Published in final edited form as:
*Stat Med*. 2014 December 20; 33(29): 5097–5110. doi:10.1002/sim.6310.

# Two-Stage Adaptive Cutoff (TACO) Design for Building and Validating a Prognostic Biomarker Signature

**Mei-Yin C. Polley**[*], **Eric C. Polley**, **Erich P. Huang**, **Boris Freidlin**, and **Richard Simon**
Biometric Research Branch, National Cancer Institute, 9609 Medical Center Drive Room 5W638 MSC 9735, Bethesda, MD 20892, U.S.A

## Abstract

Cancer biomarkers are frequently evaluated using archived specimens collected from previously conducted therapeutic trials. Routine collection and banking of high quality specimens is an expensive and time-consuming process. Therefore, care should be taken to preserve these precious resources. Here we propose a novel two-stage adaptive cutoff (TACO) design that affords the possibility to stop the biomarker study early if an evaluation of the model performance is unsatisfactory at an early stage, thereby allowing one to preserve the remaining specimens for future research. In addition, our design integrates important elements necessary to meet statistical rigor and practical demands for developing and validating a prognostic biomarker signature, including maintaining strict separation between the datasets used to build and evaluate the model and producing a locked-down signature to facilitate future validation. We conduct simulation studies to evaluate the operating characteristics of the proposed design. We show that under the null hypothesis when the model performance is deemed undesirable, the proposed design maintains type I error at the nominal level, has high probabilities of terminating the study early, and results in substantial savings in specimens. Under the alternative hypothesis, power is generally high when the total sample size and the targeted degree of improvement in prediction accuracy are reasonably large. We illustrate the use of the procedure with a dataset in patients with diffuse large-B-cell lymphoma. The practical aspects of the proposed designs are discussed.

### Keywords

cancer biomarker; two-stage design; biomarker validation; cross-validation; early stopping

## 1. Introduction

Increased understanding of cancer biology and advances in biotechnology have revolutionized the landscape of oncology drug development and brought us closer to the realization of “precision medicine”. A major thrust to this new paradigm involves development of biomarkers that will improve our ability to identify patients who are at an elevated risk of developing cancer (screening biomarkers), have a poor prognosis (prognostic biomarkers), or who are more likely to benefit from specific therapies

[*] Correspondence to: Mei-Yin C. Polley, Ph.D., 9609 Medical Center Drive Room 5W638 MSC 9735, Bethesda, MD 20892. polleymc@mail.nih.gov.

(predictive biomarkers). The oncology community has committed expansive resources to biomarker studies during the last decade. Ideally the development of a biomarker would be synchronized with the development of cancer therapy, but in reality biomarker development often lags behind therapeutic development. The reasons for this asynchrony may include an incomplete understanding of the mechanism of action of a drug, the uncertainty about what form of a marker is most relevant, and technical difficulties with marker assay development. As such, it is common for cancer biomarkers to be identified using archived specimens collected from previously conducted therapeutic trials. While routine collection and banking of high quality specimens from clinical trials provides a rich resource for conducting biomarker research, the development of the specimen banking system involves an expensive and time-consuming process. Therefore, care should be taken to preserve these valuable and scarce specimens. Motivated by the use of early looks in treatment clinical trials, here we propose a two-stage adaptive cutoff (TACO) design that affords the possibility to stop the biomarker study early if an evaluation of the model performance is unsatisfactory at an early stage, thereby allowing one to preserve the remaining specimens for future research.

A prognostic signature is a collection of biomarkers that are combined through some mathematical model to provide either a continuous score or a categorical classifier. Prognostic signatures can be used to identify patients receiving standard treatment (or no treatment if that is the standard of care) into subgroups with distinct clinical outcomes. In spite of the considerable increase in the number of cancer biomarker studies in recent years, the pace of adoption of new biomarkers into clinical practice has been slow. This phenomenon can, in part, be attributable to the lack of rigorous statistical design and analysis in many biomarker studies. In a comprehensive review of published studies reporting gene expression-based prognostic signatures for non-small cell lung cancer, Subramanian and Simon [1] identified common methodological deficiencies which included the failure to maintain a strict separation between the data used for model development from the data for model evaluation and the lack of a completely specified prognostic model to allow an independent validation on a separate dataset. The performance measurement of a prognostic model computed based on the same data used to develop it is known to be severely biased in an optimistic direction (the resulting estimates of the performance measure are referred to as the “resubstitution statistics”). This is especially problematic when the number of variables is much larger than the number of samples due to the risk of overfitting the data [2]. Re-sampling methods (e.g. split sample, *K*-fold cross-validation, leave-one-out cross-validation (LOOCV), Monte Carlo cross-validation and bootstrapping), when used properly, provide an almost unbiased estimate of the generalization error of the full sample classifier built using the entire dataset [3]. There is often a desire, however, to evaluate this “final” locked-down signature on an independent dataset where the effects of other sources of variability can be appreciated. The full model specification should include a list of variables in the model (including any interaction or non-linear terms), their weights in the multivariable model (e.g. regression coefficients) and any cutoff used to define risk groups. Clinical decision-making regarding treatment options often relies on the application of a cutoff (to some continuous score) to classify patients into “high” or “low” risk groups. In practice, however, the determination of cutoffs often suffer the lack of statistical rigor. A common pitfall is to search all possible cutoffs within the entire spectrum of the continuous predicted outcome and select one that maximizes the model performance in the same dataset. This practice is well-known to give inflated measure of the model performance and should be avoided [4].

The two-stage design we propose here represents a framework that integrates all the aforementioned elements necessary to meet the practical demands and statistical rigor for a prognostic biomarker signature. Specifically, in Stage 1, a statistically valid procedure based on the unbiased 10-fold cross-validated error is employed to evaluate the model performance. If the initial model performance based on Stage 1 data proves to be unsatisfactory, the study is terminated early. Otherwise, the study proceeds to Stage 2. In Stage 2, a final prognostic signature is built based on the entire Stage 1 data including the identification of a cutoff for classifying patients into distinct risk groups. This locked-down prognostic signature (including the cutoff) is then independently validated with Stage 2 data. The technical aspects of the design focus mainly on development of prognostic signatures in the context of a binary outcome, although with suitable adjustment similar approaches may be contemplated for other types of biomarker signatures or clinical outcomes (e.g. time-to-event). To date, we are aware of only one paper in the literature that incorporates the idea of early stopping into the development of biomarker signatures [5]. However, their work does not build in a formal statistical test for model performance in an early stage and hence the decision as to whether the model holds sufficient promise to be further validated in the second stage is somewhat arbitrary.

