Tutorial

MDM Medical Decision Making

Medical Decision Making
2018, Vol. 38(8) 904–916

DOI: 10.1177/0272989X18801312
journals.sagepub.com/home/mdm
SAGE

A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making

Aasthaa Bansal and Patrick J. Heagerty

Abstract

Many medical decisions involve the use of dynamic information collected on individual patients toward predicting likely transitions in their future health status. If accurate predictions are developed, then a prognostic model can identify patients at greatest risk for future adverse events and may be used clinically to define populations appropriate for targeted intervention. In practice, a prognostic model is often used to guide decisions at multiple time points over the course of disease, and classification performance (i.e., sensitivity and specificity) for distinguishing high-risk v. low-risk individuals may vary over time as an individual's disease status and prognostic information change. In this tutorial, we detail contemporary statistical methods that can characterize the time-varying accuracy of prognostic survival models when used for dynamic decision making. Although statistical methods for evaluating prognostic models with simple binary outcomes are well established, methods appropriate for survival outcomes are less well known and require time-dependent extensions of sensitivity and specificity to fully characterize longitudinal biomarkers or models. The methods we review are particularly important in that they allow for appropriate handling of censored outcomes commonly encountered with event time data. We highlight the importance of determining whether clinical interest is in predicting cumulative (or prevalent) cases over a fixed future time interval v. predicting incident cases over a range of follow-up times and whether patient information is static or updated over time. We discuss implementation of time-dependent receiver operating characteristic approaches using relevant `R` statistical software packages. The statistical summaries are illustrated using a liver prognostic model to guide transplantation in primary biliary cirrhosis.

Keywords

dynamic information, prognosis, risk prediction, sensitivity, specificity

Date received: May 15, 2017; accepted: May 20, 2018

Many medical decisions involve using updated information on patients under surveillance to predict transitions in future health status, such as progression of disease or advancement to death. The goal is to use a patient's clinical characteristics to calculate the predicted risk of an event within a specified time period and to identify patients who are at high risk of experiencing an adverse event in the near future. If accurate predictions can be made, they could be used clinically to guide the choice and timing of interventions and enable timely action, such as starting specific preventive strategies or initiating aggressive treatments for high-risk individuals while

The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, School of Pharmacy, University of Washington, Seattle, WA (AB), and Department of Biostatistics, University of Washington, Seattle, WA (PJH). This work was presented at the Society for Medical Decision Making annual meeting, St. Louis, MO, 2015, and Joint Statistical Meetings, American Statistical Association, Seattle, WA, 2015. The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the PhRMA Foundation, the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) (under R01-HL072966) and by the National Center for Advancing Translational Sciences of the NIH (under UL1TR000423). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Corresponding Author

Aasthaa Bansal, The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, School of Pharmacy, University of Washington, H-375, Health Sciences Building, Box 357630, Seattle, WA 98195 (abansal@uw.edu).

sparing low-risk patients the side effects and costs of unnecessary intervention.

In practice, prognostic models are often used to make decisions at multiple time points over the course of patient follow-up. Consider disease screening settings, where predicted risk may be used to identify high-risk individuals as candidates for more frequent screening. Patient follow-up with updated clinical assessment also frequently occurs to monitor response to therapy. For example, a cancer patient who has previously undergone treatment and is predicted to be at substantial risk of disease recurrence may benefit from adjuvant therapy, whereas a low-risk patient may forego aggressive treatment. Finally, in an organ transplantation setting, the predicted risk of mortality may be used to guide prioritization and timing of donor organ transplantation.[1–4]

Traditional statistical models such as Cox regression focus on the prediction of disease or death times. However, underlying these standard methods are the concepts of a time-varying "risk set" of individuals, and associated time-specific "cases" or subjects who experience the clinical event (e.g., death) at a given time. At any time point, the set of individuals still alive and at risk of an event may be partitioned into imminent cases (individuals who experience the event in a defined future time frame) and current "controls" (individuals who do not yet experience the event). Ultimately, the goal of a prognostic model is to accurately predict event times or equivalently to distinguish between the time-specific cases and the controls at all follow-up times. Furthermore, in practice, an individual's disease status changes over time, and so does his or her prognostic information, such as laboratory measures updated in routine clinic visits. Accordingly, a model's ability to distinguish between cases and controls over time may also change, thus affecting its performance as a decision-making tool. For example, a prognostic model may accurately identify patients at high risk of death within 90 days, but it may have reduced accuracy for identifying later deaths.

Accuracy concepts of sensitivity and specificity are fundamental to clinical research and decision modeling. Only recently have statistical methods been developed that can generalize these traditionally cross-sectional accuracy concepts for application to the time-varying nature of disease states, and corresponding definitions of time-dependent sensitivity and specificity have been proposed for both prevalent and incident case definitions.[2,3] These new concepts and associated statistical methods are central to the evaluation of the time-varying performance of any potential prognostic model; they allow for the estimation of sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC) as functions of time, thus providing a detailed estimate of longitudinal model performance for use in practice. These methods are particularly important in that they allow for appropriate handling of right-censored outcomes commonly encountered with clinical event time data. Unfortunately, knowledge of these methods and the tools available to implement them remain limited, and investigators often resort to overly simplistic application of methods developed for binary outcomes, which can lead to biased estimates in the presence of censoring.[5,6]

Our goal in this tutorial is to demonstrate the use of modern statistical methods that address the following questions: how can the time-varying discrimination accuracy of a prognostic model be evaluated, how can the value of updated measurements be characterized, and how can different candidate models be directly compared? We highlight the importance of determining whether interest is in the fundamental epidemiologic concept of predicting cumulative (or prevalent) cases or in incident cases.

Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis

As an illustrative case study, we consider liver transplantation in primary biliary cirrhosis (PBC). PBC is an autoimmune disease in which the bile ducts are slowly destroyed, leading to liver failure in cases of advanced disease.[7] For selected patients with liver failure who are at high risk of death without transplantation, liver transplantation can be potentially life-saving. As a result, a number of prognostic models have been developed in PBC, with the goal of predicting survival probabilities and guiding decisions regarding transplantation.[8–14] Of these, the Mayo model is perhaps the most widely known,[8] with the more recent Model for End-stage Liver Disease (MELD) score[3] representing a refinement, but potentially suboptimal for use in PBC.[8] A unique characteristic of the Mayo model compared to other existing models is that it does not require liver biopsy. Instead, it is based on inexpensive, noninvasive, and readily available measurements. Additional variables from a biopsy, such as histologic stage, that are used in other models have been shown to not contribute substantially beyond the variables included in the Mayo model.[1]

We consider a well-known data set that comes from a randomized placebo-controlled trial for the treatment of PBC conducted at the Mayo Clinic between 1974 and 1984.[15] Dickson et al.[1] used these data to develop the

Mayo risk model that included patient age, total serum bilirubin and serum albumin concentrations, prothrombin time, and severity of edema. Murtaugh et al.[2] proposed a time-dependent version of this model that uses updated values of the prognostic variables. The Mayo model has been used for making individual-level decisions regarding the selection of patients for and timing of liver transplantation in PBC.[8] Decisions about transplantation are made repeatedly over time, by selecting patients who are most likely to die in a short time interval, such as 90 days, 6 months, or 1 year from the time of prediction. We will use the 5 main predictors of survival identified by Dickson et al.[1] to calculate the predicted risk of mortality within specified time periods and evaluate the accuracy of these predictions for prioritizing patients for transplantation.

Model Development

Model development typically takes place by splitting a data set into training and validation data that are used for model selection and evaluation, respectively. Using appropriate methods to avoid overfitting in the training data,[16–18] candidate biomarkers and variables are selected and combined, traditionally using a Cox proportional hazards regression model for survival outcomes.[19] One may use standard Cox regression with fixed coefficients and baseline covariates or even incorporate time-varying covariates, as well as time-varying coefficients, into the model.[20] Alternatively, one may use more flexible, modern machine-learning approaches, such as boosting, lasso, artificial neural networks, and random forests, especially in the presence of high-dimensional data.[21–27] Regardless of the chosen modeling approach, the ultimate prognostic model is then fixed and used in the validation data to provide patient predictions of the disease outcome (i.e., a risk score).

In this article, we are agnostic to model selection. We focus on methods for evaluating any single "biomarker," which may be a novel predictive measurement, such as a specific serum protein level measured in the laboratory or, more commonly, may be the risk score derived from a model that includes multiple factors (i.e., a *derived* biomarker or classifier). The approaches we discuss for evaluating a risk score in the validation data are independent of those used for model selection in the training data, in that they do not rely on the assumptions that may be necessary for the development of the risk score.

Given our focus on model evaluation, it is not our objective here to develop a new model as an alternative to the Mayo model. We simply demonstrate how to evaluate the time-varying performance of the existing Mayo risk score, as well as one variation of it where we omit a variable, to demonstrate a comparison of 2 candidate models.

Background: Standard Measures of Discrimination Accuracy

The traditional classification problem is based on a simple binary outcome, typically the presence or absence of disease. In classifying cases and controls as having disease or not, a marker is prone to 2 types of error: incorrectly classifying a case as not having disease, leading to delays in treatment, and, conversely, incorrectly classifying a control as having disease, subjecting the individual to unnecessary follow-up medical procedures. Investigators aim to minimize false-negative and false-positive errors by developing markers with high sensitivity (true-positive fraction [TPF]) and high specificity (1 minus false-positive fraction [FPF]), respectively.

By convention, larger marker values are assumed to be more indicative of disease (and if the opposite is true, the marker is transformed to fit the convention). For a continuous marker M and a fixed threshold c, we define

$$\begin{aligned}\text{sensitivity}(c) &= \mathrm{P}(M > c|\text{case}),\\ \text{specificity}(c) &= \mathrm{P}(M \le c|\text{control}).\end{aligned}$$

The receiver operating characteristic (ROC) curve is a standard tool that plots a continuous marker's sensitivity against 1 – specificity for all possible values of the threshold c.[28–31] Classification accuracy is most commonly summarized using the area under the ROC curve (AUC), which is the probability that a randomly chosen case has a higher marker value than a randomly chosen control:

$$AUC = \mathrm{P}\big(M_i > M_j | i = \text{case}, j = \text{control}\big).$$

Therefore, the AUC represents the marker's ability to rank cases above controls. An AUC of 0.5 indicates no discrimination between cases and controls, whereas an AUC of 1.0 indicates perfect discrimination.[31]

Time-Dependent Discrimination Accuracy

Implicit in the use of traditional diagnostic sensitivity and specificity are current-status definitions of disease. In settings of long-term follow-up, disease status changes with time, and precise definitions are necessary to include event (disease) timing in definitions of prognostic error rates. Within the past 2 decades,

time-dependent ROC curve methods that extend concepts of sensitivity and specificity and characterize prognostic accuracy for survival outcomes have been proposed in the statistical literature and adopted in practice. We review 2 such time-dependent approaches, which draw upon alternative fundamental case definitions: cumulative (or prevalent) cases and incident cases.

Cumulative (Prevalent) Cases/Dynamic Controls

Often interest lies in identifying individuals at risk of an adverse event within some fixed time frame. Recall, for example, decisions about donor liver allocation in the PBC setting being made by selecting patients who are most likely to die in a short time interval, such as 90 days, 6 months, or 1 year, from the time of prediction.

A natural extension of the standard cross-sectional definitions of sensitivity and specificity to the survival context, where disease state is time dependent, is to dichotomize the outcome at a selected time of interest, t (90 days, 6 months, or 1 year), and define cases as subjects who experience the event before time t and controls as those who remain event free beyond t.[32] More formally, we let T denote survival time and s denote the start time of case ascertainment (often $s = 0$ for baseline). Then, cumulative cases (C) may be defined as subjects who experience an event prior to t, or specifically as $T_i \in (s,t)$, and dynamic controls (D) as subjects who are event free at time t, $T_i > t$ (regardless of whether or not they experience the event at a later time). Then, for a fixed threshold c, time-dependent definitions for sensitivity and specificity follow[32,33]:

$$\text{sensitivity}^{C}(c|\text{start} = s, \text{stop} = t) = P(M > c|T \geq s,\ T \leq t)$$
$$\text{specificity}^{D}(c|\text{start} = s, \text{stop} = t) = P(M \leq c|T \geq s,\ T > t)$$

Let p represent a fixed FPF. Then, for fixed $\text{specificity}^{D}(c|s,t) = 1 - p$, the time-dependent ROC value is the corresponding value of $\text{sensitivity}^{C}(c|s,t)$, or $\text{ROC}_{s,t}{}^{C/D}(p)$. Correspondingly, the time-specific AUC is defined as the area under the time-specific ROC curve across all thresholds p:

$$\text{AUC}^{C/D}(s,t) = \int \text{ROC}_{s,t}{}^{C/D}(p)dp,$$

which can be shown to be equivalent to

$$\text{AUC}^{C/D}(s,t) = P(M_j > M_k|T_j \geq s,\ T_j \leq t,\ T_k \geq s,\ T_k > t).$$

Here, $\text{AUC}^{C/D}(s,t)$ is the probability that a random subject j who experiences an event before time t (case) has a larger marker value than a random subject k who remains event free through time t (control), assuming both subjects are event free at the start of follow-up, time s.

