NIH Public Access
Author Manuscript
Stat Med. Author manuscript; available in PMC 2014 March 15.

Published in final edited form as:
Stat Med. 2013 March 15; 32(6): 1027–1037. doi:10.1002/sim.5622.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Early Termination of a Two-Stage Study to Develop and Validate a Panel of Biomarkers

Joseph S. Koopmeiners[a,b,*] and **Rachel Isaksson Vogel**[b]

[a]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A.

[b]Biostatistics and Bioinformatics Core, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Abstract

Two-stage designs to develop and validate a panel of biomarkers present a natural setting for the inclusion of stopping rules for futility in the event of poor preliminary estimates of performance. We consider the design of a two-stage study to develop and validate a panel of biomarkers where a predictive model is developed using a subset of the samples in stage 1 and the model is validated using the remainder of the samples in stage 2. First, we illustrate how a stopping rule for futility can be implemented in a standard, two-stage study for developing and validating a predictive model where samples are separated into a training and validation sample. Simulation results indicate that our design has similar type-I error rate and power to the fixed-sample design but with a substantially reduced sample size under the null hypothesis. We then illustrate how additional interim analyses can be included in stage 2 by applying existing group sequential methodology, which results in even greater savings in the number of samples required under both the null and alternative. Our simulation results also illustrate that the operating characteristics of our design are robust to changes in the underlying marker distribution.

Keywords

Group Sequential Design; Biomarker Panel; ROC Curve

1. Introduction

The scientific community has expended substantial resources over the last ten years to identify biomarkers for cancer diagnosis and prognosis. This has resulted in a large number of candidate biomarkers whose performance needs to be validated. Study design for the evaluation of a single candidate marker is well studied [1] but, in most cases, a single biomarker will not have adequate performance and a combination of biomarkers is needed to achieve performance that is useful clinically. Innovative study designs are needed to evaluate the performance of multiple markers in combination while making efficient use of the available resources.

Regardless of the method used to develop a predictive model using multiple biomarkers, it is generally accepted that the performance of any predictive model should be developed and validated in a two-stage process where the model is built in stage one and validated with an

Copyright © 0000 John Wiley & Sons, Ltd

[*]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.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

independent set of data in stage two. Simple formulas for calculating the sample size required to obtain the desired type-I error rate and power for two-stage studies to develop and evaluate the diagnostic accuracy of a panel of biomarkers are not available.

Cancer biomarkers are often evaluated using banks of stored tissue samples. The development of banks of high quality tissue samples is an expensive and time consuming process and care should be taken to preserve these scarce resources. Group sequential designs have been suggested as an approach to conserving specimens when validating biomarkers for classification or prediction [2]. In particular, two-stage designs to develop and validate a predictive model using several biomarkers are a natural setting for the inclusion of stopping rules for futility after stage one in the event of poor preliminary estimates of performance.

Early termination for futility in a two-stage study to develop and validate a predictive model would have similar statistical implications to a group sequential clinical trial but different practical implications. In a group sequential clinical trial, subjects are enrolled sequentially and the the interim analyses are used to determine if additional subjects should be enrolled in the trial. Diagnostic biomarkers are typically evaluated retrospectively using banks of stored tissue samples and the sequential aspect refers to processing the samples. In a two-stage study to develop and validate a biomarker panel, we would first assay a randomly selected subset of the stored tissue samples, develop a predictive model and evaluate our stopping rule. The remaining tissue samples would only be assayed if we do not terminate the study at the interim analysis. Early termination would allow us to save time and money by not unnecessarily processing the entire set of tissue samples and conserve the unused samples for future studies to evaluate different candidate markers.

Group sequential designs for evaluating the performance of a single marker have been discussed in the literature [3, 4, 5]. These designs could be used to evaluate the performance of a predictive model developed in previous studies but are inappropriate for the setting where development of a predictive model is of interest.

We consider early termination of a two-stage study to develop and validate a panel of biomarkers for predicting prostate cancer. First, we illustrate how a stopping rule for futility can be implemented in a standard, two-stage study for developing and validating a predictive model where samples are separated into a training and validation sample. Simulation results indicate that our design has similar type-I error rate and power to the fixed-sample design but with a substantially reduced sample size under the null hypothesis. We then illustrate how additional interim analyses can be included in stage 2 by applying existing group sequential methodology, which results in even greater savings in the number of samples required under both the null and alternative. Our simulation results also illustrate that the operating characteristics of our study are robust to changes in the underlying marker distribution.

The remainder of the paper proceeds as follows. In Section 2, we describe a two-stage study to develop and validate a panel of biomarkers that allows early termination for futility. A detailed description of a simulation study to evaluate the operating characteristics of our study can be found in Section 3 and results are presented in Section 4. We discuss how additional interim analyses can be incorporated into stage 2 in Section 5 and discuss implementation of our proposed design in Section 6. Finally, we conclude with a brief discussion in Section 7.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

2. Study Design

We present our results in the context of a study that is to be completed at the University of Minnesota to develop and validate a panel of biomarker for predicting prostate cancer recurrence. Fifteen candidate biomarkers have been identified as potential predictors of prostate cancer recurrence in previous studies and our interest lies in their performance in combination. Our study will make use of 720 stored tissue samples. We will develop a predictive model for prostate cancer using a two-stage study design where a predictive model is developed in stage 1 using a subset of the 720 samples and validated with the remaining samples in stage 2. A design that allows early termination for futility after stage 1 would allow the remaining tissue samples to be used in future studies should initial estimates of prognostic accuracy for the marker panel be inadequate.