Conservation of biospecimens is a relatively less pertinent issue in exploratory or hypothesis-generating studies involving high-dimensional data (e.g. gene-expression profiling studies) since such studies are typically based on smaller datasets readily accessible by the investigators. For this reason, the technical development of our design focuses mainly on settings in which the number of specimens is much larger than the number of biomarker variables (i.e. “large *n*, small *p*”) such that an early stopping will result in a sizable saving of specimens. In particular, we assume that biomarker discoveries have been carried out previously with smaller datasets and a small panel of biomarkers (e.g. 5 to 20) has been identified to be of prognostic value. The primary interest is to combine these promising biomarkers into a prognostic signature and to evaluate the ability of the signature to accurately predict patient outcomes. These *candidate* markers may be selected based on biological insight, data-driven approaches, or the combination of both.

This paper is organized as follows: In Section 2, we describe the two-stage adaptive cutoff design. In Section 3, we present a simulation study to evaluate the operating characteristics of the design. In Section 4, we illustrate the use of the proposed design with a real dataset. We conclude with a discussion on practical aspects of the proposed design in Section 5.

## 2. Methods

### 2.1. Theoretical Framework

We start with notational set-up and a brief overview of the theoretical framework underlying the *class prediction* problem in which the goal is to predict some clinical outcome of a patient (or associated specimen) from a set of covariates. Suppose one observes *n* independent and identically distributed random variables $O = (O_1, \ldots, O_n)$ with unknown distribution $\mathcal{P}$. Each observation in $O$ consists of a dichotomized outcome $Y \in \{0, 1\}$ and a $p$-dimensional vector ($p \ll n$) of measured covariates $X \in \mathcal{X}$, such that $O_i = (X_i, Y_i)$ for $i = 1, \ldots, n$. In a biomarker study, the outcome $Y$ may indicate some disease status (e.g. disease relapse by 5 years) and the covariates $X$ may include standard clinico-pathological variables, molecular variables such as genetic variations, or imaging variables. The observations $o = (o_1, \ldots, o_n)$ are realizations of underlying random variables $O = (O_1, \ldots, O_n)$.

Define a classifier $f(X) = I[m(X|\beta) > C]$ as a rule that maps the covariate space $\mathcal{X}$ onto the class space $\{0, 1\}$, where $m(X|\beta)$ denotes a parametric statistical model that can be built, for example, via regression techniques (linear or non-linear) and $C$ is the cutoff. Given the observations $(o_1, \ldots, o_n)$, the classifier $\hat{f}(x) = I(m(x|\hat{\beta}) > \hat{C})$ is obtained by substituting the parameters $(\beta, C)$ with their estimates $(\hat{\beta}, \hat{C})$ from the data. We are interested in the performance of the given classifier $\hat{f}(x)$ in terms of its prediction accuracy in the population. Let $L(Y, \hat{f}(X))$ denote the *loss function* for measuring errors between $Y$ and $\hat{f}(X)$. For a given loss function $L$, define the *conditional risk* for a given $\hat{f}$ as

$$\theta = E[L(Y, \hat{f}(X))] = \int L(y, \hat{f}(x)) dP(x, y). \quad (1)$$

Note that here the conditional risk refers to the expected error for a fixed classifier $\hat{f}$ constructed based on observed data $o = (o_1, \ldots, o_n)$. In the technical development that follows, we focus on the indicator loss function $L(Y, \hat{f}(X)) = I[Y \neq \hat{f}(X)]$. The indicator loss function is a common choice of a loss function with a binary outcome $Y$ although other choices are possible (e.g. the squared error loss $(Y - \hat{f}(X))^2$) [6].

A naive estimator of the conditional risk $\theta$ for the classifier $\hat{f}(x)$ is the *resubstitution error* which averages the loss over the observed data $(o_1, \ldots, o_n)$ used to build the classifier

$$\hat{\theta}_{RE} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i)). \quad (2)$$

It has been shown that the resubstitution error can severely underestimate the expected error in finite sample situations [2].

One of the most widely used methods for estimating the conditional risk $\theta$ is cross-validation. Cross-validation involves using part of the available data to build a classifier and a different part of the data to evaluate its performance. For example, in a $K$-fold cross-validation, one splits the data into $K$ roughly equal-sized and mutually exclusive subsets each with size $(s_1, \ldots, s_K)$, respectively, such that $\sum_{k=1}^{K} s_k = n$. For the *kth* subset, we develop a classifier with combined data from the remaining $K$–1 subsets, and calculate the misclassification error of the classifier when predicting the *kth* part of the data. This process is repeated for $k = 1, 2, \ldots, K$ and the $K$ estimates of the misclassification error are then combined to get the *K-fold cross-validated error*. In the case of a binary outcome and indicator loss, for example, the $K$-fold cross-validated error can be written as

$$\hat{\theta}_{KCV}=\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{s_k}\sum_{i=1}^{s_k}I\left(y_i\neq\hat{f}^{(k)}(x_i)\right)\right]\quad(3)$$

where $f^{(\hat{k})}(x)$ denotes the classifier developed with *kth* part of the data removed. Note that the case $K = n$ is known as the *leave-one-out cross-validated error*.

The performance of the classifier $\hat{f}(x)$ is evaluated in the following statistical hypothesis testing framework:

$$H_0:\theta\geq\theta_0\text{ vs. }H_1:\theta<\theta_0,\quad(4)$$

where $\theta_0$ is a pre-specified threshold below which the performance of a classifier would be deemed acceptable. In the context of a biomarker study, the alternative hypothesis would correspond to the scenario in which the probability that the classifier $\hat{f}(x)$ built on the observed data $(o_1, \ldots, o_n)$ misclassifies a future specimen is less than $\theta_0$ (see Discussion for the choice of $\theta_0$).

Dudoit and van der Laan proposed asymptotic confidence intervals for the conditional risk $\theta$ using a class of resampling methods including cross-validation [7]. They showed that these methods have good asymptotic properties when using common loss functions to assess the performance of the classifier $\hat{f}$. The asymptotic results were derived under the assumption that the size of the test set in the resampling methods converges to infinity which should apply well to the "large $n$, small $p$" setting. Based on their results, an asymptotic 100(1– $\alpha$) % confidence interval for the conditional risk $\theta$ can be written as s

$$\hat{\theta}_{KCV}\pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}\quad(5)$$

where $\hat{\theta}_{KCV}$ is the $K$-fold cross-validated error defined in (18), $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$ for the standard normal cumulative distribution function $\Phi(.)$ and the estimated variance term is

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left[I\left(y_i\neq\hat{f}(x_i)\right)-\hat{\theta}_{RE}\right]^2,\quad(6)$$

where $\hat{\theta}_{RE}=\frac{1}{n}\sum_{i=1}^{n}I[y_i\neq\hat{f}(x_i)]$ is the resubstitution error.

