Journal of Nuclear Medicine, published on February 12, 2021 as doi:10.2967/jnumed.120.251520

Statistical Considerations in the Evaluation of Continuous Biomarkers

Mei-Yin C. Polley, PhD[1,2] and James J. Dignam, PhD[1,2]

[1]Department of Public Health Sciences, The University of Chicago

[2]NRG Oncology Statistics and Data Management Center

Corresponding Author:

Mei-Yin C. Polley, PhD

Associate Professor of Biostatistics

Email: mcpolley@uchicago.edu

Phone: 773-834-2215

Abstract

Discovery of biomarkers has been steadily increasing over the past decade. While a plethora of biomarkers have been reported in the biomedical literature, few have been sufficiently validated for broader clinical applications. One particular challenge that may have hindered the adoption of biomarkers into practice is the lack of reproducible biomarker *cutpoints*. In this article, we attempt to identify some common statistical issues related to biomarker cutpoint identification and provide guidance on proper evaluation, interpretation, and validation of such cutpoints. First, we illustrate how discretization of a continuous biomarker using sample percentiles results in significant information loss and should be avoided. Second, we review the popular 'minimal p-value approach' for cutpoint identification and show that this method results in highly unstable p-values and unduly increases the chance of significant findings when the biomarker is not associated with outcome. Third, we critically review a common analysis strategy by which the selected biomarker cutpoint is used to categorize patients into different risk categories and the difference in survival curves among these risk groups in the same dataset is claimed as the evidence supporting the biomarker's prognostic strength. We show that this method yields exaggerated p-value and overestimates the prognostic impact of the biomarker. We illustrate that the degree of the optimistic bias increases with the number of variables being considered in a risk model. Finally, we discuss methods to appropriately ascertain additional prognostic contribution of the new biomarker where standard prognostic factors already exist. Throughout the article, we use real examples in oncology to highlight relevant methodological issues and when appropriate, use simulations to illustrate more abstract statistical concepts.

Introduction

Recent advances in biotechnologies have made it possible to perform extensive biological characterizations of human diseases. These efforts have resulted in the discovery of a myriad of biomarkers and generated much excitement for their potential to guide patient care. Possible uses of biomarkers in research and clinical setting include individual risk stratification, disease monitoring, and guiding the use of specific treatment regimens. Despite the large volume of published articles in biomedical journals on newly identified biomarkers, very few of these have progressed to the point of being clinically actionable. Many biomarkers may appear promising in the initial research reports but fail to retain their utility in subsequent studies. One particular challenge that may have hindered the adoption of biomarkers into practice is the lack of reproducible biomarker *cutpoints*. To aid clinical decision-making, medical practitioners are accustomed to discretizing a biomarker measured on a quantitative scale into different risk categories based on some partition of the scale, commonly called cut-point(s). This is natural as it is desirable to define patient groups sharing similar expected prognosis (say, for treatment or surveillance), and an overly precise scale is not useful in this regard. However, frequently research reports lack sufficient details on how such cutpoints are identified. Moreover, naïve use of statistical methodology for cutpoint identification, invalid methods for analysis, and overconfidence in the reliability of cutpoint-defined risk groups have hampered the ability to compare results across different studies or to generalize the results to the larger disease population of interest in an unbiased fashion. Even in the same or similar disease setting, biomarker cutpoints reported are often inconsistent and irreproducible.

Our goal in this article is to highlight some common statistical issues that arise from biomarker cutpoint identification and to provide guidance on proper evaluation, interpretation, and validation of such cutpoints. First, we illustrate how discretization of a continuous biomarker using sample percentiles (for example, sample median) results in significant information loss and should be avoided. Second, we review a popular method for cutpoint identification which entails testing a range of cutpoint values and selecting the cutpoint that yields the smallest p-value (i.e. the minimal p-value approach). We show that this approach results in highly unstable p-values and is associated with a severely inflated false discovery rate (i.e. it unduly increases the chance of significant findings when the biomarker is not associated with outcome) and estimates of the biomarker effect that are biased (suggesting a larger effect than is actually present). Some methods for correcting the p-value and biomarker effect are referenced. Third, we critically review a common analysis strategy by which the selected biomarker cutpoint is used to categorize patients into different risk categories and the difference in survival curves among these risk groups in the same dataset is claimed as the evidence supporting the biomarker's prognostic strength. We show that this method yields exaggerated p-value and overestimates the prognostic impact of the biomarker. We illustrate in a simulation study that the degree of the optimistic bias increases with the number of variables being considered in a risk model. We expand from that point to special considerations for biomarker cutpoints in disease settings where standard prognostic factors already exist. We discuss methods to appropriately ascertain additional prognostic contribution of the new biomarker and the relevance of cutpoint determination in such context.

Throughout the article, we use real examples in oncology to highlight relevant methodological issues and when appropriate, use simulations to illustrate more abstract statistical concepts. Although the examples here primarily pertain to molecular biomarkers, these principles generally apply to other type of biomarkers (e.g. imaging biomarkers, blood biomarkers) so long as they are measured on a continuous scale. Similarly, these statistical principles can be readily adopted to other non-cancer disciplines in biomedical research.

Statistical Pitfalls in Biomarker Cutpoint Search and Analysis

Loss of information due to discretization

A popular strategy for handling continuous biomarkers is to convert them into discrete variables by grouping patients into distinct risk subgroups (for example, by splitting the patients based on sample percentiles of the biomarker values). This type of categorization avoids the need to make strong assumptions about the functional relationship between the biomarker and outcome. In reality, however, the true relation between a continuous biomarker and outcome is almost always smooth. Such relations are seldom characterized by an abrupt 'jump' at a given biomarker value. **Figure 1** illustrates two true relationships between the biomarker M and some continuous outcome of interest (e.g. patient survival) - one linear (**Figure 1a)** - green line) and one quadratic (**Figure 1b)**). In **Figure 1a)**, the risk of death increases linearly with values of M. In **Figure 1b)**, the risk of death decreases with M up to the point *m*, but increases linearly after *m*. Dichotomy of biomarkers into two patient groups assumes that a discontinuity in the risk occurs at some biomarker value and that the relationship