The goal of this study is to develop a predictive model for biochemical failure. Biochemical failure for prostate cancer is defined as rising prostate specific antigen (PSA) serum levels after post-treatment low nadir. The prognostic accuracy of our model will be evaluated by the time-dependent ROC curve for 5-year biochemical failure estimated using the method proposed by Heagerty et al. [6]. This method considers men that have experienced biochemical failure within 5 years to be cases, men that have not experienced biochemical failure within 5 years to be controls and accounts for the possibility of censored observations. For the purposes of designing our study, our primary measure of prognostic accuracy will be a point, ROC(0.1), on the ROC curve for 5-year biochemical failure. ROC(0.1) is the sensitivity corresponding to 90% specificity. The null and alternative hypotheses for our study are:

$$ROC(0.1)_0=0.40 \quad \text{and} \quad ROC(0.1)_a=0.65,$$

respectively. The null and alternative hypotheses were chosen to achieve a clinically meaningful improvement in the positive and negative predictive values (PPV and NPV, respectively) for 5-year biochemical failure. Assuming that 15% of patients will experience biochemical failure within five years, the NPV and PPV under the null hypothesis will be 90% and 41%, respectively, and the NPV and PPV will be 94% and 53%, respectively, under the alternative hypothesis. While this may seem like only a modest improvement in PPV and NPV, it was felt that the PPV and NPV corresponding to the alternative hypothesis would represent a clear improvement over simply knowing the prevalence, while the PPV and NPV corresponding to the null hypothesis would not.

In stage 1, a predictive model for biochemical failure will be developed using a random subset (i.e. training subset) of P% of the total samples. A predictive model for biochemical failure will be developed from the training subset using Cox proportional hazards regression and the Lasso [7, 8]. ROC(0.1) for the predictive model will be estimated using the training subset and the study will terminate if the estimate of ROC(0.1) is less than a pre-specified cut-off, $ROC(0.1)_{co}$. We expect that ROC(0.1) estimated from the training subset will represent an optimistic estimate of performance and if the marker panel does not appear promising with an optimistic estimate of performance then there is no need to validate with an independent sample in stage 2.

Our choice of futility stopping rule deserves further discussion. Several methods exist for defining stopping rules for futility (group sequential designs, conditional power, etc.) but we feel that they are inappropriate in this setting for two reasons. First, a hypothesis test for $ROC(0.1)$ that does not account for model selection will not have the correct type-I error rate and developing a hypothesis test that does account for model selection is difficult. There is

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

no guarantee that existing methods for developing futility boundaries would perform as expected given the difficulties associated with inference on *ROC*(0.1) after stage 1. Second, the majority of group sequential methods rely on the independent increments assumptions [9]. As we will see in Section 5, the independent increments assumption may be reasonable when estimating the survival ROC curve for a single marker but the information growth across stages 1 and 2 is not well understood given model selection in stage 1. Intuitively, we expect that the data from stage 2 will provide more information about the true classification accuracy of our predictive model because these data are independent of the data used to build the model. In this case, the validity of the independent increments assumption is certainly in question. For these reasons, we take the simple approach in defining a futility stopping rule and will use simulation to evaluate the effect of varying $ROC(0.1)_{co}$ and P on the operating characteristics of our study.

If the cut-off for futility is exceeded in stage 1, the predictive model will be validated using the remaining (1 -P)% of samples in stage 2. The remaining samples used in stage 2 are independent of the training subset and estimates of prognostic accuracy in stage 2 represent an unbiased estimate of the performance of the predictive model. We will test our null hypothesis using the test statistic,

$$Z_{ROC(0.1)}=\frac{R\hat{O}C_{stg2}(0.1)-ROC(0.1)_0}{SE_{R\hat{O}C_{stg2}(0.1)}}, \quad (1)$$

where $SE_{R\hat{O}C_{stg2}(0.1)}$ is estimated by bootstrap and we will reject the null hypothesis if $Z_{ROC(0.1)}$ is greater than the .975th quantile of the standard normal distribution.

In summary, our design will proceed as follows:

1. Split data into training and validation set
2. Build a predictive model for biochemical failure using the Lasso on the training set
3. Estimate *ROC* (0.1) for this model from the training set
4. Terminate for futility if estimated *ROC* (0.1) < $ROC(0.1)_{co}$, otherwise, continue to step 5
5. Estimate *ROC* (0.1) and test the null hypothesis using the validation set

3. Simulation Study

Analytical methods exist for determining design parameters that achieve the desired operating characteristics for simple study designs but simulation is needed to evaluate the operating characteristics of a study in more complicated settings. In this section, we describe an extensive simulation study to evaluate the operating characteristics of our study.