## 2.2. The TACO Design

Assume a total of $N$ specimens are available for the study. In Stage 1, $S$% of the total $N$ samples ($n_1$ samples) are used to obtain an unbiased estimate of the conditional risk based on 10-fold cross-validation. A test for (4) is performed using a statistical procedure. If $H_0$ is not rejected (indicating an unacceptable performance of the classifier), the study is terminated early. Otherwise, the study continues to Stage 2. In Stage 2, all data in Stage 1 are combined to develop a prognostic classifier (or signature) including the identification of a cutoff. This locked-down signature is then validated using the remaining independent (1–

*S*)% samples in Stage 2 ($n_2 = N–n_1$ samples). The significance levels for Stage 1 ($\alpha_1$) and Stage 2 ($\alpha_2$) are chosen so that ($\alpha_1 \times \alpha_2$) does not exceed some pre-specified type I error $\alpha$ (e.g. 0.05). A flow chart of the design is given in Figure 1. Below we describe the technical steps of the procedure in detail.

### Stage 1

**1-1. First layer (10-fold cross-validation):** The samples are divided randomly into 10 groups with roughly equal size. Nine of the groups (i.e. 90% of the samples) serves as the "training set" while the remaining group (i.e. 10% of the samples) constitutes the "test set". For each training set, develop a prognostic signature based on *p* biomarkers using standard statistical techniques suitable for a binary outcome (e.g. logistic regression). The performance of the signature as measured by misclassification error is computed on the test set in each iteration.

**1-2. Second layer (cutoff identification):** To identify a cutoff to dichotomize the estimated probability for each specimen in the test set, each training set is further subdivided into a "learning set" (90% of the training set) and an "evaluation set" (10% of the training set) [8]. For each learning set, a prognostic signature is developed. The estimated probability of an event for each specimen in the evaluation set is then predicted based on the signature. A fine grid of cutoffs in the range [0, 1] are applied to the predicted probability for each specimen in the evaluation set and the misclassification error with each cutoff is recorded. This process is repeated for each split of data within the training set. For each fixed cutoff, the average misclassification error across the 10 learning-evaluation splits is computed. The cutoff that yields the minimum average misclassification error is the chosen cutoff to be applied to the test set. A diagram of the two-layer procedure in Stage 1 of the TACO design is given in Figure 2.

**1-3. Early test of model performance:** Repeat steps 1-1 and 1-2 until all 10 training-test splits are exhausted. The average misclassification error across the ten test sets constitutes the unbiased 10-fold cross-validated error. Perform a statistical test at significance level $\alpha_1$ based on the Dudoit and van der Laan procedure [7]. Specifically, we will reject the null hypothesis $H_0$ in (4) and proceed to Stage 2 if

$$Z_1 = \frac{\hat{\theta}_1 - \theta_0}{\hat{\sigma}_1 / \sqrt{n_1}} < z_{\alpha_1}, \quad (7)$$

where $\hat{\theta_1}$ is the 10-fold cross-validated error defined in (18) (with $K = 10$) and $\hat{\sigma_1}$ is the estimated standard error term defined in (6) using Stage 1 data. Otherwise, if $H_0$ is not rejected (the initial signature performance is unsatisfactory), terminate the study early.

### Stage 2

**2-1. Locked-down signature:** Build a prognostic signature using all Stage 1 data, including the selection of a cutoff based on the second-layer procedure described in step 1-2.

**2-2. Independent validation:** Apply the locked-down prognostic signature in step 2-1 to Stage 2 data, {( $x_i^{\prime}, y_i^{\prime}$), *i* = 1, …, $n_2$}. Since Stage 2 data are completely independent from Stage 1 data and are not used for signature building, the estimated misclassification error $\hat{\theta}_2=\frac{1}{n_2}\sum_{i=1}^{n_2} I(y_i^{\prime} \neq \hat{f}(x_i^{\prime}))$ follows a binomial distribution, where $\hat{f}$ denotes the signature built based on Stage 1 data. Based on normal approximation, we have

$$\hat{\theta}_2 \sim N\left(\theta_0, \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}\right) \quad (8)$$

We will reject $H_0$ at significance level $\alpha_2$ and conclude that the prognostic value of the signature is validated if

$$Z_2=\frac{\hat{\theta}_2-\theta_0}{\sqrt{\frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}}<z_{\alpha_2}. \quad (9)$$

Otherwise, if $H_0$ is not rejected, we will conclude that the prognostic signature fails to validate.

## 3. Simulation Studies

We conducted a simulation study to evaluate the operating characteristics of the proposed design under a variety of settings. We assumed that *X* is a 10-dimensional vector of biomarkers. We simulated the biomarkers from independently and identically distributed standard normal distributions. Assume that the collective ability of the 10-dimensional biomarkers to accurately classify patients into good (*Y* = 0) versus poor (*Y* = 1) prognosis subgroups is determined by the following logistic regression model

$$\log\left(\frac{p(y_i=1|x_{i1},\ldots,x_{i10})}{1-p(y_i=1|x_{i1},\ldots,x_{i10})}\right)=\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{10} x_{i10} \quad (10)$$

The binary outcome for the *ith* patient, $y_i$, is thus drawn from a Bernoulli distribution with the probability of being in a poor prognosis subgroup:

$$p_i=p(y_i=1|x_{i1},\ldots,x_{i10})=\frac{e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{10} x_{i10}}}{1+e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{10} x_{i10}}} \quad (11)$$

We evaluate the operating characteristics of the design under the following four scenarios:

- Null Hypothesis (NH): $\theta$ = 0.35
- Alternative Hypothesis 1 (AH1): $\theta$ = 0.20
- Alternative Hypothesis 2 (AH2): $\theta$ = 0.25
- Alternative Hypothesis 3 (AH3): $\theta$ = 0.30

Here we include three alternative scenarios representing varying magnitudes of improvement in order to evaluate the impact of the effect size on the operating characteristics of the design. In each scenario, we also vary the total sample size (n = 200, 300, 400) and the proportion of samples used in Stage 1 (*S*% = 25%, 50%, 75%).

The ability of the prognostic signature to discriminate patient status is influenced by the $\beta$'s coefficients in the logistic regression model (10). However, there are infinitely many combinations of $\beta$'s values that would give rise to a specified true conditional risk $\theta$. To simplify, we fixed the values of nine $\beta$'s coefficients in (10) and performed a fine grid search of the remaining $\beta$ coefficient (with an increment 0.02) such that the true conditional risk is equal to the desired value. This is possible by recognizing that the true conditional risk $\theta$ decreases with increasing values of the remaining $\beta$. Details on the determination of the set of ($\beta_1$, $\beta_2$, …, $\beta_{10}$) values that correspond to the true conditional risk in each of the above scenario are given in the Appendix (see Appendix 1).

