Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting

DAVID M. VOCK, Division of Biostatistics, School of Public Health, University of Minnesota
JULIAN WOLFSON, Division of Biostatistics, School of Public Health, University of Minnesota
SUNAYAN BANDYOPADHYAY, Department of Computer Science, University of Minnesota
GEDIMINAS ADOMAVICIUS, Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota
PAUL E. JOHNSON, Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota
GABRIELA VAZQUEZ-BENITEZ, HealthPartners Institute for Education and Research
PATRICK J. O'CONNOR, HealthPartners Institute for Education and Research

Models for predicting the probability of experiencing various health outcomes over a certain time frame (e.g., having a heart attack in the next 5 years) based on individual patient characteristics are important tools for managing patient care. Because electronic health data (EHD) from healthcare systems provide access to large amounts of individual-level data from contemporaneous patient populations, they are appealing sources of training data for building risk prediction models. Machine learning approaches to estimate risk are attractive because of their ability to capture complex relationships between individual characteristics and health outcomes, thereby allowing the population to be partitioned into distinct subgroups based on their risk. However, since EHD are derived by extracting information from administrative databases, some fraction of subjects will not be under observation for the entire time frame over which one wants to make predictions; this loss to follow-up is often due to disenrollment from the health system. For subjects without complete follow-up, the event status is unknown, and in statistical terms the event time is said to be right-censored. While there is a well-developed statistical literature on regression models which account for right-censored data, most machine learning approaches to the problem have been relatively *ad hoc*, for example, discarding the censored observations or treating them as non-events. In this paper, we present a rigorous, general-purpose approach to account for right-censored outcomes using the inverse probability of censoring weighting (IPCW). We illustrate how IPCW can easily be incorporated into a number of existing machine learning algorithms, and show that our approach leads to better predictive performance than *ad hoc* approaches. Our techniques are motivated by and illustrated on the problem of predicting the 5-year risk of experiencing a cardiovascular event, using EHD from a large U.S. Midwestern healthcare system.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Survival analysis; H.2.8 [**Database Management**]: Database Applications Data mining; J.3 [**Life and Medical Sciences**]: Medical information systems

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Machine learning, censored data, electronic health data, survival analysis, inverse probability of censoring weights, risk prediction, medical decision support.

ACM Reference Format:
David M. Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E. Johnson, Gabriela Vazquez-Benitez, and Patrick J. O'Connor, 2014. Adapting machine learning techniques to censored time-to-event data: a general approach. *ACM Trans. Knowl. Discov. Data.* 1, 1, Article 1 (January 0000), 25 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Predictions of the "personalized" risk of a patient experiencing various health outcomes (e.g., heart attack, stroke, diabetes, etc.) are critical tools in clinical practice. Risk prediction and stratification help clinicians

This work was partially supported by NHLBI grant R01HL102144-01 and AHRQ grant R21HS017622-01.
Author's addresses: DMV, JW: Division of Biostatistics, School of Public Health, A460 Mayo Building MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455. SB: Computer Science and Engineering Department, 200 Union St SE, Minneapolis, MN 55455. GA, PEJ: Department of Information and Decision Sciences, Carlson School of Management, 321 19th Ave S, Minneapolis, MN 55455. GV-B, PJO: HealthPartners Institute for Education and Research, Mail stop 21111R, P.O. Box 1524, Minneapolis, MN 55440-1524.

DOI:http://dx.doi.org/10.1145/0000000.0000000

to optimize resource allocation, to develop appropriate intervention strategies for those at high risk of an adverse health outcome, and to motivate patients to remain adherent to these strategies. Additionally, risk prediction tools can help raise awareness of the burden of various diseases and the risk factors associated with them. Given the importance of risk prediction and stratification in the clinical setting, there is currently great interest in developing machine learning methods to estimate flexibly the "personalized" risk of a patient experiencing various adverse health outcomes.

1.1. Cardiovascular risk prediction using right-censored electronic health data

For cardiovascular disease and related outcomes (e.g., heart attack, stroke), i.e., the application area which motivates our work here, recent systematic reviews have described over 100 risk models produced between 1999 and 2009 alone [Cooney et al. 2009, 2010; Matheny et al. 2011], including Framingham [D'Agostino et al. 2008], SCORE [Conroy et al. 2003], ASSIGN-SCORE [Woodward et al. 2007], QRISK1 [Hippisley-Cox et al. 2007, 2008], QRISK2 [Hippisley-Cox et al. 2008], PROCAM [Assmann et al. 2002], WHO/ISH, and Reynolds Risk Score [Ridker et al. 2007, 2008]. Most risk prediction models, including those mentioned above, have been estimated using data from carefully selected epidemiological cohorts. For example, the widely-used Framingham risk score is trained on a data set that represents a predominantly ($\approx 99\%$) Caucasian population and incorporates patient follow-up data from the late 1960s [D'Agostino et al. 2008]. As a result of estimating the risk of cardiovascular (CV) events using data from these narrowly defined cohorts, existing risk models often provide poor risk estimates for diverse, contemporaneous patient populations. For example, Collins and Altman [2009] illustrate the poor performance of the Framingham risk equations when applied to a contemporary population of over one million United Kingdom residents seen in the primary care clinic.

The increasing availability of electronic health data (EHD) represents a key opportunity to improve risk prediction models. EHD, which consist of electronic medical records (EMRs), insurance claims data, and mortality data obtained from governmental vital records, are increasingly available within the context of large healthcare systems and capture the characteristics of heterogeneous populations receiving care in a contemporary clinical setting. EHD databases typically include data on hundreds of thousands to millions of patients; therefore, a risk prediction model constructed from EHD has the potential to yield accurate and generalizable risk predictions, because clinically important sub-populations over which risk is likely to vary (e.g., racial and ethnic minorities, patients with specific pre-existing conditions) are often well-represented in such large databases.

The scale and complexity of EHD data provide an excellent opportunity to develop more accurate risk models using modern machine learning techniques [Colombet et al. 2000; Song et al. 2004; Wu et al. 2010; Kawaler et al. 2012; Sun et al. 2012; Austin et al. 2013; Kennedy et al. 2013; Wang et al. 2013; Lin et al. 2014; Stewart et al. 2014]. However, in many datasets derived from EHD, a large fraction of subjects do not have enough follow-up data available to ascertain whether or not they experienced the event of interest over a given time period (e.g., a CV event within 5 years of a defined baseline). In the language of statistical survival analysis, the event times of those subjects is said to be *right-censored* [Kalbfleisch and Prentice 2002].

1.2. Existing techniques for right-censored data

There are many regression-based methods that handle right-censored data including the Cox proportional hazards [Cox 1972] and accelerated failure time [Buckley and James 1979] models. However, the proportional hazards (or accelerated failure time) model make the potentially restrictive assumption that the risk factors have a linear relationship with the log hazard (or log time, respectively) of experiencing an event. If the analyst has *a priori* knowledge that this relationship is non-linear or differs in certain sub-groups, he or she may include non-linear transformations of predictors or interactions between predictors, but this is often based on trial and error [Kattan et al. 1998], and a more flexible machine learning approach is often warranted.

Fully supervised machine learning and classification methods typically assume that the event status is known for all subjects, while in our setting the event status is undetermined for subjects whose event time is censored and who are not followed for the full time period over which one wants to make predictions (e.g., 5 years). Simple approaches to dealing with this issue, such as discarding censored observations (see, e.g., Larranaga et al. [1997]; Sierra and Larranaga [1998]; Blanco et al. [2005]) or treating them as zeroes (non-events), are known to induce bias in the estimation of class probabilities [Kattan et al. 1998], making typical fully supervised classification employing these *ad hoc* approaches unsuitable. For example, Štajduhar et al. [2009] demonstrated the impact of unaccounted-for censoring on the construction and performance of

Bayesian networks. Semi-supervised approaches are also generally not applicable since the labeled (known event status) and unlabeled (unknown event status) observations are not samples from the same underlying population, and censored observations are not truly 'unlabeled' since they carry useful *partial* information about the outcome.

There has been increasing interest in adapting machine learning tools to censored, time-to-event data. Several authors including Segal [1988]; Hothorn et al. [2004]; Ishwaran et al. [2008]; Zhu and Kosorok [2012] describe versions of classification trees and random forests to estimate the survival distribution. Lucas et al. [2004] and Bandyopadhyay et al. [2014] discuss the application of Bayesian networks to right-censored data. Zupan et al. [1999] and Štajduhar and Dalbelo-Bašić [2010] have proposed approaches in which censored observations are repeated twice in the dataset, one as experiencing the event and one event-free. Each of these observations is assigned a weight based on the marginal probability of experiencing an event between the censoring time the time the event status will be assessed. This approach, although intuitive, is provably biased and inconsistent. Štajduhar and Dalbelo-Bašić [2012] adopts a more principled likelihood-based approach to imputing event times, but their imputation technique may perform poorly if the assumed parametric distribution of event times is incorrect. A few authors have considered applying neural networks to survival data but typically assume that the possible censoring and event times are few in number [Biganzoli et al. 1998; Ripley and Ripley 2001]. Additionally, several have considered adapting support vector machines to censored outcomes by altering the loss function to account for censoring [Shivaswamy et al. 2007; Khan and Zubek 2008; Shim and Hwang 2009; Van Belle et al. 2011; Goldberg and Kosorok 2012]. Other approaches, including replacing the time to event with the martingale from the null model, have been proposed to handle censored data in other machine learning methods including support vector regression, recursive partitioning, and multiple adaptive regression splines [Therneau et al. 1990; Kattan et al. 1998; Kattan 2003], but this technique can only be used with learning algorithms which permit a continuous outcome and exclude classification methods. To our knowledge, a general approach to handling censored survival data in machine learning applications has not been proposed.

1.3. Inverse probability of censoring weights (IPCW)

In this paper, we propose a general-purpose technique for mining right-censored time-to-event data using inverse probability of censoring weights (IPCW). The technique properly accounts for censoring and can be easily integrated into many existing machine learning techniques for class probability estimation, allowing sophisticated new (possibly ensemble-based) machine learning tools for censored data to be created with minimal programming effort. The advantage of our proposed approach is that it may be incorporated in any software package that allows for "observation weights" (also referred to as "instance weights" or "case weights"). We illustrate how IPCW can be applied to create "censoring-aware" versions of several popular prediction methods: logistic regression, generalized additive models, Bayesian networks, binary decision trees, and k-nearest neighbors. We also argue that traditional evaluation metrics for assessing model calibration and classification accuracy (e.g., AUC and net reclassification improvement) may be misleading in the presence of censoring, and describe alternatives which are more appropriate for use in selecting model tuning parameters. We conclude by applying IPCW-aware machine learning methods to predict the occurrence of cardiovascular events from electronic health data collected by a large Midwest healthcare delivery organization.