between the biomarker and outcome is flat for patients whose biomarker values are within the same intervals, as defined by the point of dichotomy. For example, consider the two orange dashed lines in **Figure 1a)**, representing two groups of patients with distinct risks defined by the cutpoint *c*. Such dichotomy presumes that there is a notable change in prognosis at the cutpoint in that patients whose biomarker values are below *c* confer the same risk which is lower (by the magnitude Δ) than those patients whose biomarker values exceed *c*. This risk stratification based on dichotomizing the biomarker clearly does not adequately reflect the true linear relationship between the biomarker and outcome. In addition, categorizing a continuous biomarker causes considerable loss of valuable information, which may in turn increase the chance of missing a real association. For example, the dichotomy of biomarker values in **Figure 1a)** designates patients into two risk groups (e.g. “low risk” vs. “high risk”). Therefore, a patient whose true risk is highest in the “high risk” subgroup (i.e. whose biomarker value is *c**) is assumed to have the same prognosis as a patient whose true risk is lowest in the same risk category (whose biomarker value is *c*).

Consider the following example. In early stage triple-negative breast cancer, elevated level of neutrophil to lymphocyte ratio (NLR), a peripheral indicator of systematic inflammation, has been shown to be associated with poor outcomes in small retrospective patient cohorts [1-3]. In a recent report by investigators at the Mayo Clinic, six hundred and five patients who underwent breast surgery for stage I-III breast cancer between 1985 and 2012 at Mayo Clinic and met the criteria for triple-negative breast cancer phenotype were identified [4]. Clinicopathologic factors and biomarkers (including NLR) were collected to assess their impact on clinical outcomes. In that

study, the median NLR was 2.52. A common strategy of handling continuous biomarkers such as NLR is to dichotomize the biomarker at its sample median since this guarantees equal sample size between the 'low' and 'high' risk groups. **Figure 2 a)** displays the relationship between NLR and patient survival using restricted spline [5]. Clearly, there is a non-linear relationship between NLR and risk of death. If we apply a quadratic transformation to NLR (by including a continuous NLR term and its squared term in the regression model), there is a highly significant statistical association between NLR and risk of death (likelihood ratio test = 37.91, $p < 0.0001$). However, this association dissipates if NLR is dichotomized at its sample median (see **Figure 2b)**; HR = 1.16 (95% CI: 0.89 – 1.52), log-rank p = 0.27). This example illustrates that arbitrary dichotomization of a continuous biomarker can distort its true relationship with outcome, resulting in significant information loss. Note that the HR estimate of 1.16 suggests that patients whose NLR is above the sample median (NLR = "high") confers a 16% increase in the hazard of death, comapred to those with "low" NLR. In contrast, if NLR is modeled as a continuous variable in the Cox regression model, the resultant HR is 1.23 suggesting that a *one-unit increase* in NLR is associated with 23% increase in the hazard of death. When interpreting the prognostic effect of a continuous biomarker, it is important to pay attention to its range (in the Mayo TNBC dataset, NLR ranges from 0.14 to 10.50) since how 'large' a one-unit increase is relevant to the underlying scale of the biomarker.

Due to the haphazard discretization of continuous biomarkers, the literature is plagued with biomarker cutpoints that are rarely reproducible. This makes comparison of biomarker effects across different studies impossible. For example, S-phase fraction

(SPF), the percentage of tumor cells in the S phase obtained by cell cycle analysis, was of considerable scientific interest as a potential prognostic biomarker in breast cancer, but in a review by Altman et al. a wide range of SPF cutpoints from 2.6 to 15.0 have been reported as ‘optimal’ in the literature, rendering the effect of SPF inconsistent among studies [6]. Another example is the nuclear proliferation biomarker Ki67. Ki67 is of interest for various applications in research and clinical management of breast cancer. For instance, clinical decision-making regarding treatment options for breast cancer often replies on the application of a Ki67 cutpoint to classify patients into “Ki67-high” or “Ki67-low” risk groups. However, in a review of meta-analysis of 85 studies in 32,825 patients in early breast cancer, Stuart-Harris et al. reported that Ki67 cutpoints ranging from 0% to 28.6% have been investigated [7]. This lack of consensus regarding the ‘optimal’ cutpoint for Ki67 in various settings has hindered its ability to facilitate clinical decision making or direct comparisons of Ki67 results across laboratories and clinical trials [8].

In general, when the goal is to explore whether a biomarker is singly prognostic, it would be preferable not to categorize the biomarker at all. A preferred approach to characterizing the relationship between a continuous biomarker and time-to-event outcome is by modeling the biomarker as a continuous variable in a univariate Cox regression model without introducing any cutpoint. This method has considerable advantage of retaining valuable information in the data and will improve the ability to directly compare results from different studies. When linearity assumption (that is, the risk increases or decreases linearly as the biomarker increases) is called into question, modern statistical techniques such as regression splines or fractional polynomial

models can be used to effectively model non-linear relationships between values of the biomarker and risk [5, 9]; the relationship between biomarker values and risk is represented by the fitted regression function and its associated confidence bands. Cutpoints for the biomarker, if desired, can then be defined based on the nature of the relationship.

Cutpoint search via the minimal p-value approach

Another common approach for identifying biomarker cutpoint is to examine a range of biomarker values and select the cutpoint that yields the smallest p-value. Altman et al. referred to this method as the “minimum p-value approach” [6]. Several authors have demonstrated that this naïve approach is associated with a considerable inflation of the type I error due to the well-known problem of multiple testing [6, 9-10]. Using the NLR example above, **Figure 3 a)** displays the log-rank p-values (testing the association between dichotomized NLR and recurrence-free survival) based on a range of NLR cutpoints. We excluded the top and bottom 20% of NLR and used 200 cutpoints. The NLR cutpoint associated with the smallest p-value is 3.95. It can be seen that the p-values are highly unstable (range: 0 - 0.53) and minor change in the NRL cutpoint can lead to drastically different p-values. As such, if p-value was to be reported, some statistical adjustment for multiplicity is necessary. Altman described a formula to compute a corrected p-value [6]. When we apply this adjustment to the NLR example, the resulting p-value is 4.7*10(-5), substantially larger than the uncorrected p-value 0.14*10^(-6).