In the absence of censoring, the above dichotomization at time t is equivalent to using a simple derived binary disease outcome. However, when follow-up is incomplete, as is often the case with longitudinal data, censoring needs to be addressed and can be handled using nonparametric estimation of the bivariate distribution of (M,T).[32] (See online Appendix A for a description of estimation methods.) Estimation is based on (Z_i, δ_i), where Z_i is the observed follow-up time (i.e., the minimum of the survival time T_i and the right-censoring time C_i), and δ_i denotes the event indicator.

In this tutorial, we seek to characterize time-varying performance over a meaningful range of times. To this end, we suggest obtaining a sequence of accuracy assessments over time by defining cases as events occurring cumulatively in successive windows of time. Specifically, we subset data at a sequence of index times $s = t_1, t_2, \ldots, t_K$ to include only subjects who are event free at time t_k (i.e., $Z \geq t_k$, $k = 1, \ldots, K$). These index times can represent any time points of interest and do not have to fall at constant time intervals. For each subsetted data set, we suggest conducting a separate analysis, treating t_k, $k = 1, \ldots, K$, as the new baseline s and defining cases cumulatively as subjects who have events over the following, say, 1-year span, so that $Z_i \in (s = t_k, t = t_k + 1)$ and $\delta_i = 1$, and defining controls such that $Z_i > t_k + 1$ (Figure 1). A series of accuracy summaries, such as $\text{AUC}^{C/D}(0, 1)$, $\text{AUC}^{C/D}(2, 3)$, $\text{AUC}^{C/D}(4, 5)$, . . ., is obtained, and time-varying accuracy is indicated by a change in AUCs over time. The same idea can be applied to obtain time-varying sensitivity and specificity.

If prognostic information changes over time, updated information can be included in each subsetted analysis by using the last measured information to obtain updated risk predictions. Although we chose a 1-year cumulative window for illustration, the window is flexible and may be chosen to be more clinically meaningful depending on the disease setting. Alternatively, the incident/dynamic approach, discussed next, provides a finer time scale, allowing for a smoother characterization of performance over time without having to specify a window of time over which cases accumulate.

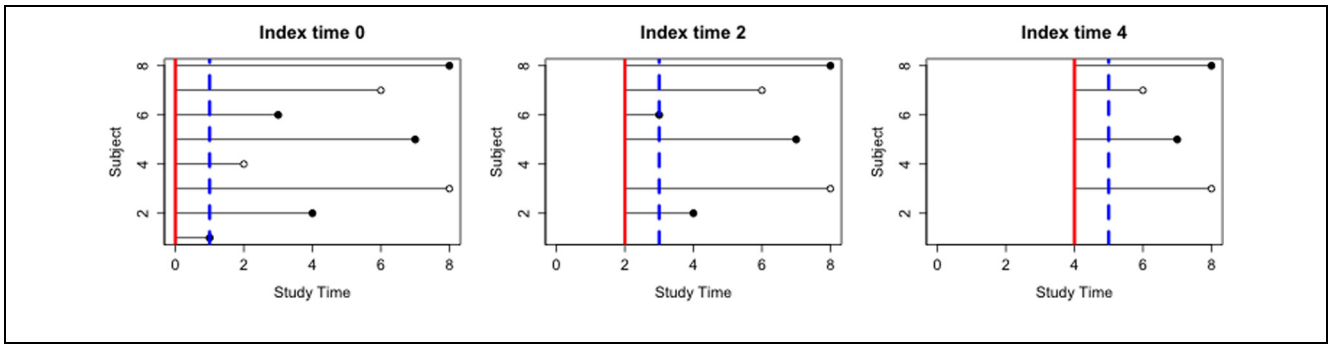

Figure 1 An illustration of assessments at sequential baseline time points. Solid circles represent events and hollow circles represent censored subjects. At each starting time point, subjects that remain event free are used for analysis. The solid red vertical line represents this cutoff. The dashed blue vertical line represents the subsequent 1-year cutoff, which is used to define cases v. controls.

Incident Cases/Dynamic Controls

Survival analysis using Cox regression is based on the fundamental concept of a risk set: a risk set at time t consists of the cases experiencing events *at* time t and the additional individuals who are under study (alive) but do not yet experience the clinical event. Extension of binary classification error concepts to risk sets leads naturally to adopting an incident (I) case definition where subjects who experience an event *at* time t or have survival time $T_i = t$ are the time-specific cases of interest. Dynamic controls (D) can be compared to incident cases and are subjects with $T_i > t$ (regardless of whether or not they experience the event or get censored at a later time). In this scenario, time-dependent definitions for sensitivity and specificity are as follows[34]:

$$\text{sensitivity}^{\text{I}}(c|t) = \text{P}(M > c|T = t)$$
$$\text{specificity}^{\text{D}}(c|t) = \text{P}(M \leq c|T > t)$$

For fixed $\text{specificity}^{\text{D}}(c|t) = 1 - p$, the time-dependent ROC value is the corresponding value of $\text{sensitivity}^{\text{I}}(c|t)$, or $\text{ROC}_{\text{t}}^{\text{I/D}}(p)$. The time-dependent AUC can be defined as the area under the time-specific ROC curve across all thresholds p:

$$\text{AUC}^{\text{I/D}}(t) = \int \text{ROC}_{\text{t}}^{\text{I/D}}(p)dp,$$

which can be shown to be equivalent to

$$\text{AUC}^{\text{I/D}}(t) = \text{P}(M_j > M_k|T_j = t, T_k > t).$$

Here, $\text{AUC}^{\text{I/D}}(t)$ is the probability that a random subject j who experiences an event at time t (case) has a larger marker value than a random subject k who remains event free through time t (control), assuming both subjects are event-free up to time t.