2. INVERSE PROBABILITY OF CENSORING WEIGHTING

2.1. Notation

Let E be the indicator that an event occurs between a pre-defined baseline $t = 0$ and some fixed time $t = \tau$. Throughout the paper, we refer to E as the τ-year event status. In our setting, for example, we are interested in whether or not a cardiovascular event occurs within 5 years of "baseline" (e.g., an "index" clinic visit) where risk factor data are available. Binary classification methods typically assume that E is fully observed for all patients, but this is unlikely to be true when using information from contemporary EHD. When a patient leaves the health system or the study ends before τ years of follow-up and before the subject experiences the event of interest, the subject's event status at τ (i.e., E) is unknown. To establish notation which is standard in the statistical literature, for individual i define T_i as the time between baseline and the event of interest, and define C_i as the time between baseline and when the patient is lost to follow-up (e.g., in our context, disenrolls from the health plan or reaches the end of the data capture period without experiencing an event). We observe $V_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{I}(T_i < C_i)$, the indicator for whether or not an event occurred during the follow-up period, which may extend beyond the interval $[0, \tau]$. If $\delta_i = 0$, the subject's event time is right-censored. We can only ascertain that an event occurred ($E_i = 1$) if $\delta_i = 1$, or

that an event did not occur ($E_i = 0$) if $\delta_i = 0$ and $V_i > \tau$. In other words, the value of E_i is unknown if subject i is censored prior to τ, or equivalently $\min(T_i, \tau) \geq C_i$. We will denote the set of features available on individual i by $\mathbf{X}_i$; it is assumed that these features are fully observed at the beginning of the follow-up period and, hence, are not subject to censoring and do not vary over time. The target of prediction is $\pi(\mathbf{X}_i) = P(E_i = 1|\mathbf{X}_i) \equiv P(T_i \leq \tau|\mathbf{X}_i)$, and predictions are denoted by $\hat{\pi}(\mathbf{X}_i)$. We assume throughout that our data are partitioned into a training and test set, with the training set used to fit the models described below and the test set used to estimate prediction error.

2.2. The IPCW method

Naive approaches to handling the subjects for whom E is unknown (e.g., excluding them from our training data set or setting $E = 0$ for all) lead to biased estimators of the risk and hence potentially poor classification performance. Instead, we propose to use an inverse probability of censoring weighting (IPCW) approach for censored event times which is well-established in the statistical literature but to our knowledge has not been broadly applied for machine learning. Intuitive, excluding subjects for whom E is unknown leads to poor risk prediction because subjects with small event times are less likely to be censored than those with event times beyond τ. Therefore, we "over-sample" subjects with $E = 1$ if we exclude patients for whom E is unknown. In IPCW, only those subjects for whom E is known contribute directly to the analysis, but they are reweighted to accurately "represent" the subjects with unknown E. For example, supposing that 1/3 of subjects have censoring times greater than 3 years, a subject who experiences an event at $t = 3$ years and hence has (known) $E = 1$ can be thought of as representing 3 individuals: 2 similar or "shadow" subjects censored prior to $t = 3$ for whom E is unknown, plus themselves. Thus, subjects with known event status E and a longer time to event receive larger weights as they represent a greater number of "shadow" subjects whose event status is unknown due to censoring. IPC weighting is conceptually equivalent to creating a new dataset where each subject is replicated ω_i times. However, creating such an expanded dataset is often not advisable, both for reasons of practicality (memory/storage limitations) and mathematical precision (ω_i may not be an integer or simple fraction).

The advantage of IPCW to account for censoring is that it is a general-purpose approach that may be applied to any machine learning method. The analyst can then apply several different machine learning methods for risk prediction and select the optimal one based on censoring-adjusted criteria discussed in Section 3.

The general-purpose IPCW method proceeds as follows:

(1) Using the training data, estimate the function $G(t) = P(C_i > t)$, the probability that the censoring time is greater than t, using the Kaplan-Meier estimator of the survival distribution (i.e., 1 minus the cumulative distribution function) of the censoring times [Kalbfleisch and Prentice 2002],

$$\hat{G}(t) = \prod_{j:V_j<t} \left(\frac{n_j - d_j^*}{n_j} \right) \tag{1}$$

where d_j^* is the number of subjects who were censored at time V_j, and n_j is the number of subjects "at risk" for censoring (i.e., not previously censored or experiencing an event) at time V_j. We note that, for IPCW, Kaplan-Meier is applied to estimate the distribution of *censoring times*, whereas it is much more commonly used to estimate the distribution of *event times*. Standard software functions can be used to estimate G by applying the Kaplan-Meier estimator of event times with "censoring" indicators defined by $\delta_i^* = 1 - \delta_i$.

(2) For each patient i in the training set, define an inverse probability of censoring weight,

$$\omega_i = \begin{cases} \frac{1}{\hat{G}\{\min(V_i,\tau)\}} & \text{if } \min(T_i, \tau) < C_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

(3) Apply an existing prediction method to a weighted version of the training set where each member i of the training set is assigned weight ω_i.

Step 3 is left purposefully vague, as the manner in which IPC weights are incorporated will vary according to the machine learning technique used. Section 4 illustrates how a variety of machine learning algorithms can be adapted for censoring using IPCW.

3. RISK PREDICTION EVALUATION METRICS FOR CENSORED DATA

A key feature of many machine learning techniques for risk prediction is that they can be improved by adjusting parameters to optimize a performance metric, e.g., the misclassification rate, calibration, etc. Furthermore, just as in uncensored scenarios, a single censoring-aware machine learning technique is unlikely to be superior across all possible applications. Therefore, we need some performance metric to compare the predictive ability across different methods or to efficiently combine the results from multiple techniques. For the same reasons described above that failing to account for censoring yields biased parameter estimates, the usual performance metrics applied to risk prediction problems can be misleading when outcomes are subject to censoring. Here, we discuss modifications of standard calibration (goodness-of-fit test statistic) and discrimination (concordance index and net reclassification improvement) metrics which properly account for censored data and allow model performance to be assessed more accurately. In Section 4, we describe how these modified metrics can be used to select tuning parameter values for IPC-weighted versions of machine learning techniques.

3.1. Calibration

In standard risk prediction settings, calibration is commonly assessed by ranking the predicted risks $\hat{\pi}(\mathbf{x}_i)$, partitioning the ranked predictions into bins $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_m$ (e.g., by decile or clinically relevant cut points), and comparing the average predicted risk in each bin to an empirical estimate of the risk within that bin. When E_i is known for all subjects, the empirical risk estimate for bin $\mathcal{B}_k$ is simply given by $\sum_{i \in \mathcal{B}_k} E_i / |\mathcal{B}_k|$, where $|\mathcal{B}_k|$ is the number of instances in bin $\mathcal{B}_k$. However, when the outcome E_i is unknown for some subjects within a bin, an alternative estimator of the empirical risk is needed. One option estimates the probability of experiencing an event prior to time τ within each bin using the Kaplan-Meier estimator, yielding a calibration statistic of the form:

$$K = \sum_{k=1}^{m} \frac{(\bar{\pi}_k - \hat{p}_k^{KM})^2}{var(\hat{p}_k^{KM})}, \tag{3}$$

where $var(\hat{p}_k^{KM}) = \hat{S}_k(\tau)^2 \sum_{t_i < \tau} \frac{d_{ik}}{n_{ik} - d_{ik}}$, $\bar{\pi}_k$ is the average of predicted probabilities in bin k, $\hat{p}_k^{KM}$ is the Kaplan-Meier estimate of experiencing an event before τ among test subjects in bin k, $\hat{S}_k(\tau)$ is its corresponding survival rate (which equals $1 - \hat{p}_k^{KM}$), $var\{\hat{p}_k^{KM}\}$ is the sampling variance of the Kaplan-Meier estimator calculated using Greenwood's formula [Greenwood 1926], d_{sk} is the number of events occurring at time t_s in bin k, and n_{sk} are the number of people "at risk" for an event at time t_s (i.e., not censored and not experiencing an event before time t_i) in bin k. K is analogous to the χ^2 statistic with $m - 2$ degrees of freedom for assessing the calibration of logistic models [Hosmer and Lemesbow 1980; Lemeshow and Hosmer 1982]. Calibration plots can be used to compare predicted and Kaplan-Meier (i.e., empirical) probabilities of experiencing an event before τ within bins defined by ranges of predicted probabilities.

3.2. Concordance index

The area under the ROC curve (AUC) is a widely used summary measure of predictive model performance. When the outcome is fully observed on all subjects, it is equivalent to the concordance index (C-index), the probability of correctly ordering the outcomes for a randomly chosen pair of subjects whose predicted risks are different. Standard techniques for estimating the AUC/C-index are potentially biased when data are censored. However, as described in Harrell [2001], the C-index can be adapted for censoring by considering the concordance of survival outcomes versus predicted survival probability among pairs of subjects whose survival outcomes can be ordered, i.e., among pairs where both subjects are observed to experience an event, or one subject is observed to experience an event before the other subject is censored. Pairs in which both subjects are censored or in which the censoring time of one precedes the event time of the other do not contribute to this metric. Using notation introduced previously, the C-index adapted for censoring is given by

$$C_{cens}(\tau) = \frac{\sum_{i \neq j} \delta_i \mathbb{I}(V_i < V_j) \mathbb{I}\{\hat{\pi}(\mathbf{X}_i) < \hat{\pi}(\mathbf{X}_j)\}}{\sum_{i \neq j} \delta_i \mathbb{I}(V_i < V_j)}, \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

3.3. Net Reclassification Improvement

The C-index often fails to distinguish between models that differ in modest but clinically important ways. One proposed alternative is the Net Reclassification Improvement (NRI) [Pencina et al. 2008]. The NRI

compares the number of "wins" for each of two competing models among discordant predictions. The NRI is computed by cross-tabulating predictions from two different models with table cells defined by clinically meaningful cardiovascular risk categories or bins, then comparing the agreement of discordant predictions with actual event status. Formally, the NRI for comparing prediction models M_1 and M_2 using fully observed (i.e., not censored) binary event data is given by:

$$\text{NRI}(M_1, M_2) = \frac{E^{\uparrow}_{M_1} - E^{\uparrow}_{M_2}}{n_E} + \frac{\bar{E}^{\downarrow}_{M_1} - \bar{E}^{\downarrow}_{M_2}}{n_{\bar{E}}} \tag{5}$$

Here $E^{\uparrow}_{M_1}$ is the number of individuals in the test set who experienced events and were placed in a higher risk category by M_1 than M_2 (i.e., a number of "wins" for M_1 over M_2 among patients who had events), and the opposite change in risk categorization yields $E^{\uparrow}_{M_2}$). Similarly, $\bar{E}^{\downarrow}_{M_1}$ and $\bar{E}^{\downarrow}_{M_2}$ count the number of individuals who did not experience an event and were "down-classified" by M_1 and M_2, respectively (i.e., "wins" among patients who did not have events). n_E and $n_{\bar{E}}$ are the total number of patients with events and non-events, respectively. A positive $\text{NRI}(M_1, M_2)$ means better reclassification performance for M_1, while a negative $\text{NRI}(M_1, M_2)$ favors M_2.

