PCA3 validation study.

Table 1

Simulated type-I error rate and expected sample size ($E(n)$) for testing the one-sided null hypothesis $PPV(u)$ using a fixed-sample design and a group sequential design with 2, 3 and 4 interim analyses.

n	θ	u	$PPV(u)_0$	K = 1		K = 2		K = 3		K = 4	
				α	n	α	$E(n)$	α	$E(n)$	α	$E(n)$
250	0.3	0.7	0.6	0.04	250	0.04	182	0.05	171	0.05	158
			0.7	0.04	250	0.05	178	0.05	171	0.04	156
		0.8	0.6	0.05	250	0.05	195	0.05	163	0.04	154
			0.7	0.04	250	0.05	186	0.04	167	0.04	159
	0.4	0.7	0.6	0.04	250	0.03	182	0.04	172	0.04	158
			0.7	0.04	250	0.04	177	0.04	167	0.04	155
		0.8	0.6	0.06	250	0.05	197	0.05	161	0.05	154
			0.7	0.04	250	0.05	187	0.04	166	0.04	158
500	0.3	0.7	0.6	0.04	500	0.05	367	0.05	343	0.05	315
			0.7	0.05	500	0.05	371	0.05	341	0.04	321
		0.8	0.6	0.04	500	0.04	387	0.05	338	0.04	315
			0.7	0.05	500	0.05	384	0.04	337	0.05	313
	0.4	0.7	0.6	0.05	500	0.04	371	0.04	345	0.05	315
			0.7	0.05	500	0.05	373	0.05	342	0.04	322
		0.8	0.6	0.04	500	0.04	391	0.04	336	0.04	313
			0.7	0.05	500	0.05	387	0.05	337	0.05	313

Presented are results for one-sided symmetric boundaries with shape parameter equal to 0. 10,000 simulations were completed for each scenario. Note: The sample size is fixed for $K = 1$.

Table 2

Simulated power for testing the one-sided null hypothesis $PPV(u)$ under the alternative hypothesis $PPV(u)_a = PPV(u)_0 + 0.10$ using a fixed-sample design and a group sequential design with 2, 3 and 4 interim analyses.

		K = 1		K = 2		K = 3		K = 4	
1 - β	θ	u	$PPV(u)_0$	1 - β	n	1 - β	$E(n)$	1 - β	$E(n)$
0.8	0.3	0.7	0.6	0.79	554	0.78	482	0.77	457
								0.77	435
		0.8	0.7	0.77	541	0.76	466	0.77	444
								0.77	426
	0.4	0.7	0.6	0.77	769	0.8	661	0.78	636
								0.79	613
		0.8	0.7	0.79	682	0.79	598	0.78	573
								0.78	545
	0.8	0.7	0.6	0.79	500	0.81	433	0.77	416
								0.78	399
		0.8	0.7	0.78	445	0.78	399	0.79	374
								0.79	361
0.9	0.3	0.7	0.6	0.8	733	0.78	649	0.8	609
								0.79	591
		0.8	0.7	0.77	632	0.79	565	0.77	535
								0.8	524
	0.4	0.7	0.6	0.89	767	0.89	630	0.89	593
								0.89	570
		0.8	0.7	0.87	746	0.88	609	0.87	579
								0.88	557
	0.8	0.7	0.6	0.91	1056	0.89	879	0.9	810
								0.89	782
		0.8	0.7	0.9	927	0.89	772	0.9	729
								0.89	697
0.9	0.3	0.7	0.6	0.9	687	0.88	571	0.88	541
								0.89	511
		0.8	0.7	0.89	605	0.89	502	0.88	484
								0.88	463
	0.4	0.7	0.6	0.89	1004	0.9	825	0.9	796
								0.89	749
		0.8	0.7	0.9	856	0.9	703	0.9	676
								0.89	650

Presented are results for one-sided symmetric boundaries with shape parameter equal to 0. 10,000 simulations were completed for each scenario. Note: The sample size for $K = 1$ corresponds to the fixed sample size required to achieve the desired power for detecting a difference of 0.10 in $PPV(u)$.