Let X be a 15 dimensional vector of biomarkers, which is composed of n_s signal markers, X_s, (i.e. biomarkers that are truly associated with PC recurrence) and $15 - n_s$ noisy markers, X_n (i.e. biomarkers that are not associated with PC recurrence). That is, X is composed of two sub-vectors,

$$X=(X_s, X_n),$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where X_s is a n_s dimensional sub-vector of signal markers and X_n is a $(15 - n_s)$ dimensional sub-vector of noisy markers. Signal markers were drawn from a multivariate normal distribution with mean $\vec{0}$ and covariance matrix Σ_s, while noisy markers were drawn from i.i.d. standard normal distributions. We note that $N(0, \Sigma_s)$ represents the marginal distribution of the signal markers and that the mean of this distribution does not impact the quality of X_s as a classifier for recurrence. For this reason, we set the mean equal to $\vec{0}$ for simplicity. Instead, the ability of X_s to discriminate between subjects with long and short failure times is controlled by Σ_s and a regression model, which we will now describe. Biochemical failure times, Y, were sampled from an exponential distribution with rate parameter, λ, where λ is a function of the signal markers,

$$\log\lambda=\beta_0+X_s^t\beta_{1,s},$$

with $\beta_{1,s} = \beta_s W_s$, where β_s is a scalar and W_s is a vector of weights indicating the relative importance of each marker. At this point, we should clarify that our assumption of $\beta_{1,s} = \beta_s W_s$, where β_s is a scalar will only be made for the purposes of our simulation study and that this restriction will not be imposed when estimating the predictive model in stage 1 of our study. There are infinitely many vectors, $\beta_{1,s}$ that result in a given value of $ROC(t)$ and we place a restriction on the form of $\beta_{1,s}$ in order to investigate the impact of specific patterns in the regression parameters on the operating characteristics of our study. Finally, censoring times, C, were drawn from a uniform distribution with range $(0, C_{max})$.

The following parameters must be specified: n_s, W_s, Σ_s, β_0, β_s and C_{max}. The first three (n_s, W_s and Σ_s) were varied to determine their impact on the operating characteristics of the study. Σ_s is a function of the variances of the signal markers, V_s, and the correlation between markers, ρ, and was varied by considering different combinations of V_s and ρ. For our simulations, we considered the following possible values of n_s, W_s, V_s and ρ:

- n_s: 3,5 and 7
- W_s: $(1, \ldots, 1)$, $\left(\frac{n_s}{n_s},\frac{n_s-1}{n_s},\ldots,\frac{1}{n_s}\right)$
- V_s: $(1, \ldots, 1)$, $\left(\frac{n_s}{n_s},\frac{n_s-1}{n_s},\ldots,\frac{1}{n_s}\right)$ and $\left(\frac{1}{n_s},\frac{2}{n_s},\ldots,\frac{n_s}{n_s}\right)$
- ρ: 0, 0.2, 0.4 and 0.6

The remaining parameters were set as follows: β_0 and β_s were fixed to achieve the desired prevalence (5-year PC recurrence rate equal to 0.15) and true value of $ROC(0.1)$ (either $ROC(0.1)_0$ or $ROC(0.1)_1$) by solving the set of equations found in the Supplementary Materials and C_{max} was set equal to 50 years because we expect 5-year follow-up to be available for 90% of subjects.

In addition to the biomarker parameters listed above, we will also consider the effect of varying the proportion of samples used in stage 1, P, and the cut-off for continuing to stage 2, $ROC(0.1)_{co}$, on the operating characteristics of our study. Values for P and $ROC(0.1)$ considered in our simulation include:

- P: 0.3, 0.4, 0.5, 0.6 and 0.7
- $ROC(0.1)_{co}$: 0 (fixed-sample design), 0.40, 0.45 and 0.50

We will evaluate the standard operating characteristics of type-I error and power. Given that we are considering a group sequential design, we will also consider the expected sample size

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of our study under the null and alternative hypotheses. At this point, we only allow early termination for futility and hope to observe a substantial reduction in the expected sample size under the null hypothesis compared to the fixed-sample design but expect that the expected sample size under the alternative will be close to that of the fixed-sample design.

Finally, it is important to note that the hypothesis test after stage 2 is *not* testing whether *ROC* (0.1) for $X_s^t\beta_{1,s}$ is greater than the null hypothesis but is instead testing whether *ROC* (0.1) for $X^t\hat{\beta}_1$ is greater than the null hypothesis where $\hat{\beta}_1$ is the estimated vector of regression coefficients for all markers estimated in stage 1. It is likely that the classification accuracy of $X^t\hat{\beta}_1$ will be worse than the classification accuracy of $X_s^t\beta_{1,s}$ if for no other reason than $\hat{\beta}_1$ is likely to include a non-zero coefficient for one of the noisy markers. We expect that $\hat{\beta}_1$ will approach $(\beta_{1,s}, \vec{0})$ as the sample size in stage 1 increases and anticipate that the classification accuracy of $X^t\hat{\beta}_1$ will improve as the proportion of samples used in stage 1 increases. For this reason, we must also consider the effect of varying P and $ROC(0.1)_{co}$ on the quality of the predictive model developed in stage one in addition to the type-I error rate and power when testing the classification accuracy of our model in stage two. We are able to calculate the true value of *ROC* (0.1) for $X^t\hat{\beta}_1$ from each simulated study using the distributional assumptions described above and this value will be reported as an operating characteristic of our study in addition to the standard operating characteristics of type-I error, power and expected sample size.

4. Simulation Results

4.1. Design Parameters

The parameters varied in our simulation can be separated into two groups: the design parameters (P and $ROC\ (0.1)_{co}$), which we can control when designing our study, and the biomarker parameters (n_s, W_s, V_s and ρ), which are inherent characteristics of the biomarkers and not within our control. Ideally, varying the biomarker parameters while holding the true value of *ROC* (0.1) constant would have limited impact on the operating characteristics of our study, leaving us only to worry about the design parameters. We begin by considering the impact of the design parameters on the operating characteristics of our study.