A semiparametric method based on the Cox model,[34] as well as a nonparametric rank-based method,[35] has been proposed for estimating $\text{ROC}_{\text{t}}^{\text{I/D}}(p)$ and $\text{AUC}^{\text{I/D}}(t)$ with censored outcomes. Both methods estimate $\text{FPF}_{\text{t}}^{\text{D}}$ nonparametrically; the difference comes from their estimation of $\text{TPF}_{\text{t}}^{\text{I}}$, which requires smoothing since the observed subset with $T_i = t$ may only contain 1 observation. The semiparametric method achieves smoothing by fitting a hazard model, whereas the nonparametric method uses kernel-based smoothing (see online Appendix A for additional details). The nonparametric approach is generally preferable as it relies on fewer assumptions than the semiparametric approach. In addition, the nonparametric method has been developed to provide a simple summary curve that graphically characterizes accuracy over time.

The performance of updated prognostic information can also be evaluated by using the semiparametric[34] or nonparametric[35] approach to accommodate time-varying markers.[36] At any time t, the last measured information may be used to obtain updated risk predictions from the prognostic model, as discussed in the previous section.

Global summary of marker performance. In many applications, there is no specific time t of interest, and a global accuracy summary of time-varying performance is desired. Furthermore, it may also be of interest to compare the overall performance of different markers or models. The incident/dynamic approach lends itself easily to addressing such questions, since marker performance can be summarized into a single-number global

summary called the survival concordance index (c-index)[34]:

$$\text{c-index} = \mathrm{P}\left(M_j > M_k | T_j < T_k\right).$$

The c-index is interpreted as the probability that the predictions for a random pair of subjects are concordant with their outcomes. In other words, it is the probability that the subject who died at an earlier time had a larger marker value. The c-index can also be expressed as a weighted average of time-specific AUCs[34] and is therefore easy to estimate using the incident/dynamic methods described above. The above definition of the basic c-index for survival outcomes applies to a baseline marker M. However, the definition and associated estimation methods can easily be generalized to accommodate updated prognostic information to estimate the generalized c-index for a time-varying marker, $M(t)$, expressed as

$$\text{generalized c-index} = \int \mathrm{AUC}^{\mathrm{I/D}}(t)\mathrm{w(t)dt}$$

using the weighted average representation, which allows time-varying markers to be use for each $\mathrm{AUC}^{\mathrm{I/D}}(t)$ (see online Appendix A for a definition of w(t) with further details and "Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis" for an illustration).

Extension to Competing Risk Outcomes

Often a subject's event time can be classified by one of several distinct causes, and interest may lie in events of a specific type. For example, in breast cancer studies, distant metastasis may be the event of interest; however, other clinical events, such as death, may preclude the researcher from observing distant metastases for particular patients.[37] The definitions of time-dependent sensitivity, specificity, ROC, and AUC presented in "Cumulative (Prevalent) Cases/Dynamic Controls" and "Incident Cases/Dynamic Controls" have been extended to incorporate cause of failure for competing risk outcomes for both the cumulative and incident case definitions, and we direct the reader to the associated literature.[38]

Software

The above methods have been implemented in publicly available `R` statistical software packages `survivalROC` (for cumulative/dynamic methods), `risksetROC` (for incident/dynamic methods with semiparametric estimation), and `meanrankROC` (for incident/dynamic methods with nonparametric estimation). The cumulative/dynamic methods have also been implemented as part of the `PHREG` procedure in the commercial software SAS. These software options are summarized in Table 1. In addition, the `survivalROC` and `risksetROC` packages have been extended to include updated definitions for competing risk outcomes.

We note that the choice of `R` package should depend on the chosen method, which should depend on the scientific question of interest, as discussed in "Comparison of Cumulative v. Incident Case Approaches" and illustrated using the `survivalROC` and `meanrankROC` packages in "Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis" (with accompanying code in online Appendix B).

Comparison of Cumulative v. Incident Case Approaches

Use of incident events naturally facilitates evaluation of time-varying prognostic performance, whereas the use of cumulative events in a sequential manner can also enable such evaluation. In practice, patterns in $\mathrm{AUC}^{\mathrm{I/D}}(t)$ tend to match $\mathrm{AUC}^{\mathrm{C/D}}(t,t+1)$ closely when the gap between t and $t+1$ is small, although $\mathrm{AUC}^{\mathrm{C/D}}(t,t+1)$ uses a coarser time scale and averages the performance over a fixed time interval.

In a descriptive context, $\mathrm{AUC}^{\mathrm{I/D}}$ may be preferable because it provides a simple graphical approach and a global summary using the c-index, without having to specify a time interval over which cases accumulate. In contrast, sequential use of cumulative cases based on $\mathrm{AUC}^{\mathrm{C/D}}$ may better align with clinical settings where prediction of short-term survival is needed at a specific decision time (or a small collection of times). For example, time intervals of 6 months, 1 year, and 5 years are commonly used for defining high-risk v. low-risk patients for targeted intervention. Methods for meaningfully averaging time-varying performance into a global performance summary using the cumulative case definition have not been developed.

Computationally, $\mathrm{AUC}^{\mathrm{I/D}}(t)$ is more straightforward to estimate and visualize for a series of time points. $\mathrm{AUC}^{\mathrm{C/D}}(t)$ requires the generation of a new subsetted data set for each time point of interest, and therefore if interest lies in several time points, then a series of $\mathrm{AUC}^{\mathrm{C/D}}(t)$ estimates may be more cumbersome to obtain.

Table 1 Guide to Available Software for Conducting Analyses Using the Cumulative/Dynamic and Incident/Dynamic Methods