To evaluate risk reclassification on test data which are subject to censoring, a "censoring-adjusted" NRI (cNRI) due to Pencina et al. [2011] takes the form:

$$\text{cNRI}(M_1, M_2) = \frac{E^{*,\uparrow}_{M_1} - E^{*,\uparrow}_{M_2}}{n^*_E} + \frac{\bar{E}^{*,\downarrow}_{M_1} - \bar{E}^{*,\downarrow}_{M_2}}{n^*_{\bar{E}}}, \tag{6}$$

where $E^{*,\uparrow}_{M_1}, E^{*,\downarrow}_{M_1}, E^{*,\uparrow}_{M_2}, E^{*,\downarrow}_{M_2}, n^*_E$ and $n^*_{\bar{E}}$ are analogous to the quantities in (5), but correspond to the expected number of subjects in each category, with the expectations computed using the Kaplan-Meier estimator to account for censoring.

One drawback of the NRI in (5) and (6) is that it weights reclassification improvement among events equal to that of non-events. While this may be reasonable in some applications, an alternative is to weight the reclassification improvement among events and non-events to the proportion of subjects experiencing the event and not experiencing the event, respectively.

4. APPLYING IPCW WITH EXISTING MACHINE LEARNING TECHNIQUES: 4 ILLUSTRATIONS

4.1. Logistic regression and generalized additive logistic regression

Logistic regression is a simple and popular technique for modeling binary or binomial data. The goal is a find a linear combination of features to approximate the log-odds, i.e.,

$$\log\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{7}$$

where $\pi(\mathbf{x}) = P(E = 1|\mathbf{X} = \mathbf{x})$ for the vector of features $\mathbf{X}$. The features may take any form, but in risk prediction the standard or "base" model often includes the main effects of each risk factor, i.e., the value of the risk factor itself. Given features $\mathbf{X}$ and a corresponding vector of event indicators $\mathbf{E}$, the logistic regression log-likelihood takes the form

$$\ell(\beta; \mathbf{X}, \mathbf{E}) = \sum_{i=1}^{n} \left[E_i \log \pi(\mathbf{X}_i) + (1 - E_i) \log\{1 - \pi(\mathbf{X}_i)\}\right], \tag{8}$$

where n is the number of observations in the training set. This log-likelihood can be maximized using a variety of techniques, the most common of which is iteratively reweighted least squares [Agresti 2012]. The solution of $\partial\ell/\partial\beta = \mathbf{0}$ is the unique maximum likelihood estimator of β.

Logistic regression using the "base" model with only the main (linear) effects of various risk factors is unlikely to produce a well-fitting model when the log odds of experiencing the event has a non-linear relationship with the features. Enlarging the feature set by considering a basis expansion of the continuous features may improve prediction. If $\mathbf{z}_j$ is the basis expansion of the j^{th} feature and $\boldsymbol{\beta}_j$ is vector of the same dimension as $\mathbf{z}_j$, then the generalized additive logistic model assumes that

$$\log\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \mathbf{z}_j, \tag{9}$$

The most intuitive expansion is the polynomial expansion where higher order moments of the j^{th} feature X_j, e.g. X_j^2, X_j^3 are included in the model. However, this particular expansion can be unstable, so restricted cubic smoothing splines, B-splines, or thin-plate regression splines are frequently used in practice. Since expanding the feature space involves estimating many more parameters, it is common to penalize the smoothness of the linear predictor $\sum_{j=1}^{p} \boldsymbol{\beta}_j^T \mathbf{z}_j$, and maximize the resulting penalized log-likelihood

$$\ell^P(\beta; \mathbf{X}, \mathbf{E}) = \sum_{i=1}^{n} \left[E_i \log \pi(\mathbf{x}) + (1-E_i)\log\{1-\pi(\mathbf{x})\}\right] - \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^T \mathbf{S}_j \boldsymbol{\beta}_j, \tag{10}$$

where $\beta = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T$, $\mathbf{S}_j$, $j = 1, \ldots, p$, are appropriately chosen smoothing matrices, and λ_j, $j = 1, \ldots, p$ are tuning parameters which control the degree of penalization/smoothness. λ_j are typically selected to minimize the unbiased risk estimator (UBRE), which in the case of logistic regression is proportional to the Akaike Information Criterion [Akaike 1974] given by $AIC = 2k - 2\ell$, where ℓ is the (log-)likelihood given in (8).

4.1.1. IPC-weighted logistic regression. IPC-weighted logistic regression maximizes the *weighted* log-likelihood

$$\ell^{\omega}(\beta; \mathbf{X}, \mathbf{E}) = \sum_{i=1}^{n} \omega_i \left[E_i \log \pi_i(\mathbf{x}) + (1-E_i)\log(1-\pi_i(\mathbf{x}))\right] \tag{11}$$

The weights are easily incorporated in standard statistical software. For example, in MATLAB, IPC weights can be used in the `weights` argument of the `glmfit` function. In R [R Core Team 2014], the `weights` argument of the `glm` command can be used, or IPC weights can be specified as sampling weights in `svyglm` from the `survey` package [Lumley 2004].

Similarly, the IPC-weighted generalized additive logistic regression maximizes the following weighted penalized log-likelihood:

$$\ell^{P,\omega}(\beta; \mathbf{X}, \mathbf{E}) = \sum_{i=1}^{n} \omega_i \left[E_i \log \pi(\mathbf{X}_i) + (1-E_i)\log\{1-\pi(\mathbf{X}_i)\}\right] - \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta_j}^T \mathbf{S}_j \boldsymbol{\beta_j}. \tag{12}$$

Most software programs which fit generalized additive models also easily incorporate observation weights. For example, in R, the `weights` argument of the `gam` function in the `mgcv` package [Wood 2006] can be used. The scores used to select the tuning parameters in the generalized additive model are also easily modified using IPC-weights to account for right censoring. In particular, the weighted AIC becomes $AIC^{\omega} = 2k - 2\ell^{\omega}$, with ℓ^{ω} given in (11).

As an alternative to estimating the parameters in the generalized additive logistic model using penalized maximum likelihood, Friedman et al. [2000] derived boosting procedures to construct flexible additive predictive models by iteratively maximizing ℓ. These approaches are easily generalized by maximizing ℓ^{ω} at each step instead of ℓ. Note that the IPC weights applied to the likelihood are distinct from the iteratively updated "case weights" used in boosting algorithms to increase the influence of poorly-classified instances, and in practice these two types of weights can be used in combination.

4.2. Bayesian networks

Bayesian networks have been used extensively in biomedical applications to: aid in understanding of disease prognosis and clinical prediction [Andreassen et al. 1999; Verduijn et al. 2007; Lipsky and Lewis 2005; Sarkar and Koehler 2013; Vila-Francés et al. 2013; Lappenschaar et al. 2013]; guide the selection of the appropriate treatment [Lucas et al. 2000; Kazmierska and Malicki 2008; Smith et al. 2009; Yet et al. 2013; Velikova et al. 2014]; and improve clinical decision support systems [Lucas et al. 1998; Sesen et al. 2013]. See Lucas et al. [2004] for a review.

The key to Bayesian network techniques is that using Bayes theorem one can rewrite $\pi(\mathbf{x})$ as

$$\pi(\mathbf{x}) = \frac{P_{\mathbf{X}|E}(\mathbf{x}|e=1)P_E(e=1)}{\sum_{e\in\{0,1\}} P_{\mathbf{X}|E}(\mathbf{x}|e)P_E(e)}, \tag{13}$$

so that focus is now shifted to estimation of the conditional density/probability $P_{\mathbf{X}|E}(\mathbf{x}|e)$ and the probability $P_E(e)$ for $e = 0, 1$. When E is observed on all subjects (i.e., there is no censoring), the maximum likelihood estimate of $P_E(e)$ is given by the sample mean of event indicators. To simplify the task of modeling $P_{\mathbf{X}|E}$, one can represent the joint distributions of $\mathbf{X}|E$ using a directed acyclic graph (DAG), i.e., a Bayesian network. One advantage of the Bayesian network approach is that clinical knowledge and data can be combined to suggest and refine DAG structures. The DAG encodes conditional independence relationships between variables, allowing the joint distribution to be decomposed into a product of individual terms conditioned on their parent variables [Russell and Norvig 2003]:

$$P_{\mathbf{X}|E}(\mathbf{x}|e) = \prod_{j=1}^{p} P_{X_j|\mathrm{Pa}(X_j),E}\{x_j|\mathrm{Pa}(x_j), e\} \tag{14}$$

where $\mathrm{Pa}(X_j)$ are the parents of X_j. Several approaches have been proposed to modeling the terms $P_{X_j|\mathrm{Pa}(X_j),E}\{x_j|\mathrm{Pa}(x_j), e\}$. In many applications, continuous covariates are discretized to allow us to learn the joint density of $P_{\mathbf{X}|E}(\mathbf{x}|e)$ nonparametrically and more easily. In the application considered in this paper, all parent nodes are discrete (or have been discretized) which simplifies the modeling considerably. If the j^{th} feature X_j is discrete, then $P_{X_j|\mathrm{Pa}(X_j),E}\{x_j|\mathrm{Pa}(x_j), e\}$ is estimated by computing the proportion of observations in each unique state of $\mathbf{X}_j$ separately for each level of $\mathrm{Pa}(X_j)$ and each level of E via

$$\hat{P}_{X_j|\mathrm{Pa}(X_j),E}\{x_j|\mathrm{Pa}(x_j), e\} = \frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_{ij} = x_j, \mathrm{Pa}(X_{ij}) = \mathrm{Pa}(x_j), E_i = e\}}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\mathrm{Pa}(X_{ij}) = \mathrm{Pa}(x_j), E_i = e\}}, \tag{15}$$