For each hypothesis scenario and (*N*, *S*%) configuration, we simulated 1,000 independent realizations of $\{(x_{i1}, \ldots, x_{i10}, y_i), i = 1, \ldots, N\}$. With each simulated dataset, we applied the proposed procedure using significance levels $\alpha_1 = 0.25$ and $\alpha_2 = 0.2$ for Stage 1 and Stage 2, respectively. A fine grid with 0.01 increment was used in the search of the cutoff value. In each setting, we recorded the empirical probability of stopping the study early, the probability of declaring that the prognostic signature is independently validated (this would represent the *type I error* under the null hypothesis and *power* under the alternative hypothesis), and the expected sample size E(SS). A reasonable design should terminate the study early with a high probability under the null hypothesis and hence minimizes the expected sample size but provides sufficient statistical power to detect a meaningful improvement in the prediction error of the classifier under the alternative hypothesis.

Table 1 and Table 2 present the results of the simulation studies. Under the null hypothesis, we note that the type I error is maintained under the overall level of 0.05 in all settings. Further, the probability of terminating the study early ranges from 77% to 84%. The percentage of saving in sample size (computed as the percent reduction in sample size compared with the total sample size) ranges from 19% to 62%, with an average of about 39%. Under AH1 (an improvement of 15% in prediction error), the power exceeds 92% in all settings. Under AH2 (an 10% improvement), the power is generally acceptable (exceeds 81% in most cases) except when the sample size in either stage is only 50 (power is 67% and 68%). Power is generally low when the targeted degree of improvement is as low as 5% (AH3). In AH1 and AH2, we note that for a fixed sample size *N*, the power tends to be maximized when sample size is equally split between the two stages. Intuitively this is because the power of the procedure depends on allocating adequate sample sizes to reject the null hypothesis in both stages. Overall, our simulation results demonstrate that desirable operating characteristics of the proposed design can be achieved when the total sample size is reasonably large (i.e. $N > 200$) with an equal split of sample size between the two stages. The power of the procedure in most cases is reasonable when the targeted degree of improvement in prediction error is at least 10%.

## 4. Example: An Application to a Dataset

We illustrate the proposed design with a real dataset in patients with diffuse large-B-cell lymphoma (DLBCL). DLBCL is a curable disease for many patients using anthracycline-based chemotherapy regimens such as a combination of cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP) [9]. Recently, the addition of rituximab immunotherapy (R-CHOP) has also been shown to substantially prolong patient survival [10]. DLBCL is a biologically diverse disease and much research has been devoted to identify molecularly distinct subtypes. Rosenwald et al. described a study in which gene-expression profiling was performed in pre-treatment tumor-biopsy specimens of 181 newly-diagnosed DLBCL patients who received CHOP or similar treatment regimens [11]. In that study, two biologically and clinically distinct molecular subtypes of DLBCL were identified, namely the germinal-center B-cell-like (GCB) subtype and the activated B-cell-like (ABC) subtype. Patients with the GCB disease tend to have better long-term prognosis than those with the ABC disease (5-year overall survival were 60% and 30%, respectively). We refer to this signature as the “subtype” signature henceforth. Using the same *training* dataset but different analytical approaches, the same investigators discovered a number of other gene expression-based signatures: the “germinal-center B-cell” signature, the “proliferation” signature, the “major histocompatibility complex class II” (MHC) signature, the “stromal-1” signature and “stromal-2” signature. For a detailed description of the biological basis of these signatures, see Lenz et al. [12]. In addition to molecular signatures, many clinical variables such as patient age, ECOG status and Ann Arbor Stage have also been shown to have prognostic value for this disease population [13].

We obtained access from the investigators to an independent dataset of 175 newly diagnosed DLBCL patients. No aspect of identification of the gene-expression signatures or development of the survival model was performed with this dataset. These patients constituted an R-CHOP cohort that received similar chemotherapy plus rituximab. We illustrate the use of the proposed TACO design in this dataset with the goal to combine putative clinical prognostic variables with discoveries from the previous training dataset into one signature. The clinical outcome of interest is death at 18 months (a time point by which 25% of the patients have deceased). ECOG status and Ann Arbor stage were analyzed as binary variables (≤ 1 or > 1). The *subtype* signature was binary in nature (GCB vs. ABC). Age and all other gene-expression signatures were analyzed as continuous variables. We test the null hypothesis that the conditional risk of the prognostic signature is greater than or equal to 35% ($H_0$: $\theta \geq 0.35$) against the alternative hypothesis that the conditional risk is smaller than 35% ($H_1$: $\theta < 0.35$).

We split the dataset roughly equally between the two stages. Significance levels 0.25 and 0.2 were used to test the performance of the prognostic signature in Stage 1 and Stage 2, respectively. Stepwise logistic regression, implemented with the step function in the R software [14], was used to build all prognostic models. In Stage 1, the resubstitution prediction error was 0.25 which represented an optimistically downwards estimate of the true conditional error. The 10-fold cross-validated error was 0.29. The test statistic in Stage 1 based on (7) was $Z_1 = -1.14$ (p-value = 0.13) leading to the rejection of $H_0$ and continuation to Stage 2. A final prognostic signature was built based on all Stage 1 data.

Table 3 gives the final locked-down multivariable logistic regression model. Stage 1 data were also used to identify the cutoff based on the procedure described in step 1-2 of Section 2.2. A fine grid with an 0.01 increment was applied and the resulting cutoff was 0.79.

In Stage 2, the probability of death by 18 months for each specimen was estimated using the logistic regression model developed in Stage 1, i.e.

$$\hat{p}(y_i=1)=\text{logit}^{-1}(-8.48+1.61\times I[\text{ECOG}>1]-1.74\times I[\text{Subtype}=\text{GCB}]+0.96\times\text{Stromal}-2). \quad (12)$$

The cutoff 0.79 was applied to the estimated probabilities to classify each specimen in Stage 2 into either a “short survivor” ($y_i$ = 1) or a “long survivor” ($y_i$ = 0). The estimated prediction error using Stage 2 data was 0.26. The test statistic in Stage 2 based on (9) was $Z_2$ =−1.81 (p-value = 0.04) leading to the rejection of $H_0$. Thus, the prognostic signature in (12) (including the cutoff 0.79) was independently validated.

## 5. Discussion

In practice, the suitable choice of the conditional risk $\theta_0$ depends on the particular clinical scenario and the targeted degree of improvement in signature performance. For example, suppose that the goal of the study is to develop a prognostic signature to predict 10-year disease recurrence. Assume that 35% of the patients are expected to experience disease recurrence within 10 years. The investigators consider a positive predictive value (PPV) (i.e. the probability that a patient truly experiences an event given that the signature predicts an event) and a negative predictive value (NPV) (i.e. the probability that a patient does not experience an event given that the signature predicts the patient to be event-free) of 70% to be unsatisfactory. Suppose an improvement in PPV by 10% and in NPV by 20% (i.e. PPV of 80% and NPV of 90%, respectively) would be deemed clinically meaningful to warrant further assay development. In this case, the conditional risk under the null and the alternative hypothesis would be 30% and 14%, respectively (see Appendix 2).