We conduct simulation studies to investigate the severity of type I error inflation and how type I error rate changes as a function of the number of cutpoints and sample size. Specifically, we simulate a continuous biomarker which follows a uniform distribution between 0 and 1 (the biomarker takes any value between 0 and 1 with equal probabilities) and a survival outcome that follow an exponential distribution with rate 0.0289 (translating to median survival of 24 months) with no censoring. Note that this data-generating mechanism ensures that the continuous biomarker and the survival outcome have no association. In each simulated dataset, we exclude 10% of smallest and of largest biomarker values as potential cutpoints, apply a fixed number of biomarker cutpoints, compute the 2-sided p-value from the log-rank test associated with each cutpoint, and identify the cutoff that yields the minimum p-value. We consider a variety of scenarios, varying the sample size (100, 300, 500) and number of biomarker cutpoints (50, 150, 300). For each sample size, 5,000 datasets are simulated as described above and the type I error (the percentage of simulations for which the minimal p-value is less than a nominal level of 5%) is recorded. The results of these simulations are shown in **Figure 4**. It can be seen that for a fixed number of cutpoints, the type I error hardly changes with the sample size. However, for a fixed sample size, the type I error increases with increasing number of biomarker cutpoints. For example, when sample size of 300, the type I error increases from 37.3% with 50 cutpoints to 43.3% with 300 cutpoints. Notably, in all scenarios considered, the type I errors exceed 37%. These simulations confirm that when a series of significance test is performed on the same dataset each with a pre-specified nominal type I error rate of, for example 5%, the minimal p-value approach leads to a global false discovery rate that may be much

higher than 5%. In particular, this approach may yield a ‘statistically significant’ result ($p < 0.05$) with a probability greater than 37% for a biomarker that has no association with outcome at all when the number of cutpoints tried exceeds 50.

Another problem with the minimum p-value approach is concerned with the estimation of the biomarker effect. Specifically, this approach gives an exaggerated sense of association between the biomarker and outcome. This is because when there is an association between the (continuous) biomarker and outcome, the p-values derived from the significance tests (e.g. log-rank) are associated with the effect estimates (e.g. hazard ratio). As such, the smallest p-value would correspond to the most extreme hazard ratio (HR) estimate (e.g., positive association for HR < 1; negative association for HR >1). **Figure 3 b)** illustrates the association between HR estimates and p-values using the NLR example. The minimal p-value corresponds to a HR estimate of 0.45 (that is, patients with NLR values above the cutpoint of 3.95 confer a 55% reduction in the hazard of death compared with patients whose NLR values are below 3.95) - this effect is overestimated. Several authors have proposed strategies to correct for the overestimation of the effect of a biomarker using the same dataset [11-12]. The best and clearly unbiased approach to estimating the biomarker effect is to apply the cutpoint identified from the current study to other independent datasets. This approach guarantees that no optimistic bias is introduced to the effect estimation by the data-derived cutpoint.

Comparison of clinical outcomes using data-driven cutpoint

Other methods exist for identifying cutpoints of continuous biomarkers. In radiology literature, for example, a common measure of discrimination for binary outcomes (e.g. alive vs. dead, cancer vs. non-cancer) is the receiver operating characteristic (ROC) curve. *Discrimination* quantifies how well a biomarker differentiates subjects at higher risk of having an event from those at lower risk. More specifically, a biomarker with good discrimination would predict having an event with a higher probability among subjects who will develop an event. The ROC curve consists of plotting the pairs of sensitivity and (1-specificity) [13] with a natural tradeoff between these two quantities. The area under the ROC curve (AUC) is a measure of discrimination, with values close to 0.5 indicating the discrimination no better than chance alone (i.e. having equal probability of classifying subjects with vs. those without events to an 'event' category). AUC values close to 0 or 1 indicate that the biomarker almost always correctly predict subject's event status. Many methods are available for identifying a biomarker cutpoint that 'optimizes' its discriminant performance. The index proposed by Youden [14], defined as (sensitivity + specificity – 1), is an example. This index, ranging from 0 to 1, gives equal weight to false positive and false negative values. Graphically, the Youden's index represents the height above the 45-degree chance line (representing AUC = 0.5). The biomarker value associated with the largest Youden's index may be chosen as the 'optimal' cutpoint. Other methods exist for identifying cutpoints from the ROC [15].

In some disease settings, a multitude of biomarkers or clinicopathologic variable may be of prognostic potential. It is sometimes useful to combine these prognostic factors via statistical modeling strategy (e.g. logistic regression model for binary

endpoints, Cox proportional hazards model for time-to-event endpoints) to form a risk system (also sometimes referred to as a prognostic signature). For individual patients, the composite *risk score* (or prognostic index) can be computed by adding up the 'weighted' factors (with the weights being the estimated regression coefficients). The prognostic indices then represents a new variable combining the information from all prognostic factors that can be used for prognostication purposes. For example, Haybittle et al. developed a prognostic index, the Nottingham Prognostic Index (NPI), from a Cox proportional hazards model for patients with primary operable breast cancer. The prognostic index for each patient was expressed as a linear function *0.17 x (tumor size in cm) + 0.76 x (lymph-node stage) + 0.81 x (tumor grade)*, where tumor grade = 1 or 2 or 3, and lymph-node stage = 1 or 2 or 3; see Haybittle et al. for the definition of lymph-node stage [16]. The larger the value of NPI, the worse the patient prognosis. Three risk groups were then defined based on the range of the NPI's. A cutpoint for the continuous prognostic index can be chosen based on the ROC methodology as described above for a single continuous biomarker.