A number of parametric and semi-parametric approaches to modeling the joint covariate distributions are possible and have been described elsewhere [Domingos and Pazzani 1997; John and Langley 1995]; one common assumption is that the density of X_j given $\mathrm{Pa}(X_j)$ and E is a normal density (or a mixture of normal densities). If the number of parents is small, then one could estimate a separate mean and variance parameter for each level of $\mathrm{Pa}(X_j)$ and each level of E. In this case the maximum likelihood estimators of the mean, $\mu_{j,m,\mathrm{Pa}(x_j),e}$, and variance, $\Sigma_{j,m,\mathrm{Pa}(x_j),e}$, given parents $\mathrm{Pa}(x_j)$ and event status e are given by:

$$\begin{aligned}
\mu_{j,\mathrm{Pa}(x_j),e} &= \frac{\sum_{i=1}^{n} X_{ij}}{\sum_{i=1}^{n}\mathbb{I}\{\mathrm{Pa}(X_{ij}) = \mathrm{Pa}(x_{ij}), E_i = e\}} \\
\Sigma_{j,\mathrm{Pa}(x_j),e} &= \frac{\sum_{i=1}^{n}(X_{ij} - \mu_{j,\mathrm{Pa}(x_j),e})(X_{ij} - \mu_{j,\mathrm{Pa}(x_j),e})^T}{\sum_{i=1}^{n}\mathbb{I}\{\mathrm{Pa}(X_{ij}) = \mathrm{Pa}(x_{ij}), E_i = e\}},
\end{aligned} \tag{16}$$

If the number of parent nodes is large (or the nodes have several levels) there may be few observations in same combinations of $\mathrm{Pa}(X_j)$. Therefore, we may obtain more efficient estimators of the density of $X_j|\mathrm{Pa}(X_j), E$ by using a regression model. In particular, we might assume that the mean of the conditional density of X_j given $\mathrm{Pa}(X_j)$ and E is related to the levels of the parent nodes and event status through an additive model and that the conditional variance is constant across all levels $\mathrm{Pa}(X_j)$ and E. For example, if the m_j parents of X_j are denoted by $PX_{j1}, \ldots PX_{j,m_j}$ and the k^{th} parent has p_k levels denoted generically as $1, \ldots, p_k$ (recall that in our application the parent nodes are discrete), we could assume that $X_j|\mathrm{Pa}(X_j), E$ has mean $\beta_{0j} + \sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_k = l)$ and that the conditional variance σ_j^2 is constant across all levels of the parents. Then the log-likelihood takes the following form which can be solved to obtain the maximum likelihood estimators of $\beta_j = [\beta_{0j}, \{\beta_{jkl}\}_{k=1,\ldots,m_j,l=1,\ldots,p_k}]$ and σ_j^2:

$$\ell(\beta_j, \sigma_j^2; \mathbf{X}, \mathbf{E}) = \sum_{i=1}^{n}\left(\frac{-1}{2\sigma_j^2}\left[X_{ij} - \{\beta_{0j} + \sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_k = l)\}\right]^2 - \frac{1}{2}\log(2\pi\sigma_j^2)\right), \tag{17}$$

4.2.1. IPC-weighted Bayesian networks. To fit the Bayesian network using IPCW, we make the following modifications as discussed in Bandyopadhyay et al. [2014]. We first estimate IPC weights ω_j as described in Section 2.2. We can obtain the IPCW maximum likelihood estimator of $P_E(e)$ using the weighted mean $\hat{P}_E(e) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbb{I}(E_i = e)$. We note that this is equivalent to the Kaplan-Meier estimator of $P_E(e)$. Similarly, we can then obtain an IPCW maximum likelihood estimator of the distribution for the discrete

variables X_j separately for each level of $\text{Pa}(X_j)$ and each level of E:

$$\hat{P}_{X_j|\text{Pa}(X_j),E}\{x_j|\text{Pa}(x_j),e\} = \frac{\frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbb{I}\{X_{ij}=x_j,\text{Pa}(X_{ij})=\text{Pa}(x_j),E_i=e\}}{\frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_j)E_i=e\}}. \tag{18}$$

The IPCW estimators of the mean, variance, and mixing parameters for continuous variables X_j can be obtained using a weighted maximum likelihood where the contribution of the i^{th} subject to the likelihood is weighted by ω_i. The formulas for the parameter estimates previously given in Equation (16) become

$$\begin{aligned}\mu_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^{n}\omega_i X_{ij}}{\sum_{i=1}^{n}\omega_i\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}} \\ \Sigma_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^{n}\omega_i(X_{ij}-\mu_{j,\text{Pa}(x_j),e})(X_{ij}-\mu_{j,\text{Pa}(x_j),e})^T}{\sum_{i=1}^{n}\omega_i\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}.\end{aligned} \tag{19}$$

Similarly, the weighted log-likelihood for the regression parameters becomes:

$$\ell^{\omega}(\beta_j,\sigma_j^2;\mathbf{X},\mathbf{E}) = \sum_{i=1}^{n}\omega_i\left(-\frac{1}{2\sigma_j^2}\left[X_{ij}-\{\beta_{0j}+\sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_{ik}=l)\}\right]^2-\frac{1}{2}\log(2\pi\sigma_j^2)\right), \tag{20}$$

We note that tuning a Bayesian network for optimal performance may involve determining the network structure and/or controlling model complexity for a given structure. In the Bayesian network implementation for our data application, we consider only a single network structure which is informed by discussions with our clinical colleagues (see Figure 3); however, a set of feasible structures could easily be compared on a test set or via cross-validation using the calibration and reclassification metrics described in Section 3.

A more flexible modeling approach for modeling the continuous density of X_j conditional on the levels of $\text{Pa}(X_j)$ and E is to assume that the conditional density can be represented as a mixture of M multivariate normal densities. We could consider M to be a tunable parameter, and either select a single value or, as suggested in Bandyopadhyay et al. [2014], use a model averaging procedure to combine results across multiple values of M.

4.3. Decision trees

Recursive partitioning is a powerful and flexible way to build predictive models for both discrete and continuous outcomes, and decision tree algorithms are widely applied in biomedicine [see Mansiaux and Carrat 2014; Liu et al. 2014; Hartney et al. 2014; Muñoz-Moreno et al. 2014; Abdollah et al. 2014, and references therein]. Decision trees aim to partition training data into subgroups with homogeneous outcomes, with subgroups defined by a set of binary splits of the features. The prediction for a given test instance is made by identifying the partition or node it belongs to, then computing a summary statistic (e.g., the sample average) for training instances in that partition or node.

Many techniques have been proposed to grow decision trees, mostly differing in the criteria used to decide how/if to split a node and to prevent overfitting. One popular technique, CART [Breiman et al. 1984], uses the decrease in Gini impurity to determine which feature and at what level to split a node. The change in Gini impurity for a splitting rule is given by

$$\Delta I_G(S) = \hat{\pi}(S)\{1-\hat{\pi}(S)\} - \frac{1}{N_S}\left[N_{S_L}\hat{\pi}(S_L)\{1-\hat{\pi}(S_L)\}+N_{S_R}\hat{\pi}(S_R)\{1-\hat{\pi}(S_R)\}\right]$$

where $\hat{\pi}(S), \hat{\pi}(S_L)$, and $\hat{\pi}(S_R)$ are respectively the sample proportion of outcomes in a node S, the node's left-hand children S_L, and the node's right-hand children S_R for the particular splitting rule; N_{S_L} and N_{S_R} are the number of instances in each child; and $N_S = N_{S_L} + N_{S_R}$. The C4.5 and C5.0 decision trees [Quinlan 1993; Kuhn and Johnson 2013] use the information gain metric instead of the Gini impurity to make decisions on how to split each node:

$$\Delta I_E(S) = \hat{I}_E(S) - \frac{1}{N_S}\{N_{S_L}\hat{I}_E(S_L)+N_{S_R}\hat{I}_E(S_R)\}$$

where the information in node S is given by

$$\hat{I}_E(S) = -\hat{\pi}(S)\log\hat{\pi}(S) - \{1-\hat{\pi}(S)\}\log\{1-\hat{\pi}(S)\}$$

and similarly for $I_E(S_L)$ and $I_E(S_R)$. In the unweighted case, the sample proportions $\hat{\pi}(S)$ for node S are computed using the usual nonparametric maximum likelihood estimators, i.e.:

$$\hat{\pi}(S) = \frac{1}{N_S}\sum_{i\in S} E_i, \quad \hat{p}(S_L) = \frac{1}{N_{S_L}}\sum_{i\in S_L} E_i, \quad \hat{p}(S_R) = \frac{1}{N_{S_R}}\sum_{i\in S_R} E_i$$

where E_i is the binary event indicator. For a test instance with features $\mathbf{x}$ falling in tree node z, we can estimate the risk $\pi(\mathbf{x})$ as

$$\hat{\pi}(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(E_i = 1, Z_i = z)}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Z_i = z)}$$

where $Z_i = z$ indicates that training instance i belongs to node z.