Various components of the proposed design may be altered although further simulation studies are needed to evaluate the impact such changes have on the overall properties of the design. For example, the 10-fold cross-validation in Stage 1 may be substituted with other resampling methods such as bootstrap. In a more confirmatory setting where a reasonably large number of specimens is available, the difference in signature performance among different resampling methods may not be profound [3]. The use of stepwise logistic regression as a model building technique may also be replaced by other methods suitable for binary outcomes. For example, when a large number of input variables is present in the dataset, penalized regression methods such as LASSO or the Elastic Net are useful to help avoid overfitting [15, 16]. A vast literature on statistical model building exists [6]. While it is beyond the scope of this work to compare among different model building approaches, we stipulate that simplicity or complexity of the model should not dictate choice of a particular approach; rather, the overall performance of the model should serve as the primary consideration. We have chosen an indicator loss function for its practical appeal and ease of interpretation, but other loss functions are possible. For example, the mean squared error can be used to measure the difference between the true binary outcomes and the predicted probabilities. In this case, a cutoff to dichotomize the predicted probability of an event into a binary outcome is not needed, hence obviating the need for the cutoff search in step 1-2. The choice of the signature performance metric may be context dependent. In this work, we focus on the conditional risk mainly because of its mathematical simplicity. For a binary classifier, other choices of performance metrics may include NPV, PPV and Area Under the ROC Curve (AUC). Minimizing the conditional risk $\theta$ may not be relevant to some clinical questions. In those settings, other performance metrics may be considered. Our general TACO framework may be tailored to test other performance metrics as long as a loss function can be properly defined. It should be noted, however, that the theoretical development of the TACO procedure is greatly simplified due to the choices of the loss function and performance metric necessary to construct the statistical testing procedure in Stage 1 and Stage 2. Any adaptations may not be mathematically trivial and will require further methodological development. Finally, the overall framework of the proposed design may be applied to "large *p*, small *n*" settings, although again the operating characteristics of the design will require further investigation.

Note that when $\alpha_1 \to 1$ and $\alpha_2 \to \alpha$, the TACO design reduces to a simple split-sample approach in which Stage 1 data are used to build a prognostic signature without the possibility for early stopping and Stage 2 data are used to evaluate the signature performance. We evaluated the operating characteristics of the contrasting one-stage design ($\alpha_1 = 1$, $\alpha_2 = 0.05$) for each scenario in Table 1 and Table 2. In general, we found that the one-stage competing designs control the type I error at the nominal 0.05 level in all scenarios. Under all of the alternative hypotheses considered, the one-stage designs have slightly higher statistical power compared with their TACO counterparts when a larger percentage of specimens are allocated to Stage 2 ($S$ = 25%). However, this power gain is accompanied by an increase in the sample size, which is equal to the total planned sample size since the one-stage design does not allow the possibility of stopping the study early. The TACO designs have higher statistical power than their competing one-stage designs in all other scenarios when the sample split is equal between the two stages ($S$ = 50%) or when a larger percentage of specimens is allocated to Stage 1 ($S$ = 75%).

In the data example and simulation studies, we chose the significance levels $\alpha_1$ and $\alpha_2$ such that the multiplication of the two is equal to some desired overall type I error (e.g. $\alpha = 0.05$) (as a consequence of the independence between the datasets in two stages). In practice, the choice of $\alpha_1$ may be guided by the scarcity and the difficulty involved in obtaining the specimens. For example, for certain specimens that are extremely rare, one may choose a more stringent $\alpha_1$ (e.g. 0.25) to ensure that a reasonably promising prognostic value of the signature needs to be established at the early stage before attempting the rest of the specimens. The choice of $\alpha_1$ should be made in conjunction with $\alpha_2$ to ensure that sufficient rigor can be achieved in the validation of the signature at the second stage. Under the guiding principle that a more stringent statistical criterion should be reserved for the final validation of the signature, we recommend using the general rule: $\alpha_1 \leq 0.50$ and $\alpha_2 \leq 0.20$. A few examples that may be useful in practice are: $(\alpha_1, \alpha_2)$ = (0.25, 0.2), (0.30, 0.17), (0.35, 0.14). Further, when sample size permits, $(\alpha_1, \alpha_2)$ may be chosen so that the product of the two values may be smaller than the conventional 0.05 level.

A larger sample size in Stage 1 will on average result in a locked-down signature with higher classification accuracy in independent data. However, the trade-off is an inevitably smaller sample size for validating the signature performance in Stage 2. While the impact this trade-off has on the statistical power of the overall procedure may be context-dependent, in the limited number of scenarios we considered in our simulation studies, an equal split of sample size between the two stages appeared to yield optimal power. This finding is consistent with Dobbin and Simon who suggested that “the optimal proportion of cases for the training set tended to be in the range of 40% to 80% for the wide range of conditions studied”, although their investigation was focused on high dimensional data problems and did not consider a formal statistical test for the training data [17]. Furthermore, the marginal gain in classification accuracy as a result of a larger Stage 1 sample size may be less profound when the true conditional risk is closer to the zero boundary (or equivalently, when the true classification accuracy is closer to 100 %). This implies that the effect a larger Stage 1 sample size has on the observed effect size (and consequently the Stage 2 test statistic) may depend on the value of the true conditional risk. In general, for fixed $\alpha_1$ and $\alpha_2$, the operating characteristics of TACO are jointly influenced by the true conditional risk, the Stage 1 sample size, and the Stage 2 sample size. Therefore, while the classification accuracy of the locked-down signature may be improved by maximizing the sample size in Stage 1, it is important to note that this strategy does not always lead to an optimal design. In practice, we encourage the investigators to conduct simulation studies based on the particular clinical setting of interest to calibrate the design parameters (e.g. total sample size, sample split between the two stages) so as to optimize the operating characteristics of the design (i.e. power, type I error and expected sample size). The R functions for implementing our procedure and programs for simulating the operating characteristics of the design are available upon request from the authors.

In this article, we have studied the properties of the proposed design in settings where a relatively large sample size is available ($N > 200$). In reality it may be practically infeasible to set up large prospective studies with building a prognostic signature as the primary objective since the collection of clinical outcomes is a costly, resource-intensive and time-consuming process. As such, we envision that the most suitable setting for the TACO design would be retrospective studies of biomarker signature development using archived specimens and clinical data previously collected in large single-arm therapeutic trials where patients are uniformly treated with some treatment regimen (or a sub-study on one treatment arm of a large randomized trial). Alternatively, our design can also be built into a prospective therapeutic clinical trial as a secondary objective with the goal to develop a prognostic biomarker signature using data collected during the trial. While the general framework of the two-stage design may be extended to include more than one treatment arms, the development of such *predictive* signatures requires more thought and is beyond the scope of this work.