Table 1 presents simulation results evaluating the effect of the proportion of subjects in stage 1 and the cut-off for early termination on the operating characteristics of the study. We hope to find a combination that minimizes the expected sample size under the null hypothesis but provides adequate power under the alternative. A higher proportion of subjects in stage 1 is expected to lead to a better predictive model and increase the likelihood of early termination, while a higher proportion of subjects in stage 2 will increase power. Similarly, a higher threshold for continuing to stage 2 will limit the expected sample size under the null but increase the chance of early termination under the alternative and decrease power.

We first consider the impact of P on the operating characteristics of our study in the fixed-sample design as this serves as a baseline for which to compare the designs that allow early termination for futility. Increasing P will result in a larger sample size for building our predictive model in stage 1 but a smaller sample size for testing the classification accuracy of our model in stage 2. We see that the power decreases dramatically as P increases. In some sense, this should not be surprising. As P increases, the sample size available for stage 2 decreases, which results in a smaller sample size available for testing our predictive model. For example, with $P = 0.30$, there are 504 samples available for testing our model in stage 2 but with $P = 0.70$, there are only 216 samples available in stage 2, resulting in a substantial reduction in power. What is somewhat surprising, though, is that the power

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

decreases even though the true classification accuracy of the predictive model developed in stage 1 increases. This implies that the marginal improvement in the classification accuracy of our fitted model is not worth the decrease in power due to an inadequate sample size for validation in stage 2. The same phenomenon is observed for the three scenarios where early termination is allowed. Increasing P when early termination was allowed resulted in an increased probability of early termination but also resulted in an increased expected sample size under the null. This may seem counter-intuitive but it simply reflects the fact that the increased probability of early termination did not offset the increased minimum sample size as the proportion of samples used in stage 1 increased.

The cut-off for early termination had the expected effect on the sample size, type-I error rate and power. Increasing the cut-off for early termination resulted in an increased probability of early termination under the null hypothesis which decreased the expected sample size. The cut-off for early termination did not affect the type-I error rate conditional on reaching stage 2 but the unconditional type-I error rate decreased as $ROC(0.1)_{co}$ increased due to the increased probability of early termination. The probability of early termination under the alternative was small when $ROC(0.1)_{co}$ equaled 0.40 and 0.45, resulting in power similar to the fixed-sample design, but increasing $ROC(0.1)_{co}$ to 0.50 resulted in approximately a 5% decrease in power when P equaled 0.30 and 0.40 due to an increase in the probability of early termination for futility. Somewhat surprisingly, $ROC(0.1)_{co}$ had little impact on the true value of ROC (0.1) for $X_t\hat{\beta_1}$. We expected that the true value of ROC (0.1) would increase with ROC $(0.1)_{co}$ because only the best models would proceed to stage 2. Instead, we observed that ROC $(0.1)_{co}$ had little impact on the true value of ROC (0.1), which suggests that variability in the estimate of ROC (0.1) after stage 1 is primarily due to variability in the estimation procedure and not variability in the true classification accuracy of $X_t\hat{\beta_1}$.

Going forward, we will consider a design with $P = 0.30$ and $ROC(0.1)_{co} = 0.45$. This design was chosen because it provides a substantial reduction in the sample size under the null hypothesis, while providing similar power to the fixed-sample design.

4.2. Biomarker Parameters

We next consider the effect of varying the biomarker parameters on the operating characteristics of our study. This can be thought of as a sensitivity analysis. We have no control over these parameters and it would be reassuring if the operating characteristics of our study were robust to variation in these parameters while holding the true value of ROC (0.1) constant.

Table 2 presents simulation results evaluating the effect of various biomarker parameters on the operating characteristics of our study. Increasing the number of signal markers (n_s) resulted in a slight increase in the expected sample size under the null (n_s = 3, average E(SS) = 383, n_s = 5, average E(SS) = 389, n_s = 7, average E(SS) = 394) and a slight decrease in power (n_s = 3, average power = 0.91, n_s = 5, average power = 0.90, n_s = 7, average power = 0.89). Increasing the correlation between signal markers had the opposite effect, decreasing the expected sample size under the null (ρ = 0.0, average E(SS) = 393, ρ = 0.6, average E(SS) = 385) and increasing power (ρ = 0.0, average power = 0.89, ρ = 0.6, average power = 0.91). Varying W_s and V_s did not appear to have a systematic effect on the operating characteristics of our study.

Overall, though, the results observed in Table 2 are reassuring. The average sample size under the null hypothesis was 388 with a range of 376 - 407 and the average power was 0.90 with a range of 0.88 - 0.92. These differences are modest compared to the differences observed in Table 1, which suggests that the operating characteristics of our study are robust

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

to changes in the underlying marker distribution as long as the performance of the markers in combination remains constant.

5. Additional Interim Analyses During Stage 2

To this point, we have only considered one interim analysis for futility at the end of stage 1. Additional savings in the number of samples could be achieved by including additional interim analyses during stage 2. Furthermore, while we are unwilling to terminate the study for superiority after stage 1 because of a desire to validate our predictive model on an independent set of data, we have no such concerns during stage 2 and can implement stopping rules that allow early termination for futility and superiority during stage 2.