4.3.1. IPC-weighted decision trees. It is straightforward to extend decision trees to incorporate IPC weighting: individual cases in the training set are assigned weights ω_i as described above, and the ω_i are used as "case weights" in the decision tree algorithm. With IPC weighting, we calculate a weighted decrease in Gini impurity,

$$\Delta I_G^{\omega}(S) = \hat{\pi}^{\omega}(S)\{1-\hat{\pi}^{\omega}(S)\} - \frac{1}{N_S^{\omega}}\left[N_{S_L}^{\omega}\hat{\pi}^{\omega}(S_L)\{1-\hat{\pi}^{\omega}(S_L)\} + N_{S_R}^{\omega}\hat{\pi}^{\omega}(S_R)\{1-\hat{\pi}^{\omega}(S_R)\}\right]$$

where

$$N_S^{\omega} = \sum_{i\in S}\omega_i, \quad N_{S_L}^{\omega} = \sum_{i\in S_L}\omega_i, \quad N_{S_R}^{\omega} = \sum_{i\in S_R}\omega_i,$$

and

$$\hat{\pi}^{\omega}(S) = \frac{\sum_{i\in S}\omega_i E_i}{N_S^{\omega}}, \quad \hat{\pi}^{\omega}(S_L) = \frac{\sum_{i\in S_L}\omega_i E_i}{N_{S_L}^{\omega}}, \quad \hat{\pi}^{\omega}(S_R) = \frac{\sum_{i\in S_R}\omega_i E_i}{N_{S_R}^{\omega}},$$

The identical approach can be applied to estimate a weighted version of the information gain metric:

$$\Delta I_E^{\omega}(S) = \hat{I}_E^{\omega}(S) - \frac{1}{N_S^{\omega}}\{N_{S_L}^{\omega}\hat{I}_E^{\omega}(S_L) + N_{S_R}^{\omega}\hat{I}_E^{\omega}(S_R)\}$$

with

$$\hat{I}_E^{\omega}(S) = -\hat{\pi}^{\omega}(S)\log\hat{\pi}^{\omega}(S) - \{1-\hat{\pi}^{\omega}(S)\}\log\{1-\hat{\pi}^{\omega}(S)\}$$

and similarly for $I_E^{\omega}(S_L)$ and $I_E^{\omega}(S_R)$.

Once the structure of the tree has been determined, the predicted risk of a test instance with features $\mathbf{x}$ falling in terminal node z is estimated as the weighted (non-parametric) maximum likelihood estimator:

$$\hat{\pi}(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbb{I}(E_i = 1, Z_i = z)}{\frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbb{I}(Z_i = z)}$$

Because of their flexibility, classification trees often overfit training data. Many overfitting avoidance techniques have been proposed, with most involving a tuning parameter which restricts the complexity of the tree. One strategy consists of setting a lower limit m on the number of individuals assigned to a terminal node; in our notation above, the node S would not be split according to a given rule unless $\min(N_{S_L}, N_{S_R}) \geq m$. This strategy is easily generalized to the case with censoring by requiring that $\min(N_{S_L}^{\omega}, N_{S_R}^{\omega}) \geq m$; however we note that $N_{S_L} \approx N_{S_L}^{\omega}$ and $N_{S_R} \approx N_{S_R}^{\omega}$ as the expected value of the weights is one, so in practice $\min(N_{S_L}, N_{S_R})$ is usually sufficient. Another approach only pursues splits where the information gain exceeds a certain threshold θ, e.g., $\Delta I_E(S) \geq \theta$. Substituting $\Delta I_E^{\omega}(S)$ for $\Delta I_E(S)$ (and similarly $\Delta I_G^{\omega}(S)$ for $\Delta I_G(S)$) allows the same rule to be used in the censored data setting. Final tuning parameter values may be chosen by cross-validation, where the cross-validated criterion to optimize could involve a measure of calibration, discrimination (C-index/NRI), or a combination of both.

4.4. k-nearest neighbors

The k-nearest neighbors classifier is widely used in biomedical applications and provides a flexible, powerful, and intuitive method for risk prediction [Gürgen and Gürgen 2003; Chen et al. 2007; Parry et al. 2010; Arif

et al. 2012; Acharya et al. 2012; Amini et al. 2013; Cheng and Zhao 2014]. Define $d(\mathbf{X}_i, \mathbf{X}_j)$ to be a distance metric between two vectors of features $\mathbf{X}_i$ and $\mathbf{X}_j$. To estimate the event probability for an instance in the test set with features $\mathbf{X} = \mathbf{x}$, define $R_i(\mathbf{x})$ to be the rank of the distance between $\mathbf{x}$ and $\mathbf{X}_i$, i.e., $d(\mathbf{x}, \mathbf{X}_i)$, among all n observations in the training data set. Here smaller ranks indicate that the distance between the training instance and the test instance is smaller. In the scenario when the event status is known on all subjects in the training dataset, in the most straightforward application of k-nearest neighbors, $\pi(\mathbf{x})$ is simply the proportion of the training instances experiencing the event among those with $R_i \leq k$. That is,

$$\hat{\pi}(\mathbf{x}) = \sum_{i=1}^{n} \frac{E_i \mathbb{I}(R_i(\mathbf{x}) \leq k)}{k} = \frac{\sum_{i=1}^{n} E_i \mathbb{I}(R_i(\mathbf{x}) \leq k)}{\sum_{i=1}^{n} \mathbb{I}(R_i(\mathbf{x}) \leq k)} \tag{21}$$

The key choice for implementing a k-nearest neighbor classifier is to select an appropriate distance metric and the number of neighbors to consider. The number of neighbors to consider may be treated as a tuning parameter and chosen using cross-validated estimates of some appropriate criterion. Similarly, a wide variety of distance functions could be considered and chosen in a data-driven manner using cross-validation. Some standard choices for the distance metric include the Euclidean distance (of the standardized features) and the Mahalanobis distance, among many others. Alternatively, we could compute the distance in a projected canonical space. To compute this distance, we find the linear combination of the features that has the largest correlation with E. We can then find the distance between two observations based on the distance between the linear combination of features described above. This was the best performing distance metric in our application. More sophisticated methods for computing $\pi(\mathbf{x})$ could include down weighting the training observations in (21) which have larger distances rather than considering all k nearest neighbors equally.

4.4.1. IPC-weighted k-nearest neighbors. To adapt a k-nearest neighbor classifier to the situation when E_i may not be known for all subjects, we note that the distance between the vector of features is not affected by censoring. That is, we assume that the features do not vary with time and are all measured prior to the start of follow-up. Therefore, to predict the probability of an event for an instance in the test set, we can identify the k closest neighbors in the training set just as we did before. However, we now replace the sample average in (21) with an average weighted by the IPC weights ω_i:

$$\hat{\pi}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \omega_i E_i \mathbb{I}(R_i(\mathbf{x}) \leq k)}{\sum_{i=1}^{n} \omega_i \mathbb{I}(R_i(\mathbf{x}) \leq k)}. \tag{22}$$

Note that in this extension of the k-nearest neighbor classifier, we choose the k neighbors regardless of the value of the IPC weights, ω_i, for those neighbors in the training set. Therefore, the total sum of the weights for the k neighbors $\sum_{i=1}^{n} \omega_i \mathbb{I}(R_i(\mathbf{x}) \leq k)$ may be different and the number of training instances with non-zero weight among those k neighbors may vary depending on the features of the test instance. However, we note that the expected value of the IPC weight, ω_i, is equal to one, independent of the features $\mathbf{X}_i$ of the training instance. Therefore, for a large number of neighbors, $\sum_{i=1}^{n} \omega_i \mathbb{I}(R_i(\mathbf{x}) \leq k)$ should be approximately equal across different values of $\mathbf{x}$, and we do not have to worry about adjusting the number of neighbors across the feature space. However, the number of neighbors used in the classifier still should be chosen using cross-validation.

5. STATISTICAL VALIDITY OF IPCW

We briefly argue why inverse probability of censoring weighting appropriately handles censoring and leads to well-calibrated risk prediction across a variety of machine learning techniques. More formal proofs are given in Robins and Finkelstein [2000]; Bang and Tsiatis [2000, 2002]; Rotnitzky and Robins [2004]; Tsiatis [2006]. If there were no censoring, class probability estimates (across all the machine learning scenarios considered in this manuscript) would be obtained using some form of maximum likelihood estimator (e.g., non-parametric, penalized, etc.). For example, in logistic regression, we parameterize the log odds (i.e., $\log[\pi_i(\mathbf{X})/\{1 - \pi_i(\mathbf{x})\}]$) in terms of a linear combination of $\mathbf{x}$ and any non-linear and interaction terms specified *a priori* and estimate the coefficients using maximum likelihood. In generalized additive logistic models, the log odds are related to a linear combination of $\mathbf{z}$, a basis expansion of $\mathbf{x}$, and regression coefficients are estimated using penalized maximum likelihood. In Bayesian networks, the continuous components of are modeled via maximum likelihood using a mixture of Gaussian densities, and the discrete components are modeled non-parametrically. Once we identify the terminal nodes in binary trees or the neighbors in k-nearest neighbors, we compute probability estimates by taking sample averages within nodes; these averages are non-parametric maximum likelihood estimators. It has been well established that, as the sample size tends to

infinity and other regularity conditions hold (including the number of observations in terminal nodes and number of neighbors also increase), maximum likelihood consistently estimates the risk probability [Sundberg et al. 1972; Lehmann and Casella 1998; Loh 2008; Boos and Stefanski 2013].

Inverse probability of censoring weighted observations "works" by approximating the log-likelihood that we would have obtained had there not been censoring. Consider the general likelihood $\ell(\beta; \mathbf{X}, \mathbf{E}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\beta; \mathbf{X}_i, E_i)$ for the parameter β had no observations been right-censored. By the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \ell_i(\beta; E_i, \mathbf{Z_i}) \xrightarrow{p} \mathcal{E}\{\ell_i(\beta; E_i, \mathbf{Z_i})\} \tag{23}$$

where $\mathcal{E}(\cdot)$ is the expectation and $\xrightarrow{p}$ denotes convergence in probability.