A salient feature of our design is that it produces a single locked-down signature at the end of Stage 1 to facilitate future independent validations. In practice, the two-stage nature of the design can also help facilitate the implementation of an “honest broker” system in which the signature builder(s) in Stage 1 remains strictly blinded to the clinical outcomes in the confirmatory dataset in Stage 2 until after individual predictions have been made for each specimen. Finally, some caution is required when interpreting the results of the study. While the term “validation” is widely used in practice, it may mean different things in different contexts or to different investigators. Within the context of our design, a “validated” signature by TACO would mean that it is highly probable that the signature in hand achieves certain accuracy (i.e. true positive + true negative), as defined by the hypothesis testing set-up. In some situations, minimizing the misclassification error rate may not be most relevant to the underlying clinical question of interest. For example, in some disease settings, PPV and NPV may be more relevant metrics that investigators will rely on to judge if a signature is satisfactory. The meaning of PPV and NPV may also be different to the investigators depending on the specific signature. For some signatures or tests, a higher PPV may be desirable while a slightly lower NPV may be tolerated. For other signatures, NPV may be deemed more important on the contrary. In real life applications, it is likely that a more comprehensive assessment of various aspects (e.g. PPV, NPV, accuracy, etc.) is necessary depending on the specific disease setting and the intended clinical use of the signature. Even if the *clinical validity* is established by demonstrating that the signature has a suitable strong association with a clinical outcome of interest, it does not imply that the signature is ready to direct patient care [18]. Establishing the *clinical utility* of a signature will require evidence that the use of the signature to direct patient care will result in favorable balance between benefits and harm, thereby leading to meaningful improvement in health outcomes such as quality of life, prolonged survival or reduced cost.

In summary, we propose a two-stage design useful for developing and validating a prognostic biomarker signature which includes an early stopping in the event of a poor signature performance. Novel statistical designs like this are needed to allow rigorous evaluation of prognostic biomarker signatures while ensuring efficient use of specimens.

## Acknowledgments

The authors are thankful to Dr. George Wright of the Biometric Research Branch at NCI for providing the diffuse large-B cell lymphoma dataset. The authors also wish to thank the editor and two referees whose helpful comments and suggestions led to an improved paper.

## References

1. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? Journal of the National Cancer Instiitute. 2010; 102(7):464–474.
2. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. Journal of the National Cancer Instiitute. 2003; 95(1):14–18.
3. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005; 21(15):3301–3307. [PubMed: 15905277]
4. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using optimal cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Instiitute. 1994; 86(11):829–835.
5. Koopmeiners JS, Vogel RI. Early termination of a two-stage study to develop and validate a panel of biomarkers. Statistics in Medicine. 2013; 32(6):1027–1037. [PubMed: 23413213]
6. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. 2. Springer; 2011.

7. Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology. 2005; 2(2):131–154.

8. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006; 7:91. [PubMed: 16504092]

9. Fisher RI, Gaynor ER, Dahlberg S, Oken MM, Grogan TM, Mize EM, Glick JH, Coltman CA Jr, Miller TP. Comparison of a standard regimen (chop) with three intensive chemotherapy regimens for advanced non-hodgkin's lymphoma. New England Journal of Medicine. 1993; 328:1002–1006. [PubMed: 7680764]

10. Coiffier B, Lepage E, Brire J, Herbrecht R, Tilly H, Bouabdallah R, Morel P, Van Den Neste E, Salles G, Gaulard P, et al. Chop chemotherapy plus rituximab compared with chop alone in elderly patients with diffuse large-b-cell lymphoma. New England Journal of Medicine. 2002; 346:235–242. [PubMed: 11807147]

11. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. New England Journal of Medicine. 2002; 346:1937–1947. [PubMed: 12075054]

12. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, et al. Stromal gene signatures in large-b-cell lymphomas. New England Journal of Medicine. 2008; 359:2313–2323. [PubMed: 19038878]

13. The international non-hodgkin's lymphoma prognostic factors project. a predictive model for aggressive non-hodgkin's lymphoma. New England Journal of Medicine. 1993; 329:987–994. [PubMed: 8141877]

14. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2012. URL http://www.R-project.org/

15. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996; 58(1):267–288.

16. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005; 67(2):301–320.

17. Dobbin KK, Simon R. Optimally splitting cases for training and testing high dimensional classifiers. BMC Medical Genomics. 2011; 4:31. [PubMed: 21477282]

18. Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, Dotson WD, Douglas MP, Berg AO. The evaluation of genomic applications in practice and prevention (egapp) initiative: methods of the egapp working group. Genetics in Medicine. 2009; 11(1):3–14. [PubMed: 18813139]

## 6. Appendix 1: Determination of β's coefficients that give rise to a true conditional risk θ

Assume there are ten biomarkers of interest and these markers are independently and identically distributed as a standard normal distribution. Assume further that the collective ability of the 10-dimensional biomarkers to accurately classify a patient into a good ($Y = 0$) or a poor ($Y = 1$) prognosis subgroup is determined by the following logistic regression model

$$log\left(\frac{p(y_i=1|x_{i1},\ldots,x_{i10})}{1-p(y_i=1|x_{i1},\ldots,x_{i10})}\right)=\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_{10}x_{i10} \quad (13)$$

The binary outcome for the *ith* patient, $y_i$, is thus drawn from a Bernoulli distribution with the probability of being in a poor prognosis subgroup:

$$p_i=p(y_i=1|x_{i1},\ldots,x_{10i})=\frac{e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{10}x_{i10}}}{1+e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{10}x_{i10}}} \quad (14)$$

Note that the ability of the prognostic signature to discriminate patient outcomes is influenced by the $\beta$'s coefficients in the logistic regression model. We wish to devise a data generating mechanism that gives rise to the true conditional risk $\theta$ in each of the following four scenarios:

- Null Hypothesis (NH): $\theta = 0.35$
- Alternative Hypothesis 1 (AH1): $\theta = 0.20$
- Alternative Hypothesis 2 (AH2): $\theta = 0.25$
- Alternative Hypothesis 3 (AH3): $\theta = 0.30$

There are infinitely many combinations of $\beta$'s values that would correspond to a specified $\theta$. To simplify, we fix the values of nine $\beta$'s coefficients and perform a fine grid search of the remaining $\beta$ coefficient (with an increment 0.02) such that the true conditional risk is equal to the desired value. This is possible by recognizing that $\theta$ decreases with increasing values of the remaining $\beta$. Specifically, in all scenarios we fix $\beta_4 = \beta_5 = \cdots = \beta_{10} = 0$. Further, we fix $\beta_1 = \beta_2 = 0.5$ for **NH** and **AH3** and $\beta_1 = \beta_2 = 1.3$ for **AH1** and **AH2** such that the only $\beta$ coefficient free to vary is $\beta_3$. The following steps are then followed to search for the solution of $\beta_3$ in each scenario:

1. For each fixed value of $\beta_3$, simulate a very large dataset $D_1 = \{(x_{i1}, \ldots, x_{i10}, y_i), i = 1, \ldots, M\}$ (e.g. $M = 10, 000$) which represents the "universe" in which the conditional risk of a signature will be evaluated.
2. For the same fixed $\beta_3$, simulate 1,000 datasets $D_2 = \{(x_{i1}, \ldots, x_{i10}, y_i), i = 1, \ldots, n_1\}$ where $n_1$ represents the sample size in Stage 1 of the TACO procedure.
3. For each simulated dataset $D_2$, build a prognostic signature, identify the cutoff based on the procedure described in step 1-2 of Section 2.2, and estimate the conditional risk of the signature (including the cutoff) in the "universe" $D_1$.
4. Average the estimated conditional risks across 1,000 simulated $D_2$ datasets.
5. If the average conditional risk in step 4 exceeds the targeted $\theta$, increment $\beta_3$ by a small amount (e.g. we used 0.02 in our simulation studies) and repeat above steps 1-4. The search stops when the smallest $\beta_3$ value that yields an average conditional risk smaller than $\theta$ is found.

## 7. Appendix 2: Translating PPV and NPV into conditional risk θ

In calculating the value of $\theta$ under $H_0$ and $H_A$, we first set up a 2 × 2 table as follows in which the rows represent the true disease status (Y = 0 if event-free, Y = 1 if event) and the columns represent the predicted disease status based on the signature (X = 0 if predicted to be event-free, X = 1 if predicted to be an event).

| | | Predicted Status | | |
|---|---|---|---|---|
| | | X = 0 | X = 1 | |
| True Status | Y = 0 | *a* | *b* | (*a* + *b*) |
| | Y = 1 | *c* | *d* | (*c* + *d*) |
| | | (*a* + *c*) | (*b* + *d*) | (*a* + *b* + *c* + *d*) |

Let $\pi_1$, $\pi_2$ and *w* denote, respectively, the PPV, NPV, and the prevalence of disease, then

$$\left\{\begin{array}{c}\pi_1=d/(b+d)\\ \pi_2=a/(a+c)\\ w=(c+d)/(a+b+c+d)\end{array}\right. \quad (15)$$

With some algebraic manipulations, we have

$$\left\{\begin{array}{c}b=\left(\frac{1-\pi_1}{\pi_1}\right)\ d\\ c=\left(\frac{1-\pi_2}{\pi_2}\right)\ a\\ a=\left(\frac{\pi_2}{1-\pi_2-\omega}\right)\ \left(\frac{w-\pi_1}{1-\pi_1}\right)\ b\end{array}\right. \quad (16)$$

Re-expressing *a*, *b* and *c* in terms of *d* in (16) gives

$$\left\{\begin{array}{c}a=\left[\frac{\pi_2(w-\pi_1)}{\pi_1(1-\pi_2-w)}\right]\ d\\ b=\left(\frac{1-\pi_1}{\pi_1}\right)\ d\\ c=\left[\frac{(1-\pi_2)\ (w-\pi_1)}{\pi_1(1-\pi_2-w)}\right]\ d\end{array}\right. \quad (17)$$

Note that since *a*, *b*, *c*, *d* > 0, the system of equations in (17) requires that either Condition (i): $1–\pi_2 < w < \pi_1$, or Condition (ii): $\pi_1 < w < 1–\pi_2$ to be satisfied.

Theoretical bounds for *w* depend on the disposition of $\pi_1$ and $\pi_2$. Specifically, consider the following scenarios:

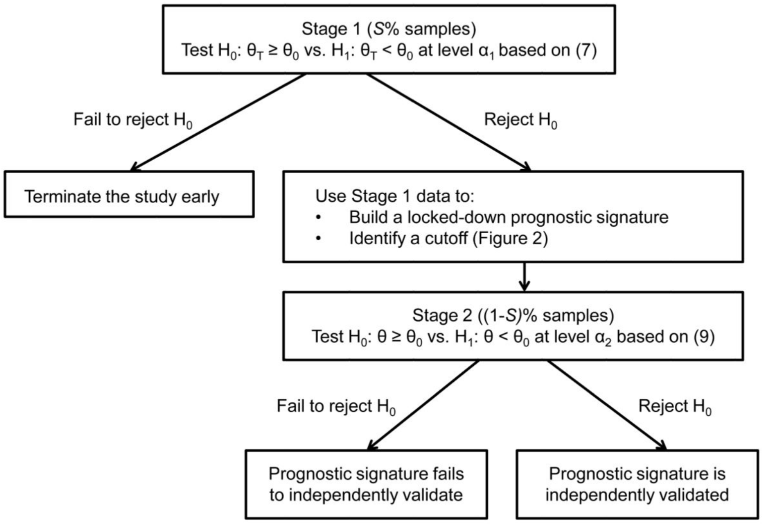
| | |
|---|---|
| **S1** | $\pi_1,\ \pi_2 \geq 0.5$ |
| **S2** | $\pi_i < 0.5 < \pi_j(i \neq = j)$ and $\pi_1 + \pi_2 > 1$ |
| **S3** | $\pi_1,\ \pi_2 \leq 0.5$ |
| **S4** | $\pi_i < 0.5 < \pi_j(i = \neq j)$ and $\pi_1 + \pi_2 < 1$ |

It can be shown that *w* is bounded by Condition (i): $1–\pi_2 < w < \pi_1$ for S1 or S2, and *w* is bounded by Condition (ii): $\pi_1 < w < 1–\pi_2$ for S3 or S4.

Finally, using expressions in (17), the conditional risk $\theta$ can be expressed as a function of $\pi_1$, $\pi_2$ and $w$ as follows:

$$\theta=\frac{b+c}{a+b+c+d}=\frac{(1-\pi_1)\,(1-\pi_2-w)+(1-\pi_2)\,(w-\pi_1)}{(1-\pi_1-\pi_2)}, \quad (18)$$

where $1-\pi_2 < w < \pi_1$ if S1 or S2, and $\pi_1 < w < 1-\pi_2$ if S3 or S4. Note a special case when $\pi_1 = \pi_2 (\neq 0.50)$, it can be easily shown that $\theta$ in (18) reduces to $(1-\pi_1)$.