Once a predictive model is developed in stage 1, stage 2 is simply a fixed-sample study to evaluate the classification accuracy of a single marker (that is a linear combination of several markers). Additional interim analyses can be incorporated into stage 2 by applying existing group sequential methodology (i.e. O'Brien-Fleming or Pocock boundaries [10,11] or error spending functions [12]). Standard group sequential methodology relies on the existence of a test statistic with an independent increments covariance structure [9]. The independent increments assumption has been verified for the standard ROC curve [5, 13] but not for the survival ROC curve. The independent increments assumption holds in a wide variety of situations and, while it has not been verified theoretically for the survival ROC curve, we can evaluate this assumption through simulation.

Adding additional interim analyses to stage 2 does not substantially change the character of our study. In stage 1, we develop a predictive model using the Lasso for Cox regression and estimate $ROC\ (0.1)$ for the predictive model. We compare our estimate of $ROC\ (0.1)$ to $ROC\ (0.1)_{co}$ and terminate for futility if our estimate is less than $ROC\ (0.1)_{co}$. If our estimate exceeds $ROC\ (0.1)_{co}$, we validate our predictive model in stage 2 but instead of a single hypothesis test at the end of stage 2, we monitor $ROC\ (0.1)$ sequentially using the test statistic from (1). For our simulations, we consider stopping boundaries using the error spending functions proposed by Hwang et al. [14], which allow early termination for superiority and futility. Considering only the studies that proceed to stage 2, implementing group sequential stopping boundaries during stage 2 should have no effect on the type-I error rate but will decrease the power (usually, sample size is increased to accommodate interim analyses in a group sequential study but our sample size is fixed). The probability of continuing to stage 2 will not be effected by the inclusion of additional interim analyses during stage 2, implying that the additional interim analyses will have no effect on the overall type-I error rate but will decrease the overall power.

Table 3 presents simulation results evaluating the operating characteristics of our study when additional interim analyses are included in stage 2. Simulations were completed with $P = 0.3$, $ROC\ (0.1)_{co} = 0.45$, $n_s = 5$, $W_s = (1, \ldots, 1)$, $V_s = (1, \ldots, 1)$ and $\rho = 0.0$. We see that the conditional type-I error rate is similar regardless of the number of interim analyses. While not definitive, this does suggest that the independent increments assumption holds for the survival ROC curve. Incorporating additional interim analyses into stage 2 results in a decrease in the expected sample size under both the null and alternative hypothesis. Under the null hypothesis, the expected sample size decreased from 397 when one, single analysis is completed at the end of stage 2 to 302 when four stopping times are considered. An even larger decrease is observed under the alternative, where the expected sample size decreases from 712 with one stopping time to 551 when four stopping times are considered. We observe a decrease in power as additional interim analyses are included, as expected, but the difference is modest when compared to the reduction in the expected sample size. Finally, simulations evaluating the effect of varying the biomarker parameters n_s, W_s, V_s and ρ can

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

be found in the Supplementary Materials and illustrate that the operating characteristics of our study are robust to changes in the biomarker parameters as long as the true value of *ROC* (0.1) is held constant.

6. Implementation

In this section, we briefly discuss the implementation of our proposed method in the context of nine simulated scenarios. Our study proceeds as described in Sections 2 and 5. First, we split the data into a training and a validation sample. We develop a predictive model for prostate cancer recurrence using the Lasso for Cox regression and estimate *ROC*(0.1) for the predictive model using the training set. The study terminates for futility if the estimated value of *ROC* (0.1) from the training set is less than $ROC\ (0.1)_{co}$. Otherwise, we validate our predictive model in the testing sample using group sequential stopping boundaries and the test statistic described in (1).

In order to proceed, we must specify the following study parameters: n, $ROC\ (0.1)_0$, P, $ROC\ (0.1)_{co}$ and the group sequential stopping boundaries used in stage 2. In our motivating example, the total sample size, n, is fixed at 720 due to the number of stored tissue samples available. We design our study to test the null hypothesis, $ROC\ (0.1)_0 = 0.40$, as described in Section 2. We set $P = 0.30$, which implies that 216 samples will be used in stage 1, and terminate for futility if the estimated *ROC* (0.1) from the training set is less than $ROC\ (0.1)_{co} = 0.45$. Finally, we will test the classification accuracy of our predictive model in stage 2 using the stopping boundaries developed by Hwang et al. [14] with three interim analysis and a one-sided type-I error rate of 0.025. This results in the following stopping boundaries:

- Superiority: 3.15, 2.82, 2.44, 1.96
- Inferiority: -0.65, 0.33, 1.17, 1.96

Our study will stop early for superiority if our test statistic exceeds the superiority stopping boundaries and terminate for futility if our test statistic is less than the inferiority boundaries.

Table 4 presents the estimated *ROC* (0.1) after stage 1($\hat{ROC}\ (0.1)_{stg1}$), the sequential test statistic ($Z_{ROC\ (0.1),1}$ - $Z_{ROC\ (0\ 1),4}$) and the total sample size for the nine simulated scenarios. The study terminates after stage 1 in scenarios 1, 2 and 4 because $\hat{ROC}\ (0.1)_{stg1}$ is less than $ROC\ (0.1)_{co}$, which results in only 216 of the 720 total samples being used. Five of the remaining six scenarios terminate early in stage 2. Scenarios 3, 5, 6 and 7 terminate early for futility, while Scenarios 8 and 9 terminate early for superiority. The total sample size in these five scenarios ranged from 342 samples in Scenario 3 and 594 samples in Scenario 6. This illustrates the strength of the proposed design. High quality tissue samples are a scarce commodity and early termination allows unused samples to remain available for future studies.