In practice, it is not uncommon for investigators to use the selected cutpoint of the model score to categorize patients and then compare the nonparametric survival curves of the two risk groups via the log-rank test using the same dataset. This approach tends to exaggerate the p-value and overestimates the effect of the model. Optimizing a biomarker or risk model based on outcome and then claiming good discriminatory value based on the survival curves on that same dataset is a prevalent problem in the medical literature. Simon et al. referred to the performance measure of a risk model (e.g. discrimination) evaluated using the same data for some form of

“optimization” (for example, cutpoint selection or model development) as “resubstitution statistics” [17]. The separation between Kaplan-Meier curves for low- and high- risk patients as defined by the cutpoint derived from the same dataset is an example of resubstitution statistics. Simon et al. maintain the importance of separating the data used for any aspect of optimization from the data used for performance assessment. Some complex statistical approaches (such as bootstrap, jackknife, and permutation tests) may be useful in providing a more unbiased assessment of the true utility of the dichotomized biomarker. These methods belong to a class of “re-sampling” methods [18]. One simple form of re-sampling method is the sample split. With sample split, one portion of the dataset is used for cutpoint optimization or model development and the remaining (independent) data are used to evaluate the discriminatory power of the biomarker or model developed with the first portion [19]. It should be recognized, however, that resampling methods represent interval validation and do not reflect many sources of variabilities present in broader practice settings. Therefore, large independent studies will still be required to confirm the results.

Consider the studies by Lin et al. [20] and Casasnovas et al. [21], both aiming to assess the prognostic value of early Fluorine-18 fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) using standardized uptake values (SUV) in patients with diffuse large B-cell lymphoma (DLBCL). A clinical endpoint of interest was event-free survival (EFS), defined as months from study enrollment until first evidence of progression, relapse, or death due to any cause in Lin et al. In order to apply standard ROC methodology, the investigators first replaced the continuous EFS variable with a binary one (i.e. event versus no-event). Of note, the approach of using a binary

outcome status (e.g. vital status = dead or alive) in place of a continuous outcome variable such as EFS suffers the drawback of information loss as it ignores the varying length of follow-up among patients. For example, a patient who has survived for 5 years would have the same binary outcome status as another patient who has survived for 1 year (i.e. for both patients, the vital status would be “alive”). Note that statistical methods exist that extend the standard ROC methodology to accommodate time-to-event outcomes such as EFS [22]. The investigators then applied the ROC methodology to identify an optimal cutpoint for SUV (65.7% in Lin et al. and 66% for Casasnovas et al.). Study subjects were then categorized into two risk groups based on the selected cutpoint, and the ‘significant’ p-values from log-rank test ($p = 0.028$ in Lin et al. and $p < 0.0001$ for Casasnovas et al.) and notable separation in the Kaplan-Meier survival curves (Figure 2 b) in both studies) were cited as strong evidence supporting the prognostic value of SUV. Again, because the cutpoint was pre-selected to distinguish outcome by some measure, the resultant estimated biomarker effect and p-value obtained from the same dataset are optimistically biased and should not be interpreted as a confirmation of SUV’s prognostic utility.

In general, the magnitude of resubstitution bias is further exacerbated with increasing number of covariates in the risk model. This problem is known as *overfitting* in that a complex statistical model containing a sufficiently large number of variables having no true association with clinical outcome at all can spuriously provide an excellent fit to a small dataset. We performed a simulation to illustrate the bias in the estimated discrimination. We simulated a “sample” dataset with n = 200 patients and a “population” dataset with a very large size (N = 10,000). The latter represents the target

population at large and hence the performance of the model evaluated in the population is regarded as the true value. In each simulated dataset, we randomly generated a set of k continuous variables, denoted as $X = (X_1, X_2, \ldots, X_k)$, each following a standard normal distribution (with mean 0 and standard deviation 1). We assumed that 2 out of the k variables are associated with the binary endpoint Y. Specifically, the correlation between $(X_1, X_2, \ldots, X_k)$ and Y is induced by a multivariable logistic regression with intercept 0 and regression coefficients $B = (\beta_1 = 1.2, \beta_2 = 1.2, \beta_3 = 0, \ldots, \beta_k = 0)$. Correspondingly, the association between (X_1, X_2) and Y is characterized by an odds ratio (OR) of exp(1.2) = 3.32 whereas the remaining $(k\text{-}2)$ variables have no association with the outcome (i.e. OR = 1). We considered two scenarios: k = 5 (small number of biomarkers) and k = 50 (large number of biomarkers). For each k, we generated 1,000 datasets as described above and compared the distributions of AUC between the sample datasets and the population datasets. To arrive at the AUC estimate in a sample dataset, we fit a multivariable regression model of X on Y and obtained k regression coefficients estimates. The prognostic scores for individual patients were calculated as the linear combination of the variables weighted by the regression coefficients. The AUC was then estimated from the ROC for the new continuous score variable. Note that the regression model was only constructed using the sample dataset; the resultant regression coefficients were then fixed and applied to the population dataset to obtain individual prognostic scores and the “true” AUC value (i.e. with no further model building or refinement).

Figure 5 displays side-by-side boxplots of the distributions of AUC’s from the simulated sample datasets and the population datasets for k = 5 (left) and k = 50 (right).

It can been seen that when $k = 5$, AUC's were slightly biased upward in the sample distribution compared to the true population (median: 0.75 and 0.73 for samples and populations, respectively). The degree of optimistic bias increases drastically when the number of variables increases to 50 (median: 0.87 and 0.64 for samples and populations, respectively). This simulation exercise underscores the fact that the performance of a risk model is overestimated when the evaluation is performed in the same dataset used to construct the model, and the degree of the optimistic bias increases with the number of variables in the model. These results highlight the importance of evaluating the performance of a risk model in dataset that are independent from that used for model development.

Biomarker Cutpoint in the Presence of Established Prognostic Factors

For many cancers, certain prognostic factors are known and well established. For example, tumor size and the number of positive lymph modes are well-known prognostic factors in breast cancer. For patients with advanced non-Hodgkin's lymphoma, the International Prognostic Index (IPI) was a risk system developed to predict patient survival [23]. The components of IPI were based on clinical features including age, tumor stage, serum lactate dehydrogenase concentration, performance status, and the number of extranodal disease sites that are easy to measure and prognostically important. In these settings, it is more pertinent to determine whether a new biomarker adds additional prognostic information to that already provided by standard prognostic factors alone. Statistical models such as Cox's proportional hazards regression model are often used to study the joint prognostic influence of multiple factors. To assess the independent prognostic influence of the new biomarker

above and beyond recognized factors, one (reduced) multivariable model can be fitted containing only the standard factors and one (full) multivariable model can be fitted that simultaneously contain the new biomarker and standard factors. The difference in how well the two nested models fit the data provides a measure of statistical significance of whether the new factor contains additional prognostic information (e.g. via the likelihood ratio test) [24]. If there are multiple 'new' factors, this approach accounts for the number of new variables in the calculation of statistical significance. For example, Cheang et al. studied the additional prognostic information of a five-biomarker panel (estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (Her2), EGFR, and cytokeratin 5/6) above and beyond a three-biomarker panel (ER, PR, and Her2) in the presence of standard clinical variables for predicting breast cancer death-specific survival [25]. To test the statistical significance of the two additional biomarkers, two Cox regression models were fitted and a likelihood ratio test of the difference between the two models was used to evaluate the additional prognostic contribution of EGFR and cytokeratin 5/6.