Table 3

Simulated type-I error rate and power for testing the one-sided null hypothesis $PPV(u)$ and power under the alternative hypothesis $PPV(u)_a = PPV(u)_0 + 0.10$ using a fixed-sample design and a group sequential design with 2, 3 and 4 interim analyses when the prevalence is estimated from the data compared to when the prevalence is assumed equal to the design prevalence.

u	$PPV(u)$	Design θ	Type-I Error (n = 250)											
			K = 1			K = 2			K = 3			K = 4		
			true θ	Est. θ	Fix θ	Est. θ	Fix θ	Est. θ	Fix θ	Est. θ	Fix θ	Est. θ	Fix θ	
0.8	0.6	0.3	0.25	0.05	0.29	0.05	0.28	0.04	0.27	0.04	0.27	0.04	0.27	
			0.3	0.05	0.04	0.06	0.04	0.05	0.04	0.05	0.04	0.04	0.04	
			0.35	0.06	0	0.06	0	0.05	0	0.05	0	0.05	0	
		0.4	0.35	0.06	0.19	0.06	0.18	0.05	0.18	0.05	0.18	0.05	0.18	
			0.4	0.06	0.04	0.06	0.04	0.05	0.04	0.05	0.04	0.05	0.04	
			0.45	0.06	0.01	0.06	0.01	0.05	0	0.05	0	0.05	0.01	
0.7		0.3	0.25	0.05	0.3	0.05	0.29	0.05	0.28	0.04	0.28	0.04	0.28	
			0.3	0.05	0.04	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	
			0.35	0.04	0	0.04	0	0.04	0	0.04	0	0.04	0	
		0.4	0.35	0.04	0.17	0.04	0.17	0.04	0.17	0.04	0.17	0.04	0.16	
			0.4	0.04	0.04	0.04	0.03	0.05	0.04	0.05	0.04	0.04	0.04	
			0.45	0.04	0	0.04	0	0.04	0.01	0.04	0.01	0.04	0.01	
Power (80% power)														
0.8	0.6	0.3	0.25	0.78	1	0.79	1	0.79	1	0.79	1	0.79	1	
			0.3	0.78	0.9	0.79	0.9	0.79	0.91	0.79	0.91	0.79	0.91	
			0.35	0.79	0.28	0.77	0.28	0.8	0.28	0.79	0.28	0.79	0.28	
		0.4	0.35	0.8	1	0.77	1	0.8	1	0.78	1	0.78	1	
			0.4	0.79	0.87	0.78	0.87	0.8	0.88	0.77	0.87	0.77	0.87	
			0.45	0.78	0.39	0.77	0.39	0.79	0.38	0.79	0.38	0.79	0.38	
0.7		0.3	0.25	0.79	1	0.78	1	0.78	1	0.79	1	0.79	1	
			0.3	0.78	0.91	0.79	0.91	0.78	0.92	0.78	0.92	0.78	0.91	
			0.35	0.8	0.32	0.78	0.3	0.78	0.3	0.78	0.3	0.8	0.3	
		0.4	0.35	0.8	0.99	0.79	1	0.78	0.99	0.79	0.99	0.79	0.99	

<i>n</i>	<i>PPV (u)</i>	<i>Design θ</i>	<i>true θ</i>	K = 1		K = 2		K = 3		K = 4	
				<i>Est. θ</i>	<i>Fix θ</i>	<i>Est. θ</i>	<i>Fix θ</i>	<i>Est. θ</i>	<i>Fix θ</i>	<i>Est. θ</i>	<i>Fix θ</i>
			0.4	0.77	0.86	0.78	0.87	0.77	0.87	0.79	0.86
			0.45	0.78	0.42	0.8	0.42	0.78	0.42	0.79	0.43

Presented are results for one-sided symmetric boundaries with shape parameter equal to 0. 10,000 simulations were completed for each scenario.

Table 4

Simulation results evaluating the performance of the adjusted estimators and confidence intervals using a one-sided symmetric design with 2, 3 and 4 interim analyses and shape parameter equal to 0.