In the case of IPCW estimators, we maximize the likelihood $\frac{1}{n}\sum_{i=1}^{n} \omega_i \ell_i(\beta; E_i, \mathbf{Z_i})$. Note that we have

$$\omega_i = \mathbb{I}[\min(T_i, \tau) < C_i] / \hat{G}\{\min(T_i, \tau)\} \xrightarrow{p} \mathbb{I}[\min(T_i, \tau) < C_i] / G\{\min(T_i, \tau)\} \tag{24}$$

and hence

$$\begin{aligned}
\mathcal{E}\left\{\frac{1}{n}\sum_{i=1}^{n} \omega_i \ell_i(\beta; \mathbf{X}_i, E_i)\right\} &= \mathcal{E}\left\{\frac{\mathbb{I}\{\min(T_i, \tau) < C_i\}}{G\{\min(T_i, \tau)\}} \ell_i(\beta; \mathbf{X}_i, E_i)\right\} \\
&= \mathcal{E}\left(\mathcal{E}\left[\frac{\mathbb{I}\{\min(T_i, \tau) < C_i\}}{G\{\min(T_i, \tau)\}} \ell_i(\beta; E_i, \mathbf{Z_i}) | \mathbf{X}_i, T_i\right]\right) \\
&= \mathcal{E}\left(\frac{\mathcal{E}[\mathbb{I}\{\min(T_i, \tau) < C_i\} | \mathbf{X}_i, T_i]}{G\{\min(T_i, \tau)\}} \ell_i(\beta; \mathbf{X}_i, E_i)\right) \\
&= \mathcal{E}\left(\frac{\mathcal{E}[\mathbb{I}\{\min(T_i, \tau) < C_i\}]}{G\{\min(T_i, \tau)\}} \ell_i(\beta; \mathbf{X}_i, E_i)\right) \\
&= \mathcal{E}\left(\frac{\mathcal{E}[G\{\min(T_i, \tau)\}]}{G\{\min(T_i, \tau)\}} \ell_i(\beta; \mathbf{X}_i, E_i)\right) \\
&= \mathcal{E}\left\{\ell_i(\beta; \mathbf{X}_i, E_i)\right\}
\end{aligned}$$

That is, for large samples the IPCW log-likelihood converges to the same quantity as the fully-observed (i.e., uncensored) likelihood. Because the difference in the IPCW and fully-observed likelihoods are (asymptotically) negligible, in large samples the IPCW approach inherits all the properties of machine learning estimators if we had full data. The above argument relies on the assumption that the censoring time C is independent of the event time T and all features $\mathbf{X}$. In our application, most patients are censored due to the end of the study or because they disenroll from the insurance plan due to a change in employment, reasons unrelated to their health status (i.e., $\mathbf{X}$ and T), so this independence assumption is reasonable. How to handle this so-called "dependent censoring" is a current area of research in statistics [Tsiatis 2006], and to our knowledge very little of this work has been applied in the machine learning domain.

6. EXAMPLE APPLICATION: PREDICTING CARDIOVASCULAR RISK USING ELECTRONIC HEALTH DATA

We now illustrate the application of IPC-weighted risk prediction methods to the problem of predicting the risk of a cardiovascular event from electronic health data. The data come from a healthcare system in the Midwestern United States and were extracted from the HMO Research Network Virtual Data Warehouse (HMORN VDW) associated with that system [Selby 1997; Platt et al. 2001; Maro et al. 2009]. The VDW stores data including insurance enrollment, demographics, pharmaceutical dispensing, utilization, vital signs, laboratory, census, and death records. This healthcare system includes both an insurance plan and a medical care network in an open system which is partially overlapping. That is, patients of the insurance plan may be served by either the internal medical care network and or by external healthcare providers, and the medical care network serves patients within and outside of the insurance plan. Patient-members who do not visit any of the clinics and hospitals in-network do not have any medical information (e.g., blood pressure information) included in the electronic medical record (EMR) of this system. Furthermore, once the patient-member disenrolls from the insurance plan, the patient is right-censored as there is no longer any information on risk factors or outcomes (i.e., CV events) recorded in the EMR or insurance claims data.

6.1. Defining the study population

The study population was initially selected from those enrolled in the insurance plan between 1999 and 2011 and who had at least one outpatient medical encounter at an in-network clinic. From this initial database of 448,306 subjects, an analysis population of was identified by applying the following inclusion/exclusion criteria:

(1) To ensure sufficient time to collect baseline risk factors on subjects, the analysis was restricted to those subjects with at least one year of continuous insurance enrollment. Some of the patients were sporadically enrolled during the period of study; however, for the purpose of our analysis, we ignored gaps in enrollment less than 90 days and considered a patient-member continuously enrolled over this period. These gaps in enrollment are likely due to administrative errors or patients changing employers but still electing coverage with the same insurance provider.
(2) We included only patients with two medical encounters in the in-network clinic with blood pressure information at least 30 days but at most 1.5 years apart and with drug coverage.
(3) Patients under the age of 40 were excluded.
(4) Subjects with pre-existing serious comorbidities other than diabetes (e.g., prior CV event, chronic kidney disease, etc.) were excluded.

These criteria ensure that the analysis population consists of a diverse group of subjects at non-trivial risk of experiencing a CV event, who were treated routinely in the primary care clinic. Patients who are only infrequently treated in the emergency room or urgent care clinics (i.e., settings where patients are unlikely to be counseled about their CV risk) were not of interest in this analysis. Subjects with comorbidities were excluded because comorbidity information was recorded inconsistently and because, as a group, these subjects have an extremely high 5-year event rate. Therefore, these subjects are of limited interest for risk prediction models because they are already almost uniformly being aggressively treated to manage their cardiovascular risk. After applying the above criteria, our final analysis dataset contained 87,363 individuals.

The available longitudinal data on each patient-member was divided into: (i) a *baseline* period, where the risk factors were ascertained, and (ii) a *follow-up* period, where we assessed whether a patient experienced a CV event (and, if so, when). The baseline period consisted of the time between the first blood pressure reading during the enrollment period and the date of the final blood pressure reading at most 1.5 years from the first measurement. The follow-up period for a patient begins at the end of the baseline period, referred to as the index date, and continues until either the patient experiences a CV event (defined below), the patient disenrolls from the insurance plan for more than 90 days, or the data capture period ends (in 2011), whichever comes first. The distribution of the follow-up periods for the resulting analysis cohort is shown in Figure 1, which illustrates that a large proportion of subjects' CV event times are censored prior to the end of follow-up. Figure 2 shows that, unless we consider a very short time horizon τ, the τ-year event status will be unknown for a substantial proportion of subjects in this cohort.

6.2. Risk factor ascertainment

Risk factors used as features in the machine learning models included age, gender, systolic blood pressure (SBP), use of blood pressure medications, cholesterol markers (HDL and total cholesterol), body mass index (BMI), smoking status, and presence/absence of diabetes. Summary statistics and brief descriptions for the risk factors are given in Table I. Missing risk factor values were filled in prior to model fitting using multiple imputation by chained equations [van Buuren and Groothuis-Oudshoorn 2011] to create a dataset with no mising values; Table I displays the percentage of missing values for each risk factor in the original (pre-imputation) data set.

6.3. Events and censoring

Cardiovascular events were defined as the first recorded stroke, myocardial infarction (MI), or procedure proximal to stroke or MI (e.g., coronary artery bypass surgery, stent for either the coronary arteries or carotid artery) after the baseline period, prior to 5 years of follow-up. This information was obtained from diagnosis codes recorded by physicians or inferred from procedures (such as bypass surgery or stent placement) performed on an individual. In addition to using procedure and diagnosis codes to infer if a CV event occurred, we considered a patient to have experienced a CV event if the cause of death listed on the death certificate included MI or stroke. The total number of first CV events recorded within 5 years of the baseline period was 3,653; the 5-year event rate for the entire analysis cohort calculated via Kaplan-Meier was 6.4%.

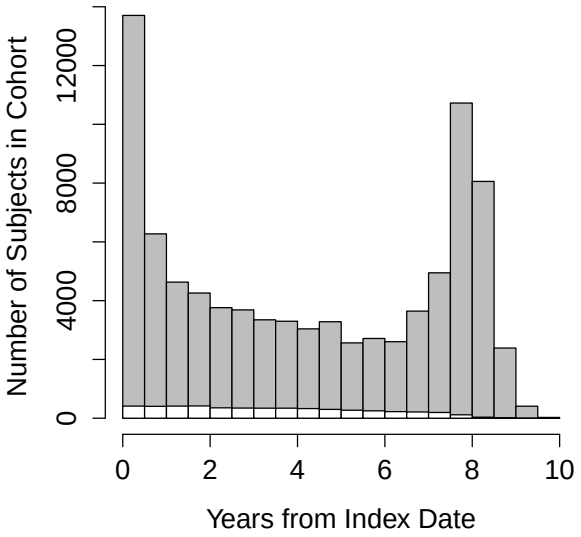

Fig. 1. Distribution of follow-up times, i.e., time from the end of the baseline period until the patient experiences a CV event, the patient disenrolls from the insurance for more than 90 days, or the study ends, in our entire cohort after applying inclusion/ exclusion criteria detailed in Section 6.1. The number of subjects whose follow-up ends in a CV event are shown in white bars while the number whose follow-up is censored is given by the gray bars.

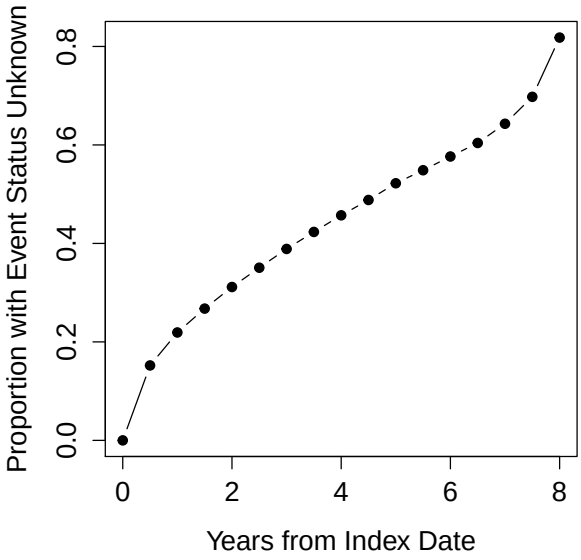

Fig. 2. Proportion of subjects with unknown τ-year event status as a function of τ, the time from index date in years.

7. MODELS AND RESULTS

Subjects who experienced an event within five years were recorded as $E = 1$, and those with at least 5 years of event-free follow-up were recorded as $E = 0$. Subjects who were event-free but censored before accruing 5

Table I. Distribution of risk factors in the analysis dataset.

Feature Name	**Median (IQR)** or **N (%)**	**% missing (in original data)**	**Description**
Gender			
Female	51,530 (59.0)	0	
Male	35,833 (41.0)	0	
Age (Years)	52 (46 - 60)	0	Age at the end of the baseline period
SBP (mm Hg)	123 (115 - 133)	0	Average systolic blood pressure during baseline period
BMI (kg/m^2)	28.0 (24.7 - 32.3)	8	Body mass index
HDL (mg/dL)	48 (40 - 59)	41	Final high density lipoprotein cholesterol during baseline period
Total cholesterol (mg/dL)	196 (172 - 222)	41	Final total cholesterol during baseline period
Smoking			Smoking status in EMR
Never or Passive	64,335 (73.6)	0	
Quit	9,829 (11.3)	0	
Current	13,199 (15.1)	0	
SBP Meds			Subject is currently taking SBP medication during baseline period
No	49,165 (56.3)	0	
Yes	38,198 (43.7)	0	
Diabetes			Subject has a current diagnosis of diabetes
No	80,921 (92.6)	0	
Yes	6,442 (7.4)	0	

years of follow-up have E unknown. We applied and evaluated four variants of each of the machine learning techniques described in Section 4 to our data. The variants differ in their handling of subjects with E unknown:

(1) **Set $E = 0$ if E is unknown**. Techniques using this strategy are denoted with the suffix *-Zero*.
(2) **Discard observations with E unknown**. Techniques using this strategy are given the suffix *-Discard*.
(3) **Use IPCW on observations with E known**. The resulting techniques, as described in Section 4, have the suffix *-IPCW*.
(4) **"Split" observations with E unknown into two observations with $E = 1$ and $E = 0$ with weights based on marginal survival probability**. The resulting techniques, as described subsequently, have the suffix *-Split*.