Measures of Interest	Software
Cumulative cases/dynamic controls	R package `survivalROC`
• ROC function	`survivalROC()` accepts censored survival data and returns a set of TPF and FPF values for construction of the ROC curve, $ROC_{s,t}^{C/D}$, where s is the "baseline" time of the subsetted data set (i.e., $T \geq s$), while t (specified using the `predict.time` argument) defines the window over which cases accumulate, so that $T \leq t$ defines cases and $T > t$ defines controls. The function calculates estimates and associated 95% confidence intervals for $ROC_{s,t}^{C/D}(p)$ on subsetted data sets based on new index (or "baseline") times and updated marker values.
• AUC function	`survivalROC()` (described above) also calculates estimates and associated 95% confidence intervals for $AUC^{C/D}(s,t)$.
• Example	The documentation for the `survivalROC` package demonstrates the above functionality on *baseline* markers in the Mayo PBC data set. Furthermore, see "Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis" of this tutorial (and online Appendix B for corresponding R code) for an illustration of the package applied to assessing *time-dependent* discrimination accuracy of both baseline and *time-varying* markers.
Cumulative cases/dynamic controls	`SAS` procedure `PHREG`
• ROC function	The PHREG procedure accepts censored survival data and allows construction of the ROC curve, $ROC_{s,t}^{C/D}$, where s is the "baseline" time of a subsetted data set (i.e., $T \geq s$). One can specify `AT` $= t$ in the `ROCOPTIONS` in the `PROC PHREG` statement to define the window over which cases accumulate, so that $T \leq t$ defines cases and $T > t$ defines controls. Specifying PLOTS = ROC in the `PROC PHREG` statement displays the ROC curve at selected time points.
• AUC function	Using the same options as above, but instead specifying `PLOTS=AUC` in the `PROC PHREG` statement displays the AUC and the 95% confidence limits with respect to time.
• Example	The SAS user's guide for the `PHREG` procedure demonstrates the above functionality on the Mayo PBC data set to assess time-varying performance and to compare models.
Incident cases/dynamic controls (semiparametric estimation)	R package `risksetROC`
• ROC function	`risksetROC()` calculates estimates and associated 95% confidence intervals for $ROC_t^{I/D}(p)$ by accommodating updated marker values using time-dependent data and appropriately specifying the `entry` and `Stime` arguments. For example, consider the illustrative data set in Table 2(a) with marker values measured only at baseline. Compare this to the time-dependent data set in Table 2(b) that includes monthly updated marker values. When a new marker value is available, the individual is censored with the old value and reenters the study with the new value at the updated entry time.
• AUC function	`risksetROC()` (described above) also calculates estimates and associated 95% confidence intervals for $AUC^{I/D}(t)$.
• c-index function	`risksetAUC()` estimates the c-index. Confidence intervals can be computed using bootstrapping, as illustrated in the annotated code of online Appendix B.
• Example	The documentation for the `risksetROC` package demonstrates the above functionality on a lung cancer data set (also freely available in R, like the Mayo PBC data set).
Incident cases/dynamic controls (nonparametric estimation)	R package `meanrankROC`
• ROC function	`dynamicTP()` accommodates updated marker values using time-dependent data as described above and appropriately specifying `start` and `stop` times for intervals with updated marker values. `dynamicTP()`, along with `nne_TPR()`, provides a smooth curve over time of sensitivity (or TPF) or $ROC_t^{I/D}(p)$ for a fixed specificity $1 - p$.
• AUC function	`MeanRank()` accommodates updated marker values using time-dependent data as described above and appropriately specifying `start` and `stop` times for intervals with updated marker values. `MeanRank()`, along with `nne.Crossvalidate()`, provides a smooth curve of $AUC^{I/D}(t)$ over time.
• c-index function	`dynamicIntegrateAUC()` estimates the c-index. Confidence intervals can be computed using bootstrapping, as illustrated in the annotated code of online Appendix B.
• Example	See "Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis" of this tutorial (and online Appendix B for corresponding R code) for an illustration of the `meanrankROC` package applied to assessing time-dependent discrimination accuracy of both baseline and time-varying markers.

AUC, area under the receiver operating characteristic curve; FPF, false-positive fraction; PBC, primary biliary cirrhosis; ROC, receiver operating characteristic; TPF, true-positive fraction.

Table 2 Illustration of Data Sets with Marker Values Measured Only at Baseline and Updated Approximately Every Month[a]

Subject	Marker	Start Time (Days)	Stop Time (Days)	Death Observed
Marker measured at baseline only				
1	m_0	0	65	1
2	m_0	0	40	0
Marker measured approximately monthly				
1	m_0	0	25	0
1	m_{25}	25	58	0
1	m_{58}	58	65	1
2	m_0	0	30	0
2	m_{30}	30	40	0

[a]Subjects are censored when a new marker value is available, and they reenter the study with the new marker value and an updated start time.

Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis

As an illustrative case study, we consider the problem of liver transplantation in PBC that was introduced in "Case Study: Liver Prognostic Model to Guide Transplantation in Primary Biliary Cirrhosis."

Description of Study Cohort

The study cohort consisted of 312 patients with PBC; 125 (40%) of these patients were observed to die during the study period; 19 subjects were recipients of liver transplantation during the study period. We censored these subjects at the time of transplantation, since the prognostic model is intended to predict the risk of mortality *without transplantation*, and we use that risk to prioritize such patients. For each patient, we had baseline demographic and diagnosis data and longitudinal data on laboratory measures. Counting multiple observations per patient, we included 1945 total records.

Risk Models

We evaluated the following models: 1) a 5-covariate model containing the same variables as those in the Mayo model[1]: log(bilirubin), albumin, log(prothrombin time), edema, and age, and 2) a 4-covariate model where we omitted log(bilirubin) to illustrate the comparison of different candidate models. Predictions from Cox models were summarized into a single baseline risk score and a separate time-varying, updated risk score to demonstrate that the methods can incorporate time-varying measurements and to show the implications of using older measurements on accuracy. For the baseline score, we used 10-fold cross-validation to protect against overfitting.[16–18] For the time-varying score, we used baseline measurements as training data to develop the Cox model and predicted the score at follow-up times using updated values of log(bilirubin), albumin, and log(prothrombin time).[16–18]

What Is the Accuracy of Baseline Measurements and the Value of Updated Measurements?

As a first step, we use the incident/dynamic approach to assess the prognostic accuracy of the baseline risk score obtained from the 4-covariate model v. the 5-covariate model. Figure 2 and Table 3 show that the 5-covariate model has consistently better performance than the 4-covariate model over time with respect to both $AUC^{I/D}(t)$ (Table 3 and Figure 2, left panel) and sensitivity for a fixed specificity of 10% (Figure 2, right panel). The estimated c-indices are 0.72 (95% confidence interval [CI], 0.66–0.76) and 0.79 (95% CI, 0.75–0.83) for the 4- and 5-covariate models, respectively, with a statistically significant difference of 0.07 (95% CI, 0.04–0.11). Table 3 also shows the sequential cumulative/dynamic approach that uses successive 1-year windows to define cases. We see similar estimates for $AUC^{I/D}$ and $AUC^{C/D}$. Any observed differences are due to $AUC^{I/D}$ reflecting performance at a given time point and $AUC^{C/D}$ averaging performance over a 1-year window.