7. Discussion

We discussed early termination in a two-stage study to develop and validate a panel of biomarkers. First, we illustrated how a stopping rule for futility can be implemented in a two-stage study where samples are separated into a training set and a validation set. Simulation results showed that this design had a smaller type-I error rate and similar power to the fixed-sample design, while decreasing the expected sample size under the null hypothesis. We then discussed how an even greater reduction in sample size can be achieved by implementing additional interim analyses during stage 2. The inclusion of additional interim analyses during stage 2 had little impact on the type-I error rate and power but reduced the expected sample size under both the null and alternative hypothesis.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

These results have important implications for the design of diagnostic biomarker studies. One of the biggest challenges to evaluating a panel of biomarkers is the scarcity of high quality tissue samples. In response, extensive resources have been used to develop banks of high quality tissue samples for evaluating biomarkers for diagnosis and prognosis. These are scarce resources and it is important to use them responsibly. Group sequential designs are an obvious approach to conserving these resources while still adequately evaluating candidate biomarkers.

Our simulation results also showed that varying characteristics of the biomarkers while holding their combined classification accuracy constant had little effect on the operating characteristics of our study. Knowing that the marker parameters have little effect on the operating characteristics of our study allows us to design our study without worrying about characteristics of the biomarkers that we can't control. Furthermore, future attempts to develop analytic methods for designing similar studies would be greatly simplified if characteristics of the underlying marker distribution are not of great concern.

An important difference between diagnostic biomarker studies and therapeutic clinical trials is that diagnostic biomarker studies are generally completed retrospectively, while clinical trials are prospective. The sequential aspect of a group sequential diagnostic biomarker study refers to the processing of the samples and not their collection. A retrospective study using stored tissue samples that is terminated early will save money by not processing the samples and conserve specimens for future studies of other candidate biomarkers. In this sense, group sequential methods are beneficial even if all samples have already been collected.

Some of the features of this design (the stopping rule at the end of stage 1, in particular) are somewhat ad hoc. Nevertheless, our simulation results indicate that our design has similar type-I error and power to the fixed-sample design but requires a substantially smaller sample size, on average. A rigorous theoretical justification and analytical methods are needed in order for group sequential designs for developing and validating a panel of biomarkers to achieve more widespread use. Our results, in particular, simulation results indicating the limited influence of the underlying marker parameters on the operating characteristics of our study, provide motivation for future theoretical work. Utilizing group sequential designs to develop and validate a panel of biomarkers will impact our estimates of marker performance at study completion. Estimation after a group sequential clinical trial has been studied extensively in the literature [15, 16, 17]. The majority of these estimators focused on unconditional estimation, where the goal is an estimator that has good statistical properties when averaged over stopping times, while a subset were developed to achieve good statistical properties conditional on the observed stopping time. In a fixed-sample design, we estimate the performance of a panel of biomarkes conditional on the model developed in the training set. In our setting, this implies that, at the very least, we are interested in estimation conditional on proceeding to stage 2. Conditional estimation of the classification accuracy of a single marker after designs that allow early termination for futility have been discussed in the literature and could provide the motivation for conditional estimators of the performance of a panel of biomarkers [18, 19].

We considered a scenario where only a small number of markers ($n = 15$) were considered as candidates for our panel. There are many situations where we would be interested in developing a marker panel from a much longer list of candidate markers. This does not fundamentally change our design but it may have an impact on the operating characteristics of our study.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the Associate Editor and two referees for their helpful comments that resulted in a greatly improved manuscript and their collaborator, Dr. Steve Schmechel, for providing the motivating example for this work. The effort of the first author is partially supported by NIH grant R01-CA131013.

References

1. Pepe, MS. The statistical evaluation of medical tests for classification and prediction, Oxford Statistical Science Series. Vol. 28. Oxford University Press; Oxford; 2003.
2. 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]
3. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. Statistics in Medicine. 2003; 22(5):727–739. [PubMed: 12587102]
4. 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]
5. 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]
6. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000; 56(2):337–344. [PubMed: 10877287]
7. Tibshirani R. The lasso method for variable selection in the cox model. Statistics in Medicine. 1997; 16(4):385–395. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380. 10.1002/(SICI)1097-0258(19970228)16:4¡385::AID-SIM380¿3.0.CO;2-3 [PubMed: 9044528]
8. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software. Mar; 2011 39(5):1–13. URL http://www.jstatsoft.org/v39/i05. [PubMed: 21572908]
9. Jennison, C.; Turnbull, BW. Group sequential methods with applications to clinical trials. CRC Press Inc; 2000.
10. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979; 35(3): 549–556. [PubMed: 497341]
11. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64(2):191–199. URL http://biomet.oxfordjournals.org/content/64/2/191.abstract. 10.1093/biomet/64.2.191
12. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983; 70(3): 659–663. URL http://biomet.oxfordjournals.org/content/70/3/659.abstract. 10.1093/biomet/70.3.659
13. Koopmeiners JS, Feng Z. Asymptotic properties of the sequential empirical ROC and PPV curves under case-control sampling. Annals of Statistics. 2011; 39(6):3234–3261. [PubMed: 24039313]
14. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type i error probability spending functions. Statistics in Medicine. 1990; 9(12):1439–1445. [PubMed: 2281231]
15. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. Biometrika. 1990; 77(4):875–892. URL http://biomet.oxfordjournals.org/cgi/content/abstract/77/4/875. 10.1093/biomet/77.4.875
16. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. Biometrika. 1986; 73(3):573–581. URL http://biomet.oxfordjournals.org/cgi/content/abstract/73/3/573. 10.1093/biomet/73.3.573

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

17. Troendle JF, Yu KF. Conditional estimation following a group sequential clinical trial. Communications in Statistics: Theory and Methods. 1999; 28:1617–1634.

18. Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. Statistics in Medicine. 2009; 28(5):762–779. [PubMed: 19097251]

19. Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a group sequential diagnostic biomarker study that allows early termination for futility. Statistics in Medicine. 2012; 31(5):420–435. [PubMed: 22238117]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Simulation results evaluating the effect of the proportion of samples used in stage 1, P, and the threshold for continuing to stage 2, ROC_{co} (0.1) on the operating characteristics of our study. Simulations were completed using $n_s = 5$, $W_s = (1, \ldots, 1)$, $V_s = (1, \ldots, 1)$ and $\rho = 0.0$. 10000 simulations were completed for each scenario.

	Null hypothesis[1]					Alternative hypothesis[2]				
P	**% Early Stop**	**E(SS)**	**Cond α**	**Uncond α**	***ROC* (0.1) for $X^t\hat{\beta}_1$**	**% Early Stop**	**E(SS)**	**Cond $1-\beta$**	**Uncond $1-\beta$**	***ROC* (0.1) for $X^t\hat{\beta}_1$**
Fixed-sample										
0.3	0.00	720	0.004	0.004	0.371	0.00	720	0.905	0.905	0.628
0.4	0.00	720	0.007	0.007	0.379	0.00	720	0.888	0.888	0.635
0.5	0.00	720	0.011	0.011	0.384	0.00	720	0.838	0.838	0.638
0.6	0.00	720	0.012	0.012	0.387	0.00	720	0.754	0.754	0.640
0.7	0.00	720	0.012	0.012	0.389	0.00	720	0.646	0.646	0.642
$ROC_{co}(0.1) = 0.40$										
0.3	0.43	502	0.006	0.003	0.372	0.00	718	0.908	0.904	0.629
0.4	0.46	521	0.007	0.004	0.380	0.00	719	0.883	0.882	0.635
0.5	0.48	546	0.011	0.006	0.385	0.00	720	0.832	0.832	0.638
0.6	0.50	576	0.011	0.006	0.387	0.00	720	0.756	0.756	0.640
0.7	0.50	612	0.016	0.008	0.389	0.00	720	0.635	0.635	0.642
$ROC_{co}(0.1) = 0.45$										
0.3	0.64	397	0.004	0.002	0.372	0.02	712	0.909	0.894	0.628
0.4	0.69	423	0.010	0.003	0.380	0.01	717	0.887	0.880	0.635
0.5	0.72	459	0.010	0.003	0.385	0.00	719	0.834	0.830	0.638
0.6	0.76	502	0.015	0.004	0.387	0.00	719	0.754	0.752	0.640
0.7	0.79	550	0.017	0.004	0.389	0.00	720	0.637	0.637	0.642
$ROC_{co}(0.1) = 0.50$										
0.3	0.81	311	0.005	0.001	0.373	0.05	697	0.904	0.863	0.629
0.4	0.85	352	0.009	0.001	0.380	0.03	706	0.879	0.851	0.635

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

	Null hypothesis[1]					Alternative hypothesis[2]				
P	% Early Stop	E(SS)	Cond α	Uncond α	*ROC* (0.1) for $X^t\hat{\beta}_1$	% Early Stop	E(SS)	Cond $1-\beta$	Uncond $1-\beta$	*ROC* (0.1) for $X^t\hat{\beta}_1$
0.5	0.89	398	0.009	0.001	0.384	0.02	711	0.837	0.817	0.638
0.6	0.92	454	0.008	0.001	0.388	0.01	716	0.753	0.743	0.640
0.7	0.94	516	0.013	0.001	0.389	0.01	718	0.634	0.627	0.642

[1] $ROC\,(0.1)_0 = 0.40$

[2] $ROC\,(0.1)_a = 0.65$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Simulation results evaluating the impact of marker parameters n_s, W_s, V_s and ρ on the operating characteristics of our study. Simulations were completed using P = 0.3 and ROC $(0.1)_{co}$ = 0.45. 10000 simulations were completed for each scenario.