When the cutpoint of a biomarker is pre-selected based on clinical outcome (e.g. via the minimal p-value approach or the ROC methodology), the corresponding dichotomized biomarker will impart an inflated effect in the multivariable regression model and thus diminishing the relative importance of other known prognostic factors. It is important to note that display of Kaplan-Meier curves showing the difference in survival between risk groups correspond to univariate statistical tests (e.g. log-rank), and thus does not indicate the effect of the biomarker after accounting for the other variables that may influence survival. In fact, in the presence of existing prognostic

factors, determination of a cutpoint for the new biomarker alone is not as relevant. Instead, a more holistic approach would be to develop a prognostic model incorporating both known prognostic factors and the new biomarker. Prognostic categories can then be defined based on the model-predicted prognostic indices of individual patients. For example, Paik et al. developed the *Oncotype Dx* assay, a 21-gene recurrence score (RS), to quantify the likelihood of distant recurrence in women with node-negative, estrogen-receptor positive breast cancer who have been treated with tamoxifen [26]. The cutpoints were determined on the basis of the results of NSABP trial B-20 and validated using trial data B-14. The cutpoints classify patients into three risk categories base on predicted 10-year distant recurrence rate: low risk (RS < 18), intermediate risk (18 <= RS < 31), and high risk (>= 31). The authors also demonstrated that the model based on age, tumor size, and recurrence score provided significantly independent prognostic information compared to the model including age and tumor size alone ($p<0.001$ by the likelihood ratio test).

Conclusions

Discovery of biomarkers has been steadily increasing over the past decade. While a plethora of biomarkers and associated cutpoints have been reported in the biomedical literature, few have been sufficiently validated for broader clinical applications. In contrast to the abundance of classical clinical trial principles for guiding the design, conduct, analysis, and reporting of studies, relatively fewer guidelines exist for biomarker research [27-28]. In this article, we have attempted to identify some common methodological issues related to biomarker cutpoint identification and evaluation. We strongly advocate that discretization of continuous biomarkers be avoided. If cutpoint

identification is performed, it should be handled with statistical care. Biased resubstitution should either not be reported or be clearly noted as an unreliable representation of the true discriminant value of the biomarker. When feasible, large independent datasets are ideal for confirmation of the prognostic value of the biomarker and its cutpoint. A schema for the consideration of biomarker analysis and cutpoint evaluation is proposed in **Figure 6**. We hope that the discussions here will draw attention to critical statistical issues associated with development and evaluation of biomarker cutpoints and will in turn help improve methodological rigor in this line of research.

Funding

This work was supported by the National Cancer Institute of the National Institutes of Health under Award Number U10 CA180822 (NRG Oncology - Statistics and Data Management Center) and P50 CA116201 (Mayo Clinic Breast Cancer Specialized Program of Research Excellence).

Disclosure

The authors have no conflict of interest to disclose.

References

[1] Chae S, Kang KM, Kim HJ, et al. Neutrophil-lymphocyte ratio predicts response to chemotherapy in triple-negative breast cancer. *Curr Oncol*. 2018;25(2):e113-e119.

[2] Patel DA, Xi J, Luo J, et al. Neutrophil-to-lymphocyte ratio as a predictor of survival in patients with triple-negative breast cancer. *Breast Cancer Res Treat*. 2019;174(2):443-452.

[3] Pistelli M, De Lisa M, Ballatore Z, et al. Pre-treatment neutrophil to lymphocyte ratio may be a useful tool in predicting survival in early triple negative breast cancer patients. *BMC Cancer*. 2015;15:195.

[4] Leon-Ferre RA, Polley MY, Liu H, et al. Impact of histopathology, tumor-infiltrating lymphocytes, and adjuvant chemotherapy on prognosis of triple-negative breast cancer. *Breast Cancer Res Treat*. 2018;167(1):89-99.

[5] Wahba G. Spline Models for Observational Data. SIAM. 1990.

[6] Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst*. 1994;86(11):829-835.

[7] Stuart-Harris R, Caldas C, Pinder SE, Pharoah P. Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. *Breast*. 2008;17(4):323-334.

[8] Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *J Natl Cancer Inst*. 2013;105(24):1897-1906.

[9] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer*. 1994;69(6):979-985.

[10] Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? *Breast Cancer Res Treat*. 1992;22(3):197-206.

[11] Schumacher M, Holländer N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building. *Stat Med*. 1997;16(24):2813-2827.

[12] Holländer N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Stat Med*. 2004;23(11):1701-1713.

[13] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

[14] Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35.

[15] Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7):670-5.

[16] Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. *Br J Cancer*. 1982;45(3):361-366.[17] Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst*. 2010;102(7):464-474.

[18] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21(15):3301-3307.

[19] Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep*. 1985;69(10):1071-1077.

[20] Lin C, Itti E, Haioun C, et al. Early 18F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. *J Nucl Med*. 2007;48(10):1626-1632.

[21] Casasnovas RO, Meignan M, Berriolo-Riedinger A, et al. SUVmax reduction improves early prognosis value of interim positron emission tomography scans in diffuse large B-cell lymphoma. *Blood*. 2011;118(1):37-43.

[22] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56(2):337-44.

[23] International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med*. 1993;329(14):987-994.

[24] David W. Hosmer Jr., Stanley Lemeshow, Susanne May. Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Wiley. 2nd Edition.