The final technique of splitting observations with E unknown was described by Štajduhar and Dalbelo-Bašić [2010]. For each observation i in the training set for which E_i is unknown, we create two observations, one with $E = 1$ and the other with $E = 0$, but with the same features $\mathbf{X}_i$. Let $\hat{F}(t)$ be the Kaplan-Meier estimator of the survival probability at time t,

$$\hat{F}(t) = \prod_{j:V_j<t} \left(\frac{n_j - d_j}{n_j} \right) \tag{25}$$

where d_j is the number of subjects who are observed to experience the event at time V_j, and n_j is the number of subjects "at risk" for the event (i.e., not previously censored or experiencing an event) at time V_j. Then, if E is unknown for instance i in the training set, the weight for the imputed observation with $E = 0$ is $\hat{F}(\tau)/\hat{F}(V_i)$ (an estimate of the conditional probability that $E = 0$), and the weight for the imputed observation with $E = 1$ is $1 - \hat{F}(\tau)/\hat{F}(V_i)$. The weights are implemented in the analysis in the same way as the IPC weights. These weights are advantageous because all observations receive non-zero weights and are used in the analysis.

Results from the Cox proportional hazards model [Cox 1972] are included for comparison. The Cox model is a well-established technique for estimating survival probabilities which has been used as the basis for many CV risk prediction models, including the popular Framingham risk score [D'Agostino et al. 2008]. If the Cox regression model is correctly specified, the estimated survival probabilities converge to the true probabilities as the sample size increases.

Model performance was assessed based on the calibration and discrimination metrics described in Section 3. To calculate the calibration statistic and cNRI, we defined five risk strata based on clinically relevant cutoffs for the risk of experiencing a cardiovascular event within 5 years: 0-5%, 5-10%, 10-15%, 15-20% and $> 20\%$ [Ridker et al. 2007]. For the cNRI, risk predictions for an individual were considered discordant

between two models if the predictions fell in different ranges. Asymptotically, the calibration statistic has a χ^2 distribution with 3 degrees of freedom, so a statistical test for the null hypothesis that a model is well-calibrated would fail to reject at a 5% significance level if the statistic exceeds 7.81. All models were fitted using the open-source statistical software program R [R Core Team 2014]. Code is available from the first author's website, https://sites.google.com/site/dmvock/.

We now provide some implementation details for the various machine learning techniques.

7.1. Logistic regression and generalized additive logistic regression

For the logistic regression models, all (unscaled) risk factors described in Section 6.2 were included as additive factors in the model for the log odds of having a CV event. The reported results are for models with a single "main effect" term for each predictor (i.e., no interactions or transformations); predictive performance did not markedly improve when second-order interaction terms were included (data not shown). Models were fitted using the `glm` function in R; IPC weights were incorporated using the `weights` argument.

The generalized additive models included the same risk factors as those in the linear regression model. However, we allowed the effect of the continuous covariates on the log odds to vary smoothly by using low rank thin plate regression splines for each covariate. The smoothing penalty was chosen using generalized cross-validation to minimize UBRE (which is related to AIC). Models were fitted using the `gam` function in the `mgcv` package in R; IPC weights were incorporated using the `weights` argument.

7.2. Bayesian networks

Figure 3 displays the structure of the Bayesian network that we used to construct our prediction models. The structure was determined by combining known relationships from the medical literature with input from our clinical colleagues. As noted in Section 4.2, it is possible to use IPCW to account for censoring when building and comparing different graph structures.

Given the complex relationship between age and body mass index and other covariates, we discretized those covariates. In particular, we considered the age categories 40-50, 50-60, 60-70, 70-80, and >80 and BMI categories < 25, 25-30 (overweight), 30-35 (class I obesity), 35-40 (class II obesity), > 40 (class III obesity).

Nodes were jointly modeled as described in Section 4.2. To model the distribution of SBP, HDL, TC, we considered linear regression models with the parents of those nodes as additive predictors in the model. Additionally, the model for SBP included an interaction between BMI category and SBP medication.

7.3. Classification trees

Classification trees were built using the `rpart` package in R, which implements the classification and regression trees described in Breiman et al. [1984]. Nodes are split based on the Gini loss criterion. We considered the ratio of the loss between misclassifying events and non-events, the minimum number of subjects in each terminal node, and the cost complexity parameter as tuning parameters which were chosen using five-fold cross-validation over a grid of values for those parameters. We selected the most parsimonious tree which had an average C-index in the hold-out sets within one standard error of the best combination of those tuning parameters. The loss matrix was modified to give more weight to incorrect non-event prediction among those experiencing events than the reverse, to induce additional splits and improve discrimination among the large fraction of the population with a relatively low (e.g., $< 5\%$) 5-year CV event risk. The ratio of the loss for incorrect non-event predictions to incorrect event prediction considered in the grid search ranged from 2.5 to 10. The minimum number of subjects in each terminal node was also varied among 50, 100, and 200. Finally, in the cross-validation analysis the cost complexity parameter ranged across a fine grid between 10^{-1} and 10^{-4}. Risk factors were not scaled prior to fitting the tree. IPC weights were incorporated via the `weights` argument in `rpart`, which treats them as case weights.

7.4. k-nearest neighbors

Classification using k-nearest neighbors was done using the `yaImpute` package in R to identify efficiently the k neighbors for each instance in the test set. We found that computing the distance between the features in the projected canonical space works well in this application, and those results are reported here. The number of neighbors was considered as a tuning parameter and selected using five-fold cross-validation. In particular, we selected the largest number of neighbors (more neighbors is equivalent to a more parsimonious model) which had an average C-index in the hold-out sets within one standard error of the best C-index. A maximum of 1,000 neighbors was considered to improve computational speed as would likely be done in common practice.

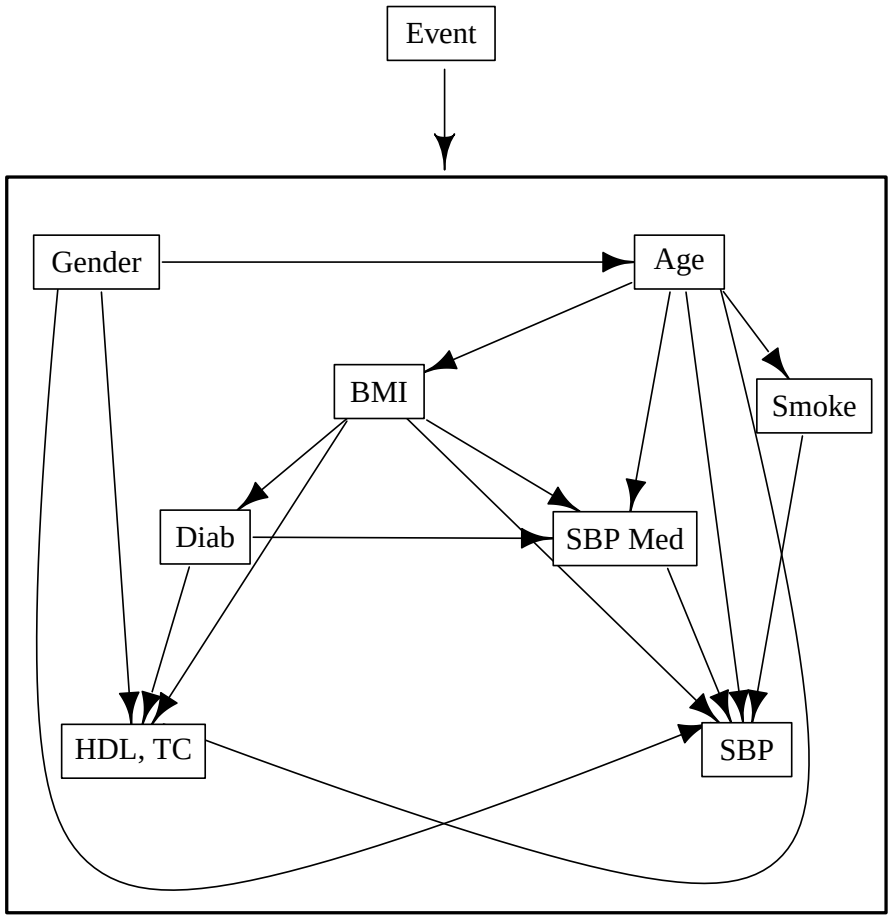

Fig. 3. The graphical model for our Bayesian network for CV risk prediction. The figure includes the structure of risk factors, conditioned on the CV event status. In particular, nodes represent input variables and edges represent conditional dependencies between the variables. The edge between subgraphs indicates an edge from every node in the source subgraph to every node in the destination subgraph or node. That is, the reader should assume that our outcome variable (Event) is connected to every node in the graph. Features in the same nodes indicate those features are modeled jointly. The full description of each of the features appears in Section 6. *Smoke*: current smoking status of patient; *BMI*: body mass index of patient; *Diab*: indicator for whether or not patient is diabetic *SBP*: systolic blood pressure; *SBP Med*: indicator for whether or not patient is prescribed blood pressure medication; *HDL*: high density lipoprotein cholesterol; *TC*: total cholesterol (note that HDL and TC are modelled jointly).

7.5. Cox regression

The Cox model specifies the relationship between the risk factors and the conditional hazard function $\lambda(t; \mathbf{X}) = f_{T|\mathbf{X}}(t|\mathbf{x})/S_{T|\mathbf{X}}(t|\mathbf{x})$, where $f_{T|\mathbf{X}}$ is the conditional density function and $S_{T|\mathbf{X}}$ the conditional survival function of T. As is standard, we take

$$\lambda(t; \mathbf{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots)$$

where the risk factors $X_1, X_2, \dots$ are the same as for the logistic regression model. The resulting model is fairly similar to the one used to compute the Framingham risk score, providing an established standard against which to compare our models' calibration and reclassification performance.