Stage 1 ($S$% samples)
Test $H_0$: $\theta_T \geq \theta_0$ vs. $H_1$: $\theta_T < \theta_0$ at level $\alpha_1$ based on (7)
Fail to reject $H_0$
Reject $H_0$
Terminate the study early
Use Stage 1 data to:
• Build a locked-down prognostic signature
• Identify a cutoff (Figure 2)
Stage 2 ((1-$S$)% samples)
Test $H_0$: $\theta \geq \theta_0$ vs. $H_1$: $\theta < \theta_0$ at level $\alpha_2$ based on (9)
Fail to reject $H_0$
Reject $H_0$
Prognostic signature fails to independently validate
Prognostic signature is independently validated

**Figure 1.**

Flow chart of the two-stage adaptive cutoff (TACO) procedure. $\theta_0$ denote a pre-specified error rate below which a model would be deemed acceptable. The significance levels $\alpha_2$ and $\alpha_2$ are chosen so that ($\alpha_1 \times \alpha_2$) does not exceed some pre-specified type I error $\alpha$ (e.g. 0.05).

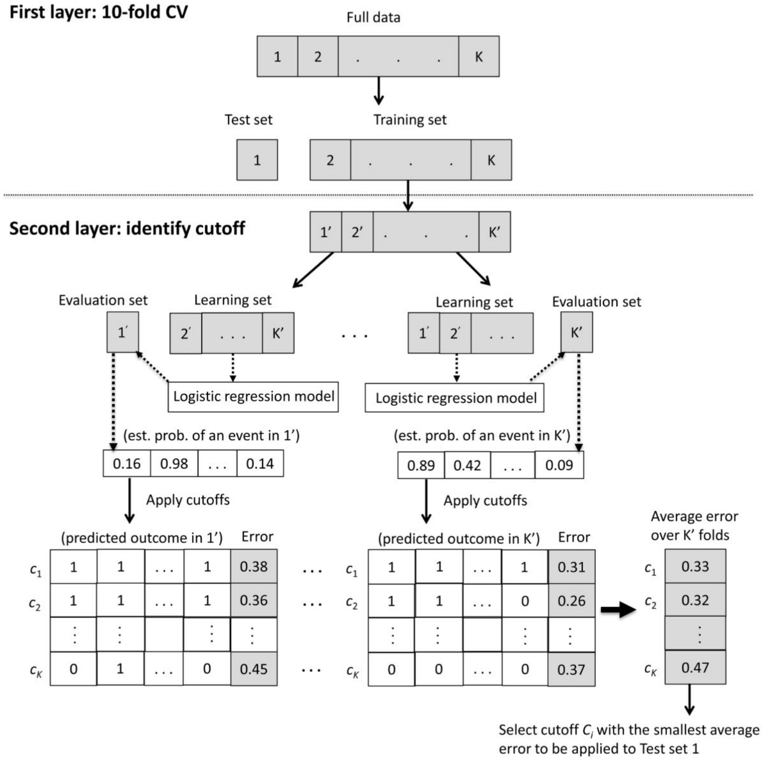


**Figure 2.**

A diagram of the two-layer procedure in Stage 1 of the TACO design.

## Table 1

Operating characteristics of the TACO design under various hypotheses

|  |  |  |  | Null Hypothesis: $\theta = 0.35$ |  |  |  | Alternative Hypothesis 1: $\theta = 0.20$ |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $S$(%) | $n_1$ | $n_2$ | $\beta_3$ | % Early Stop | Type I error | E(SS) | $\beta_3$ | % Early Stop | Power | E(SS) |
| 200 | 25% | 50 | 150 | 1.26 | 0.77 | 0.02 | 97 | 3.38 | 0.06 | 0.93 | 191 |
|  | 50% | 100 | 100 | 0.90 | 0.80 | 0.03 | 120 | 2.26 | 0.01 | 0.98 | 199 |
|  | 75% | 150 | 50 | 0.76 | 0.84 | 0.02 | 158 | 1.82 | 0.001 | 0.92 | 200 |
| 300 | 25% | 75 | 225 | 1.02 | 0.83 | 0.02 | 114 | 2.54 | 0.02 | 0.97 | 295 |
|  | 50% | 150 | 150 | 0.76 | 0.81 | 0.03 | 179 | 1.82 | 0.002 | 0.99 | 300 |
|  | 75% | 225 | 75 | 0.66 | 0.77 | 0.04 | 242 | 1.78 | 0.001 | 0.98 | 300 |
| 400 | 25% | 100 | 300 | 0.90 | 0.80 | 0.02 | 159 | 2.26 | 0.004 | 0.99 | 399 |
|  | 50% | 200 | 200 | 0.70 | 0.79 | 0.03 | 243 | 1.82 | 0 | 1 | 400 |
|  | 75% | 300 | 100 | 0.62 | 0.81 | 0.04 | 319 | 1.70 | 0 | 1 | 400 |

## Table 2

Operating characteristics of the TACO design under various hypotheses

| | | | | **Alternative Hypothesis 2: $\theta = 0.25$** | | | | **Alternative Hypothesis 3: $\theta = 0.30$** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $S$**(%)** | $n_1$ | $n_2$ | $\beta_3$ | **% Early Stop** | **Power** | **E(SS)** | $\beta_3$ | **% Early Stop** | **Power** | **E(SS)** |
| 200 | 25% | 50 | 150 | 1.84 | 0.24 | 0.68 | 164 | 1.78 | 0.45 | 0.40 | 133 |
| | 50% | 100 | 100 | 0.94 | 0.08 | 0.85 | 192 | 1.38 | 0.34 | 0.43 | 166 |
| | 75% | 150 | 50 | 0.40 | 0.06 | 0.67 | 197 | 1.18 | 0.37 | 0.25 | 182 |
| 300 | 25% | 75 | 225 | 1.20 | 0.16 | 0.81 | 265 | 1.52 | 0.42 | 0.45 | 205 |
| | 50% | 150 | 150 | 0.40 | 0.05 | 0.90 | 293 | 1.18 | 0.40 | 0.39 | 241 |
| | 75% | 225 | 75 | 0.24 | 0.02 | 0.81 | 299 | 1.14 | 0.23 | 0.41 | 283 |
| 400 | 25% | 100 | 300 | 0.94 | 0.10 | 0.89 | 371 | 1.38 | 0.37 | 0.52 | 289 |
| | 50% | 200 | 200 | 0.40 | 0.02 | 0.96 | 397 | 1.18 | 0.22 | 0.59 | 357 |
| | 75% | 300 | 100 | 1.24* | 0.01 | 0.92 | 399 | 1.10 | 0.16 | 0.51 | 384 |

Note: The unusual $\beta_3$ value (*) may be due to the relatively small number of simulation runs (1,000 simulations).

## Table 3

The final locked-down prognostic signature based on Stage 1 data of the DLBCL dataset

| Coefficients | Estimate | Standard Error | p-value |
|---|---|---|---|
| Intercept | −8.48 | 2.97 | 0.004 |
| ECOG (> 1 vs. ≤1) | 1.61 | 0.58 | 0.005 |
| Subtype (GCB vs. ABC) | −1.74 | 0.61 | 0.004 |
| Stromal-2 (continuous) | 0.96 | 0.37 | 0.009 |