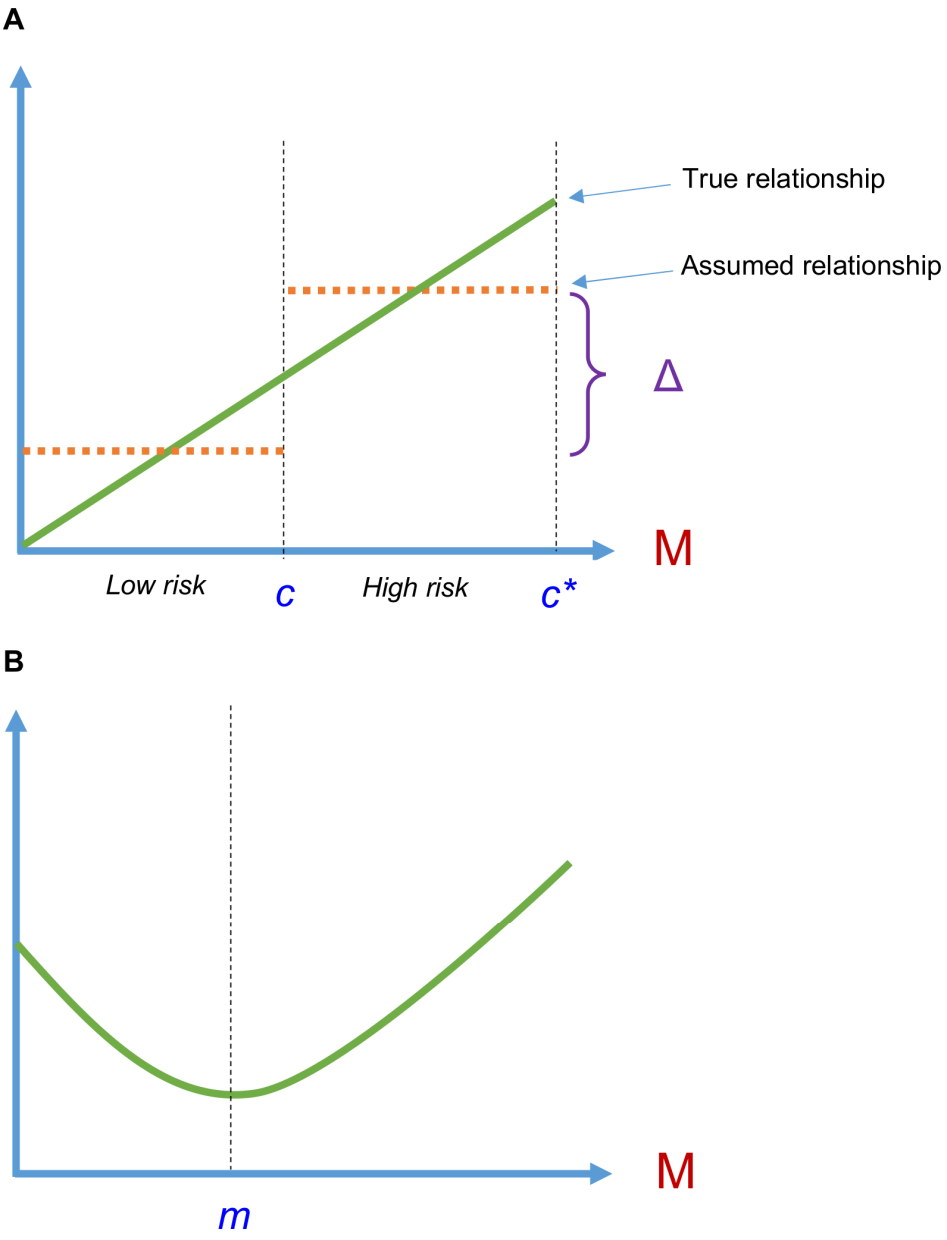
[25] Cheang MC, Voduc D, Bajdik C, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res*. 2008;14(5):1368-1376.

[26] Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351(27):2817-2826.

[27] McShane LM, Altman DG, Sauerbrei W, et al. Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer. 2005;93(4):387-91.

[28] McShane LM, Polley MY. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. Clin Trials. 2013;10(5):653-65.

Figure 1: Hypothetical relationship between biomarker (M) and clinical outcome



In Figure 1a), the green line depicts a linear relationship between biomarker (M) and outcome – the risk of outcome increases linearly with increasing biomarker values. The orange dashed line in Figure 1a) illustrates the effect of dichotomizing M. It assumes that a discontinuity in the risk occurs at a cutpoint *c* (patients whose biomarker values are below *c* confer the same risk, which is lower by the magnitude Δ than those patients whose biomarker values exceed c. Figure 1b) depicts a quadratic relationship between M and outcome - the risk of outcome decreases with M up to the point *m*, and increases linearly after *m*.

Figure 2: Dichotomy of a continuous biomarker (neutrophil lymphocyte ratio example)