Looking at the 5-covariate model, the performance of the baseline score declines over time with $AUC^{I/D} = 0.88$ (95% CI, 80–0.90) at 1 year v. 0.66 (95% CI, 0.62–0.78) at 6 years. In contrast, fairly consistent performance is maintained using a risk score that is updated over time ($AUC^{I/D}(t) = 0.92$ [95% CI, 0.88–0.96] at 1 year, 0.89 [95% CI, 0.84–0.92] at 6 years) (Table 3 and Figure 3). The 95% CIs are included in Table 3 and can also be

Table 3 Time-Varying Performance of Baseline and Updated Risk Scores from the 4-Covariate and 5-Covariate Models Using $AUC^{I/D}$ and $AUC^{C/D}$

	$AUC^{I/D}(t)$ (95% CI)				$AUC^{C/D}(t,t+1$ year) (95% CI)		
	$t = 1$ Year	$t = 4$ Years	$t = 6$ Years	c-index (95% CI)	$t = 1$ Year	$t = 4$ Years	$t = 6$ Years
Baseline risk scores							
4-covariate model	0.84 (0.79–0.89)	0.69 (0.60–0.76)	0.64 (0.55–0.70)	0.72 (0.66–0.74)	0.77 (0.56–0.95)	0.72 (0.55–0.87)	0.77 (0.60–0.88)
5-covariate model	0.88 (0.80–0.91)	0.85 (0.74–0.86)	0.66 (0.62–0.78)	0.79 (0.76–0.83)	0.80 (0.57–0.93)	0.78 (0.66–0.91)	0.65 (0.44–0.89)
Updated risk scores							
4-covariate model	0.90 (0.86–0.96)	0.86 (0.80–0.91)	0.84 (0.77–0.90)	0.86 (0.80–0.89)	0.79 (0.61–0.95)	0.81 (0.63–0.91)	0.84 (0.63–0.95)
5-covariate model	0.92 (0.88–0.96)	0.92 (0.86–0.95)	0.88 (0.82–0.93)	0.89 (0.84–0.92)	0.82 (0.70–0.94)	0.84 (0.68–0.94)	0.87 (0.66–0.99)

AUC, area under the receiver operating characteristic curve; CI, confidence interval.

included in plots, as shown in Figure 4 for baseline and updated risk scores from the 5-covariate model.

Similar patterns are observed for the 4-covariate model, with the baseline score's performance declining over time and the updated risk score's performance staying fairly steady. Interestingly, the updated 4-covariate risk score performs almost as well as the updated 5-covariate risk score, indicating that some of the loss of accuracy due to the omission of log(bilirubin) can be recovered by using updated measurements on other variables.

Implications for Decision Making in PBC

This Mayo risk score has been used for individual-level decision making about transplantation over time, by selecting patients who are most likely to die in a short time interval from the time of prediction. We used the 5 main predictors of survival identified by Dickson et al.[1] to calculate the predicted risk of mortality and evaluate the accuracy of these predictions toward prioritizing patients for transplantation. It is clear from the results that patient information should be updated regularly in practice to maintain prognostic accuracy. The updated 5-covariate Mayo model maintains an $AUC^{I/D}$ of around 0.90 over time, with a high generalized c-index of 0.89 (95% CI, 0.84–0.92), indicating that it is a strong prognostic model for use in practice. In addition, we used $AUC^{C/D}$ sequentially with 1-year windows to evaluate the use of the Mayo model as a decision-making tool in practice. We found that $AUC^{C/D}$ is consistently above 0.80 at all chosen time points, indicating that the model identifies high-risk patients for transplantation with high accuracy.

Discussion

The American Heart Association's 2009 criteria for evaluating a risk prediction model categorize performance measures into those of calibration, association, discrimination, and risk reclassification.[39] Similarly, Steyerberg et al.[40] differentiated the roles of various performance measures for assessing prediction models, defining them as measures of overall performance, discrimination, calibration, reclassification, and clinical usefulness. They explained that these measures serve different purposes and suggested that "reporting discrimination and calibration will always be important for a prediction model." Although their focus was on binary outcomes, the same ideas hold for survival outcomes.

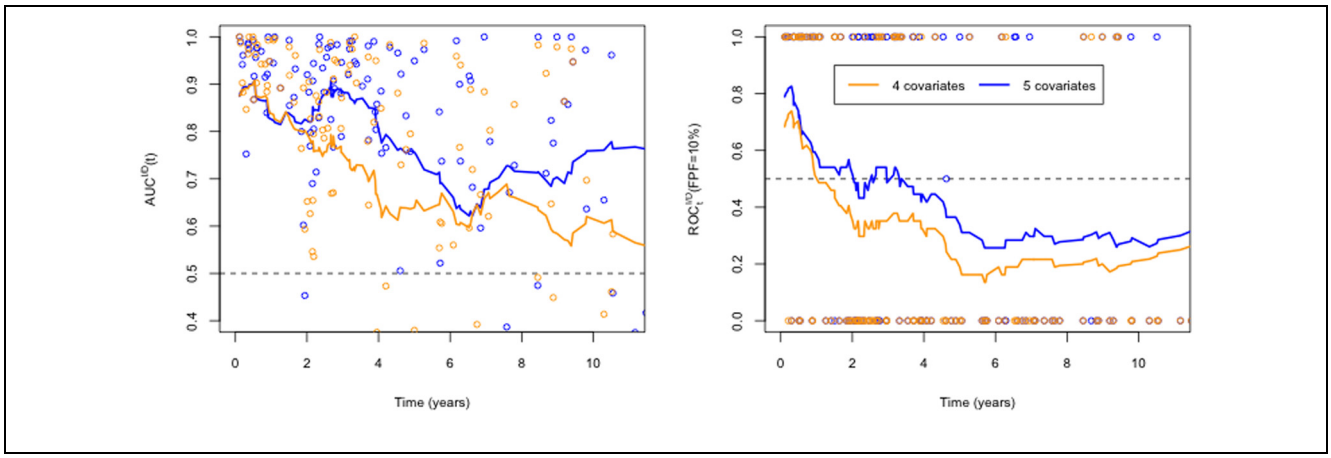

Figure 2 Time-varying prognostic accuracy of baseline risk scores obtained from the 4-covariate model v. the 5-covariate model over time using the incident/dynamic approach, with respect to $AUC^{I/D}(t)$ (left) and $ROC_t^{I/D}$ (right) for a fixed false-positive fraction (FPF) of 10% (or sensitivity for a fixed specificity of 90%).