K	θ	u	$PPV(u)$	$\widehat{PPV}(u)$	$\widehat{PPV}(u)_{rb}$	$\widehat{PPV}(u)_{mn}$	$\widehat{PPV}(u)_{md}$	Unadj Cov	Adj Cov
2	0.4	0.7	0.6	0.59 (0.07)	0.59 (0.08)	0.59 (0.08)	0.59 (0.08)	0.95	0.95
			0.65	0.63 (0.07)	0.64 (0.08)	0.64 (0.07)	0.64 (0.07)	0.94	0.95
			0.7	0.68 (0.07)	0.7 (0.08)	0.69 (0.07)	0.68 (0.07)	0.94	0.95
			0.75	0.73 (0.07)	0.75 (0.07)	0.74 (0.07)	0.74 (0.07)	0.92	0.96
	0.4	0.8	0.8	0.79 (0.07)	0.79 (0.06)	0.79 (0.06)	0.79 (0.07)	0.93	0.96
			0.6	0.59 (0.08)	0.6 (0.1)	0.6 (0.09)	0.59 (0.09)	0.94	0.96
			0.65	0.63 (0.08)	0.65 (0.1)	0.64 (0.08)	0.64 (0.09)	0.95	0.96
			0.7	0.68 (0.08)	0.7 (0.09)	0.69 (0.08)	0.69 (0.08)	0.95	0.96
	3	0.4	0.75	0.73 (0.08)	0.75 (0.08)	0.74 (0.08)	0.74 (0.08)	0.93	0.97
			0.8	0.79 (0.08)	0.8 (0.07)	0.8 (0.07)	0.79 (0.08)	0.92	0.96
			0.6	0.57 (0.08)	0.6 (0.1)	0.59 (0.09)	0.58 (0.09)	0.94	0.95
			0.65	0.62 (0.08)	0.65 (0.09)	0.64 (0.08)	0.63 (0.08)	0.95	0.95
4	0.4	0.7	0.7	0.68 (0.08)	0.7 (0.08)	0.69 (0.08)	0.68 (0.08)	0.93	0.95
			0.75	0.73 (0.08)	0.75 (0.07)	0.74 (0.07)	0.74 (0.08)	0.93	0.95
			0.8	0.8 (0.07)	0.8 (0.07)	0.8 (0.07)	0.79 (0.07)	0.9	0.94
			0.6	0.56 (0.11)	0.59 (0.13)	0.58 (0.11)	0.57 (0.11)	0.93	0.95
	0.8	0.7	0.65	0.61 (0.1)	0.64 (0.11)	0.63 (0.1)	0.62 (0.11)	0.95	0.95
			0.7	0.67 (0.1)	0.69 (0.1)	0.68 (0.1)	0.67 (0.1)	0.94	0.95
			0.75	0.72 (0.1)	0.74 (0.09)	0.74 (0.09)	0.73 (0.09)	0.92	0.96
			0.8	0.78 (0.09)	0.79 (0.08)	0.79 (0.08)	0.79 (0.08)	0.92	0.95
	4	0.4	0.6	0.56 (0.1)	0.59 (0.11)	0.58 (0.1)	0.57 (0.1)	0.92	0.95
			0.65	0.62 (0.09)	0.64 (0.09)	0.64 (0.09)	0.63 (0.09)	0.93	0.96
			0.7	0.67 (0.09)	0.69 (0.08)	0.69 (0.08)	0.68 (0.09)	0.92	0.95
			0.75	0.73 (0.08)	0.74 (0.08)	0.74 (0.07)	0.73 (0.08)	0.91	0.96
4	0.8	0.7	0.8	0.79 (0.08)	0.79 (0.07)	0.79 (0.07)	0.79 (0.07)	0.94	0.96
			0.6	0.57 (0.12)	0.6 (0.12)	0.59 (0.12)	0.58 (0.12)	0.92	0.96
			0.65	0.62 (0.11)	0.65 (0.11)	0.64 (0.1)	0.63 (0.11)	0.95	0.96

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$\mathbf{K}$	θ	n	$\widehat{PPV}(u)$	$\widehat{PPV}(u)$	$\widehat{PPV}(u)_{rb}$	$\widehat{PPV}(u)_{mn}$	$\widehat{PPV}(u)_{md}$	Unadj Cov	Adj Cov
0.7	0.75	0.7	0.67 (0.1)	0.69 (0.1)	0.69 (0.1)	0.69 (0.09)	0.68 (0.1)	0.91	0.95
		0.75	0.72 (0.1)	0.74 (0.09)	0.74 (0.09)	0.74 (0.09)	0.73 (0.09)	0.92	0.96
		0.8	0.79 (0.09)	0.8 (0.08)	0.79 (0.08)	0.79 (0.08)	0.79 (0.09)	0.92	0.96

Presented are the simulated mean(SE) for the unadjusted estimator ($\widehat{PPV}(u)_{rb}$), the Rao-Blackwell estimator ($\widehat{PPV}(u)_{mn}$) and the median adjusted estimator ($\widehat{PPV}(u)_{md}$), as well as the coverage probability for the unadjusted and adjusted 95% confidence interval. The analysis-time ordering was used for the median-adjusted estimator and adjusted confidence interval. 10,000 simulations were completed for each scenario.

Table 5

Unadjusted and bias-adjusted estimates of $PPV(u)$ after a one-sided symmetric test with 4 interim analyses and shape parameter equal to either 0 or 0.5 for the PCA3 validation study.

Shape Parameter	N at Term	$\widehat{PPV}(0.75)$	$\widehat{PPV}(0.75)_{rb}$	$\widehat{PPV}(0.75)_{mn}$	$\widehat{PPV}(0.75)_{md}$	Unadj 95% CI	Adj 95% CI
0	141	0.936	0.936	0.923	0.936	(0.806, 1.0)	(0.789, 1.0)
0.5	141	0.936	0.936	0.936	0.936	(0.806, 1.0)	(0.789, 1.0)