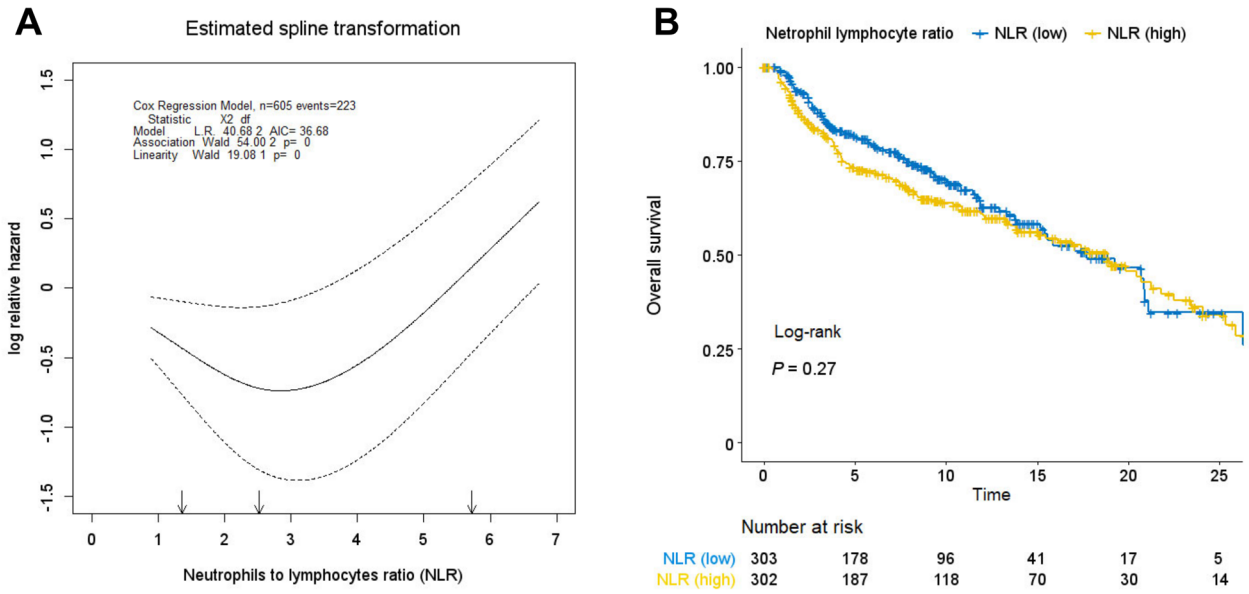


Figure 2 a): Non-linear relationship between neutrophil to lymphocyte ratio (NLR) and patient survival in the Mayo Clinic triple-negative breast cancer (TNBC) dataset using restricted spline method.

Figure 2b): The effect of dichotomizing NLR at its sample median - the association between NLR and survival is no longer significant (log-rank p-value = 0.27).

Figure 3: Minimal p-value approach (NLR example): (a) highly unstable p-values; b) inverse correlation between hazard ratios and p-values

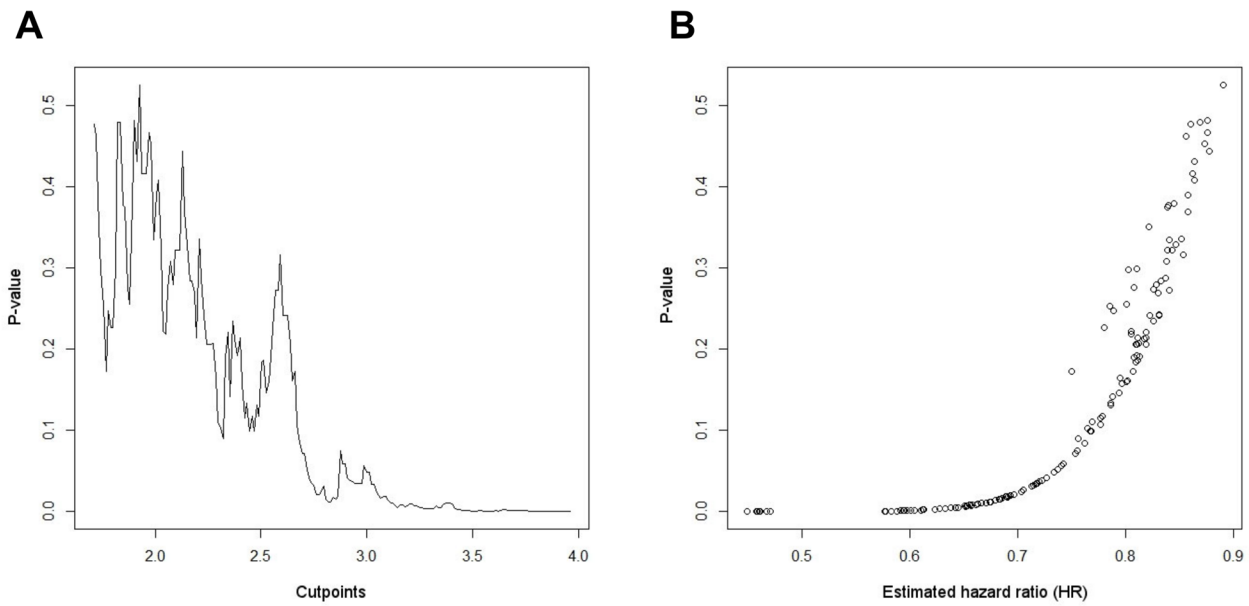
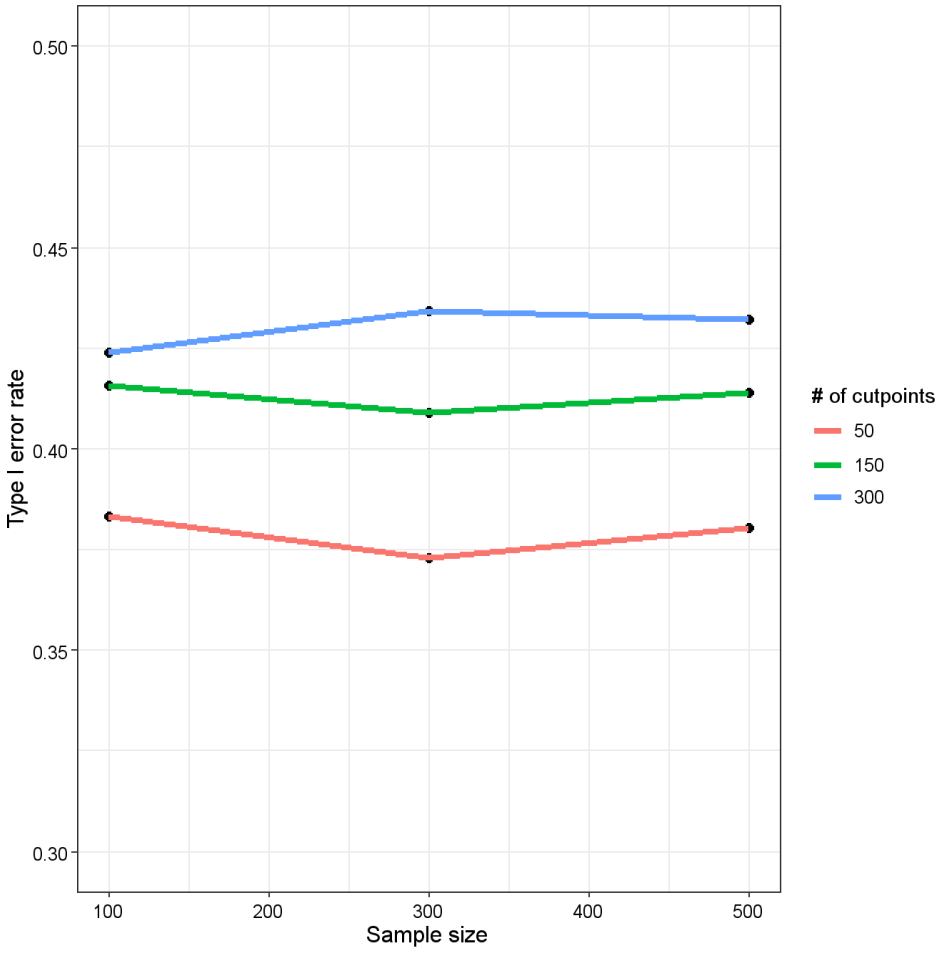