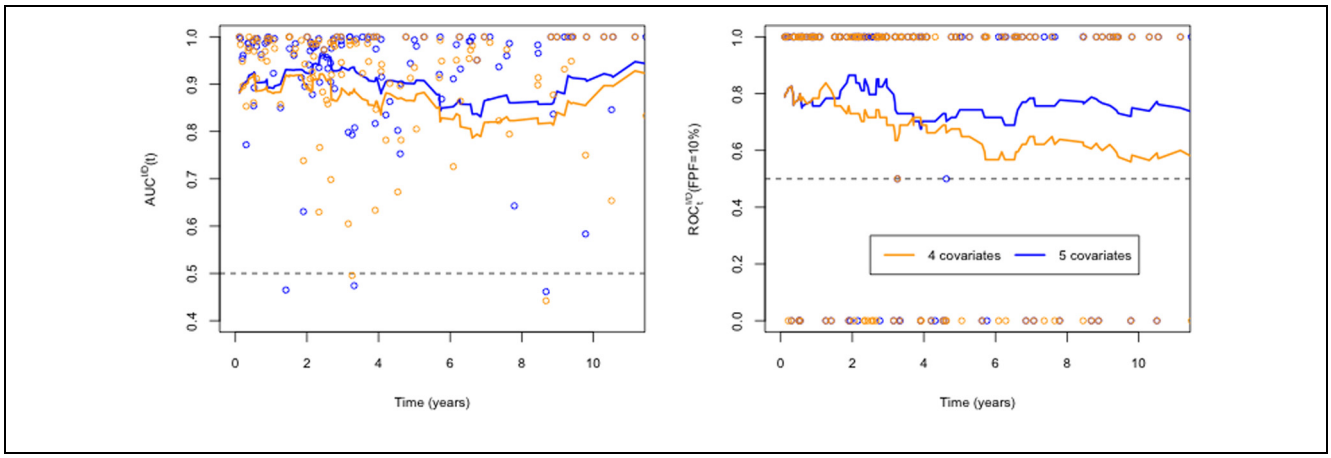

Figure 3 Time-varying prognostic accuracy of updated risk scores obtained from the 4-covariate model v. the 5-covariate model over time using the incident/dynamic approach, with respect to $AUC^{I/D}(t)$ (left) and $ROC_t^{I/D}$ (right) for a fixed false-positive fraction (FPF) of 10% (or sensitivity for a fixed specificity of 90%).

In this tutorial, we focused on discrimination accuracy (other work has demonstrated calibration for prognostic models for survival outcomes[41]). We presented methods that extend standard diagnostic definitions of sensitivity and specificity and develop key summaries for evaluating the time-varying prognostic performance of a marker or model measured at baseline only or updated in routine clinical care. A basic epidemiologic concept that distinguishes alternative summaries is the idea of cumulative v. incident events to define cases. $AUC^{I/D}(t)$ is a convenient descriptive and graphical summary that characterizes time-varying performance without having to select a particular timeframe over which cases accrue, whereas sequential use of $AUC^{C/D}(t)$ may be useful in clinical settings where predictions of short-term survival are needed at select times to identify high-risk patients for targeted intervention.

In addition to allowing for evaluation of time-varying discrimination accuracy of prognostic models, there are other implications for how these methods could be applied in practice. First, these methods may guide practice and policy with regard to the frequency of updating patient information, by comparing the performance of risk scores updated using different measurement

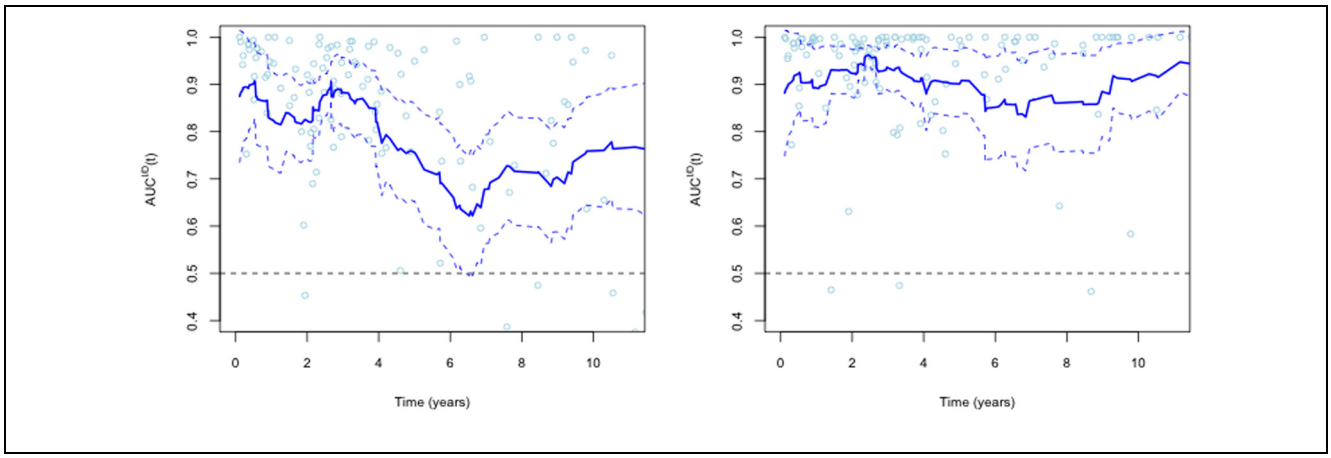

Figure 4 Time-varying prognostic accuracy with 95% confidence intervals of baseline (left) and updated (right) risk scores obtained from the 5-covariate model using the incident/dynamic approach.

schedules to assess how often patient information should be updated before it becomes outdated and affects accuracy. Second, although we compared the 5-covariate Mayo model to a simple 4-covariate variation of the model for illustration, in practice, one may choose more clinically relevant variables, such as more expensive measures, to omit or replace and assess the impact on prognostic accuracy. Finally, one may choose to explore the performance of a risk model in subsets of patients, say older v. younger patients, to assess whether the model is a better decision-making tool for particular subgroups.

One limitation of this tutorial is that we do not discuss model selection in detail, focusing on the evaluation of a given model. However, the methods for model evaluation that we discuss could also be used at the stage of model selection to guide identification of a model with optimal performance. For example, with variable selection in high-dimensional settings, one may use the c-index, which is a global summary of time-varying performance, as a way of initially screening the strongest markers as candidates for combining into a multivariate risk score. One may also use the c-index as the optimization criterion in model selection, instead of the typically used likelihood-based criteria.[42–44] For example, approaches that optimize the c-index have been developed using boosting.[45,46]

A potential limitation of the case study is that in the absence of an independent data set on PBC, our evaluation uses the same data set that was used by Dickson et al.[1] to develop the Mayo model. As discussed in "Model Development," the standard approach is to use separate training and validation data sets to fairly assess model performance. We used cross-validation to mitigate the potential issue of an optimistic assessment. In practice, an independent validation data set is important if the results may have clinical implications. However, this case study was meant to illustrate methods, rather than inform clinical practice. In addition, the case study uses data from a trial conducted between 1974 and 1984. Again, a newer data set would not add substantially to our primary goal of illustrating methods. Furthermore, the Mayo model, which is widely used in practice today, was developed using the same data set.