	n_s = 3				n_s = 5				n_s = 7			
	Null hypothesis[1]		Alt hypothesis[2]		Null hypothesis[1]		Alt hypothesis[2]		Null hypothesis[1]		Alt hypothesis[2]	
	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$
$W_s = (1, \ldots, 1)$, $V_s = (1, \ldots, 1)$												
$\rho = 0.0$	385	0.003	712	0.913	397	0.002	712	0.896	407	0.002	713	0.876
$\rho = 0.2$	383	0.002	710	0.914	388	0.003	711	0.900	396	0.002	712	0.889
$\rho = 0.4$	382	0.004	710	0.916	388	0.003	711	0.903	393	0.002	711	0.892
$\rho = 0.6$	376	0.003	711	0.914	386	0.003	710	0.909	390	0.002	712	0.899
$W_s = (1, \ldots,1)$, $V_s = \left(\frac{n_s}{n_s}, \frac{n_s-1}{n_s}, \ldots, \frac{1}{n_s}\right)$												
$\rho = 0.0$	391	0.002	711	0.910	396	0.002	712	0.896	400	0.001	713	0.878
$\rho = 0.2$	383	0.003	710	0.910	394	0.002	711	0.900	402	0.002	711	0.886
$\rho = 0.4$	386	0.004	710	0.909	392	0.002	712	0.905	393	0.002	711	0.885
$\rho = 0.6$	384	0.003	710	0.913	390	0.003	711	0.903	393	0.003	712	0.888
$W_s = \left(\frac{n_s}{n_s}, \frac{n_s-1}{n_s}, \ldots, \frac{1}{n_s}\right)$, $V_s = (1, \ldots,1)$												
$\rho = 0.0$	387	0.002	710	0.906	392	0.002	711	0.892	393	0.002	712	0.877
$\rho = 0.2$	380	0.003	710	0.911	388	0.002	711	0.901	390	0.002	713	0.895
$\rho = 0.4$	377	0.003	710	0.917	383	0.002	711	0.908	388	0.002	711	0.899
$\rho = 0.6$	376	0.003	710	0.919	382	0.002	711	0.910	386	0.003	712	0.905
$W_s = \left(\frac{n_s}{n_s}, \frac{n_s-1}{n_s}, \ldots, \frac{1}{n_s}\right), V_s = \left(\frac{n_s}{n_s}, \frac{n_s-1}{n_s}, \ldots, \frac{1}{n_s}\right)$												
$\rho = 0.0$	385	0.002	712	0.911	387	0.002	712	0.895	395	0.002	711	0.879

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

	n_s = 3				n_s = 5				n_s = 7			
	Null hypothesis[1]		Alt hypothesis[2]		Null hypothesis[1]		Alt hypothesis[2]		Null hypothesis[1]		Alt hypothesis[2]	
	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$	E(SS)	Uncond α	E(SS)	Uncond $1-\beta$
$\rho = 0.2$	382	0.003	711	0.908	389	0.002	711	0.897	397	0.002	712	0.885
$\rho = 0.4$	384	0.003	710	0.916	385	0.002	712	0.905	386	0.002	710	0.890
$\rho = 0.6$	378	0.003	709	0.915	384	0.002	711	0.905	384	0.003	712	0.898
$W_s=\left(\frac{n_s}{n_s},\frac{n_s-1}{n_s},\ldots,\frac{1}{n_s}\right), V_s=\left(\frac{1}{n_s},\frac{2}{n_s},\ldots,\frac{n_s}{n_s}\right)$												
$\rho = 0.0$	387	0.003	712	0.907	390	0.003	712	0.893	396	0.002	711	0.878
$\rho = 0.2$	384	0.002	710	0.910	389	0.002	710	0.897	401	0.002	712	0.886
$\rho = 0.4$	382	0.002	711	0.910	387	0.002	711	0.902	396	0.002	711	0.889
$\rho = 0.6$	380	0.004	710	0.917	391	0.003	712	0.906	393	0.002	712	0.897

[1] $ROC\,(0.1)_0 = 0.40$

[2] $ROC\,(0.1)_a = 0.65$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Simulation results evaluating the effect of implementing additional interim analyses during stage 2 on the operating characteristics of our study. Simulations were completed using $P = 0.3$, $ROC\ (0.1)_{co} = 0.45$, $n_s = 5, W_s = (1, \ldots, 1)$, $V_s = (1, \ldots, 1)$ and $\rho = 0.0$. 10000 simulations were completed for each scenario.

Stage 2	Null hypothesis[1]			Alternative Hypothesis[2]		
Stopping Times	***E* (*SS*)**	**Cond α**	**Uncond α**	***E* (*SS*)**	**Cond $1 - \beta$**	**Uncond $1 - \beta$**
J = 1	397	0.004	0.002	712	0.909	0.894
J = 2	325	0.005	0.002	609	0.901	0.887
J = 3	308	0.005	0.002	572	0.899	0.885
J = 4	302	0.004	0.002	551	0.891	0.878

[1] $ROC\ (0.1)_0 = 0.40$

[2] $ROC\ (0.1)_a = 0.65$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 4

The results of nine simulated scenarios for the proposed study. Included in the table are the estimated *ROC* (0.1) after stage 1($\hat{ROC}$ (0.1) $_{stg1}$), the sequential test statistic ($Z_{ROC(0.1),1}$ - $Z_{ROC(0.1),4}$) and the total sample size for the nine simulated scenarios.

Scenario	$\hat{ROC}_{stg1}$	$Z_{ROC(0.1),1}$	$Z_{ROC(0.1),2}$	$Z_{ROC(0.1),3}$	$Z_{ROC(0.1),4}$	Reject Null hypothesis[1]	Total n
Scenario 1	0.32	-	-	-	-	-	216
Scenario 2	0.35	-	-	-	-	-	216
Scenario 3	0.48	-3.36	-	-	-	No	342
Scenario 4	0.43	-	-	-	-	-	216
Scenario 5	0.51	0.8	-0.49	-	-	No	468
Scenario 6	0.48	0.99	0.53	1.16	-	No	594
Scenario 7	0.58	0.72	1.36	1.3	1.05	No	720
Scenario 8	0.67	1.5	2.89	-	-	Yes	468
Scenario 9	0.62	1.57	2.93	-	-	Yes	468

[1] *ROC* $(0.1)_0 = 0.40$