Figure 3 a): P-value of the log-rank test as a function of the cutpoint used for NLR in the Mayo Clinic triple-negative breast cancer (TNBC) dataset. The top and bottom 20% of the NLR values were excluded and 200 cutpoints were used. P-values are highly unstable within the NLR range.

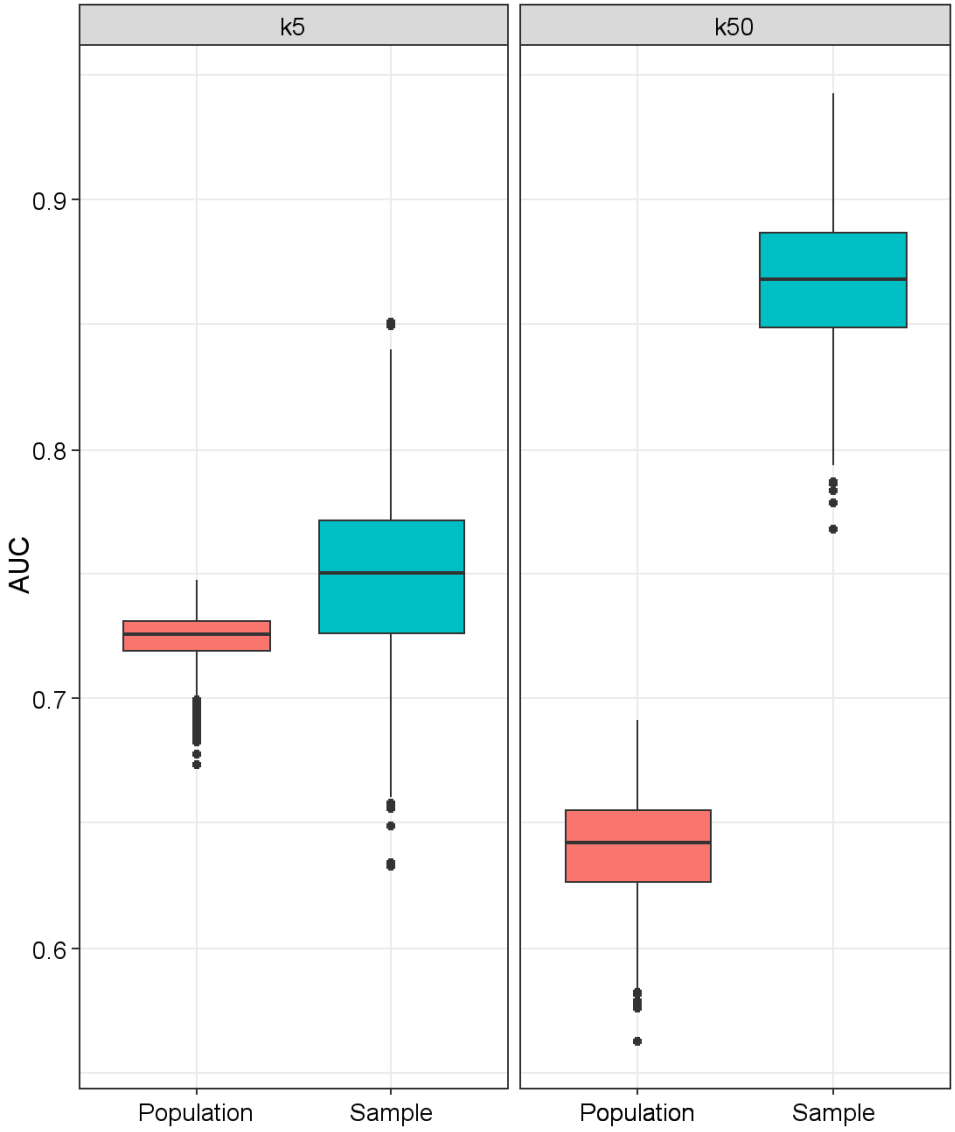
Figure 3b): Strong correlation between estimated hazard ratios (HR) and log-rank values for NLR in the Mayo Clinic TNBC dataset – smallest p-value corresponds to the most extreme HR estimate.

Figure 4: Type I error inflation associated with minimal p-value approach



Type I error as a function of number of cutpoints and sample size using the minimal p-value approach. In each simulation, 10% of smallest and largest biomarker values were not considered as potential cutpoints. A 2-sided p-value from the log-rank test was computed for each cutpoint applied. Each plotted point represents the percentage of 5,000 simulations for which the minimal p-value is less than the nominal 5% level based on the assumption that there is no association between the biomarker and time-to-event outcome (i.e. type I error). No censoring in the outcome was assumed.

Figure 5: Resubstitution bias in the area under the ROC curve



The effect of number of covariates (*k*) in the risk model on resubstitution bias. The "population" boxplot represents the true AUC distribution in the interested population at large. The 'sample' boxplot represents the distribution of AUC derived from the sample dataset used to construct the risk model. Each boxplot was based on 1,000 simulations. When *k* = 5 (left panel), there is a slight upward (optimistic) bias in the sample AUC distribution compared to the true population. The degree of optimistic bias increases drastically when *k* increases to 50 (right panel).

Figure 6: Schema for biomarker cutpoint analysis and evaluation