Finally, there is growing interest in evaluating the incremental value gained from adding a new marker(s) to an existing baseline marker or model. Difference in AUC is a popular metric for evaluating incremental value. As we illustrated using the case study, the time-varying incremental value of a marker can be evaluated by comparing the time-varying AUCs of 2 models. In addition, a number of alternative measures have been proposed in recent literature for binary outcomes—namely, the net reclassification index[47] and integrated discrimination improvement.[48] Extensions of these measures for time-dependent outcomes have been developed[49,50] and can provide alternative summaries of the time-varying incremental value of a marker.

Supplementary Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm. Supplementary R code available at http://faculty.washington.edu/abansal/software.html.

References

1. Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*. 1989;10:1–7.
2. Murtaugh PA, Dickson ER, Van Dam GM, et al. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*. 1994; 20(1): 126–34.
3. Coombes JM, Trotter JF. Development of the allocation system for deceased donor liver transplantation. *Clin Med Res*. 2005; 3:87–92.
4. Egan TM, Murray S, Bustami RT, et al. Development of the new lung allocation system in the United States. *Am J Transplant*. 2006;6:1212–27.
5. Leung KM, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health*. 1997;18:83–104.
6. Little RJA, Rubin DB. 1982. *Statistical Analysis with Missing Data*. New York: John Wiley.
7. Mayo Clinic. Primary biliary cholangitis. Available from: https://www.mayoclinic.org/diseases-conditions/primary-biliary-cirrhosis/basics/definition/con-20029377
8. Lammers WJ, Kowdley KV, van Buuren HR. Predicting outcome in primary biliary cirrhosis. *Ann Hepatol*. 2014;13(4):316–26.
9. Christensen E, Neuberger J, Crowe J, et al. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial. *Gastroenterology*. 1985;89:1084–91.
10. Roll J, Boyer JL, Barry D, Klatskin G. The prognostic importance of clinical and histologic features in asymptomatic and symptomatic primary biliary cirrhosis. *N Engl J Med*. 1983;308:1–7.
11. Bonsel GJ, Klompmaker IJ, van't Veer F, Habbema JD, Slooff MJ. Use of prognostic models for assessment of value of liver transplantation in primary biliary cirrhosis. *Lancet*. 1990;335:493–7.
12. Rydning A, Schrumpf E, Abdelnoor M, Elgjo K, Jenssen E. Factors of prognostic importance in primary biliary cirrhosis. *Scand J Gastroenterol*. 1990;25:119–26.
13. Christensen E, Altman DG, Neuberger J, De Stavola BL, Tygstrup N, Williams R. Updating prognosis in primary biliary cirrhosis using a time-dependent Cox regression model. PBC1 and PBC2 trial groups. *Gastroenterology*. 1993;105:1865–76.
14. Krzeski P, Zych W, Kraszewska E, Milewski B, Butruk E, Habior A. Is serum bilirubin concentration the only valid prognostic marker in primary biliary cirrhosis? *Hepatology*. 1999;30:865–9.
15. Dickson ER, Fleming TR, Wiesner RH, et al. Trial of penicillamine in advanced primary biliary cirrhosis. *N Engl J Med*. 1985;312:1011–5.
16. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–87.
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann; 1995. p 338–45.
18. van Belle G, Fisher LD, Heagerty PJ, Lumley T. *Biostatistics: A Methodology for the Health Sciences*. Hoboken, NJ: John Wiley.
19. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc B Met*. 1972;34:187–220.
20. Kalbleish J, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: Wiley-Interscience; 2002.
21. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. *Stat Methods Med Res*. 2010;19: 29–51.
22. Verweij P, van Houwelingen H. Penalized likelihood in cox regression. *Stat Med*. 1994;13:2427–36.
23. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med*. 1997;16:385–95.
24. Li H, Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*. 2005;21:2403–9.
25. Lisboa PJ, Wong H, Harris P, et al. A Bayesian neural network approach for modeling censored data with an application to prognosis after surgery for breast cancer. *Artif Intell Med*. 2003;28:1–25.
26. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. *Stat Med*. 2004;23(1):77–91.
27. Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. *Ann Appl Stat*. 2008;2(3):841–60.
28. Swets JA, Pickett RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press; 1982.
29. Metz CF. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;8:283–8.
30. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press; 2003.
31. Hanley JA, McNeil BJ. The meaning and use of the area under an ROC curve. *Radiology*. 1982;143:29–36.
32. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56:337–44.
33. Zheng Y, Heagerty PJ. Prospective accuracy for longitudinal markers. *Biometrics*. 2007;63:332–41.
34. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005;61:92–105.
35. Saha-Chaudhuri P, Heagerty PJ. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*. 2013;14:42–59.
36. Bansal A, Heagerty PJ, Saha-Chaudhuri P. *Dynamic Placement Values: A Basis for Evaluating Prognostic Potential*. Unpublished manuscript.
37. Buyse M, Loi S, van't Veer L, et al., on behalf of the TRANSBIG Consortium. Validation and clinical utility of a

70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Instit*. 2006;98:1183–92.
38. Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*. 2010;66: 999–1011.
39. Hlatky MA, Greenland P, Arnett DK, et al; American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009;119(17):2408–16.
40. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–38.
41. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*. 2006;113:1424–33.
42. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer Science + Business Media; 2009.
43. Hoerl AE, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12:55–67.
44. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996;58:267–88.
45. Ma S, Huang J. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*. 2005;21:4356–62.
46. Mayr A, Schmid M. Boosting the concordance index for survival data: a unified framework to derive and evaluate biomarker combinations. *PLoS ONE*. 2014;9(1):e84483.
47. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157–72.
48. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30: 11–21.
49. Liang CJ, Heagerty PJ. A risk-based measure of time-varying prognostic discrimination for survival models. *Biometrics*. 2016 Nov 28. [Epub ahead of print]. DOI: 10.1111/biom.12628.
50. French B, Saha-Chaudhuri P, Ky B, Cappola TP, Heagerty PJ. Development and evaluation of multi-marker risk scores for clinical prognosis. *Stat Methods Med Res*. 2016;25:255–71.