8. RESULTS

The full training dataset consists of 65,522 patients (75%) drawn at random from the analysis cohort. 52% were censored prior to five years; as a result *-Discard* models were trained on 31,345 subjects. The performance of all models is evaluated based on the risk predictions of the remaining 21,841 patients not included in any training set.

Table II. Calibration statistic and C-index of versions of classification trees (*Tree*), k-nearest neighbors (*k-NN*), Bayesian network models (*Bayes*), logistic regression (*Logistic*), generalized additive models (*GAM*), and Cox proportional hazards model (*Cox*) evaluated on the hold-out test set. *Predicted event rate*: Average predicted probability of experiencing a CV event within 5 years; *Calibration*: calibration test statistic K; *C-index*: Concordance index adapted for censoring.

Method	Predicted event rate (%)	Calibration	C-Index
Tree			
-IPCW	5.41	12.74	0.788
-Discard	7.13	76.92	0.784
-Zero	4.19	125.76	0.784
-Split	6.42	289.54	0.782
k-NN			
-IPCW	5.27	10.24	0.787
-Discard	7.07	49.11	0.793
-Zero	4.11	85.64	0.788
-Split	6.37	106.60	0.787
Bayes			
-IPCW	5.62	6.18	0.802
-Discard	7.40	76.82	0.802
-Zero	4.26	80.16	0.800
-Split	6.49	194.56	0.801
Logistic			
-IPCW	5.40	4.85	0.801
-Discard	7.14	63.92	0.801
-Zero	4.18	83.78	0.799
-Split	6.42	150.46	0.797
GAM			
-IPCW	5.47	6.96	0.805
-Discard	7.22	67.57	0.804
-Zero	4.17	83.04	0.801
-Split	6.42	233.07	0.802
Cox			
	5.69	8.80	0.801

Table II shows how different approaches to handling censored observations affect the predicted event rate, calibration statistic, and C-index of the techniques described in Section 4. Figure 4 displays calibration plots which compare predicted CV risk to empirical (using the Kaplan-Meier estimator) CV risk across bins defined by the predicted risk. From Table II and Figure 4, it is clear that the *-Discard*, *-Zero*, and *-Split* variants of each technique are poorly calibrated *across all methods considered in this analysis*. As expected, the *-Discard* approach consistently over-estimates risk. As noted previously, subjects with short event times are much more likely to have their event status known. For example, a subject who has a CV event one year after the index date must only stay enrolled in the health plan for one year for E to be known; those subjects for whom $E = 0$ must stay enrolled in the insurance plan for five years after baseline for the event status to be known. As expected, the *-Zero* approach underestimates the CV risk, both overall and within subgroups, across all the machine learning techniques considered here. This approach inflates the proportion of subjects not experiencing a CV event as some subjects whose follow-up was censored would have experienced a CV event prior to 5 years.

The effect of the *-Split* technique is more subtle but consistent across the methods considered in this analysis. For subjects in the training set with the event status unknown, the replicate with E set equal to 0 is assigned a weight based on the probability that $E = 0$ given that the subject was known to survive until the censoring time but not conditioned on any of the features. Similarly, the replicate with E set equal to 1 is assigned a weight based on the probability that $E = 1$ given the subject was known to survive until the censoring time. This approach necessarily attenuates the relationship between the features and the event

status. As a result, this technique tends to over-predict the risk for subjects with low risk and under-predict the risk for subjects at high risk which can be seen in Figure 4. Even though this method appears to be more refined than the *-Discard* and *-Zero* variants, the performance is just as poor.

The *-IPCW* versions are generally well-calibrated, with predicted event rate and calibration similar to and in some cases exceeding the Cox model. IPCW machine learning techniques were only versions of the machine learning methods to consistently have acceptable calibration. Discrimination performance is not dramatically affected by the way in which censoring is handled, with small gains (change in C-index of < 0.005) in most techniques due to IPCW as compared to other methods for handling censoring. That is, *ad hoc* methods for handling censoring do not substantially impact the relative ordering of patient's risk. But, simply put, risk predictions which are poorly calibrated are unlikely to be adopted in the clinical setting.

Table III compares the net reclassification improvement for the IPCW versions of various techniques, along with the Cox model. We do not consider cNRIs for the *-Discard*, *-Zero*, and *-Split* variants, as recent papers [Pepe 2011] have shown that the NRI can be a very misleading statistic when comparing poorly calibrated models. In almost all cases, reclassification performance as measured by cNRI is similar across the techniques, which is consistent with the C-index results in Table II.

Table III. Net reclassification (cNRI) comparisons for IPC weighted versions of classification tree, k-nearest neighbors, Bayesian network, logistic regression, generalized additive model and Cox proportional hazards model, evaluated on the hold-out test set. Positive numbers indicate that the bolded technique correctly reclassifies subjects more frequently than the technique preceded by "vs.". *cNRI (Events)* and *cNRI (Non-Events)* give the reclassification improvement among those who did and did not experience events, and *cNRI (Overall)* is their sum. *cNRI (Overall Weighted)* is a weighted sum where the reclassification performance among Events and Non-Events is weighted according to the event and non-event probabilities, respectively.

	cNRI (Events)	cNRI (Non-Events)	cNRI (Overall)	cNRI (Overall Weighted)
Tree				
vs. k-NN	-0.003	0.048	0.045	0.045
vs. Bayes	-0.064	0.058	-0.006	0.050
vs. Logistic	-0.065	0.045	-0.020	0.038
vs. GAM	-0.056	0.030	-0.026	0.024
vs. Cox	-0.102	0.061	-0.041	0.051
k-NN				
vs. Tree	0.003	-0.048	-0.045	-0.045
vs. Bayes	-0.065	0.015	-0.050	0.009
vs. Logistic	-0.108	0.009	-0.099	0.001
vs. GAM	-0.069	-0.013	-0.082	-0.016
vs. Cox	-0.159	0.031	-0.128	0.018
Bayes				
vs. Tree	0.064	-0.058	0.006	-0.050
vs. k-NN	0.065	-0.015	0.050	-0.009
vs. Logistic	-0.013	-0.017	-0.030	-0.017
vs. GAM	0.028	-0.040	-0.012	-0.035
vs. Cox	-0.060	0.002	-0.058	-0.002
Logistic				
vs. Tree	0.065	-0.045	0.020	-0.038
vs. k-NN	0.108	-0.009	0.099	-0.001
vs. Bayes	0.013	0.017	0.030	0.017
vs. GAM	0.037	-0.022	0.015	-0.018
vs. Cox	-0.053	0.021	-0.031	0.017
GAM				
vs. Tree	0.056	-0.030	0.026	-0.024
vs. k-NN	0.069	0.013	0.082	0.016
vs. Bayes	-0.028	0.040	0.012	0.035
vs. Logistic	-0.037	0.022	-0.015	0.018
vs. Cox	-0.085	0.043	-0.043	0.035

We have demonstrated that properly accounting for censoring using IPC weights allows machine learning risk prediction methods to perform as well as methods designed specifically for censored data including the Cox model which, in the context of CV risk prediction, has been shown repeatedly to have excellent calibration and reclassification performance [D'Agostino Sr et al. 2001; Eichler et al. 2007]. In other more complex dataset or complex subgroups in this analysis, we would expect the flexible machine learning methods to outperform the Cox model. For example, among younger patients (under 55 years of age) on blood pressure medication but with controlled blood pressure (SBP ¡ 140 mm Hg), the Cox model greatly under-predicts events (4.27 % predicted event rate, calibration statistic $K = 13.6$) as a model which is additive in the

covariates is not flexible enough to accurately predict the risk in this unique subpopulation. Conversely, the more flexible IPC weighted tree, Bayesian network, generalized additive models provide more accurate risk predictions in this subpopulation (K between 1.5 and 8.6) and slightly better discrimination.

9. DISCUSSION

We have proposed a general-purpose technique for improving the performance of machine learning methods when the binary class indicator is unknown for a subset of individuals due to censoring and have illustrated the approach within a variety of standard machine learning algorithms. Previous methods to handle unknown event statuses due to censoring have largely been developed within the context of a single machine learning technique. Technique-specific approaches to handle censored event times can be suboptimal because frequently it is not known *a priori* which techniques are likely to perform best in a particular application. Inverse probability of censoring weighting provides an encompassing tool which can be straightforwardly applied to any machine learning method.

We demonstrate that a wide variety of general-purpose machine learning techniques, when properly accounting for censoring using IPCW, can be successfully applied to predict time-to-event outcomes and can produce results that are at least comparable to state-of-the-art methods such as Cox that were designed specifically for time-to-event data. Furthermore, though motivated by an example in electronic health data, our technique is generally applicable to any situation where event outcomes are subject to censoring. For example, in economics, one might wish to predict whether recently-unemployed individuals will be re-hired within a fixed time period, an outcome which is likely to be censored in most feasible study designs. In our context, we plan to incorporate this technique into a point-of-care clinical decision support system, which will provide more accurate cardiovascular risk predictions for patients based on their individual health history.

9.1. Limitations

The statistical validity of IPCW rests on several assumptions, in particular that the censoring time is independent of both the event time and patient features. This is a plausible assumption for EHD, where censoring typically occurs for reasons unrelated to a person's health status, but the assumption is much less plausible in other contexts. For example, if data were collected from a small regional hospital, patients with severe health problems might be censored because they went to a larger facility to seek care. In practice, it is unlikely that the independence assumption is satisfied exactly, and it is an open question how the degree to and manner in which the assumption is violated affects the performance of IPCW techniques.

The benefits of IPCW also depend on the proportion of subjects who are censored and hence have an undetermined event status. In our data, approximately half of subjects were censored before 5 years, a level of censoring which is amenable to IPCW. When few subjects (e.g., $< 10\%$) are censored, the gains of IPCW over "naive" techniques will be modest. When the vast majority (e.g., $> 90\%$) are censored, only a small fraction of subjects will be used in the analysis, with many having large weights leading to variable and unstable predictions.

9.2. Extensions and future work

We have provided several examples of how to incorporate IPCW into established machine learning techniques. There are, of course, a wide variety of machine learning algorithms which we did not implement, but the simplicity of the IPCW approach means that it can be adapted to a wide range of existing tools. Indeed, due to IPCW's ease of implementation and use, it would be possible to develop ensemble-based risk prediction tools to apply to censored data. For instance, given an implementation of IPCW decision trees, constructing an IPCW random forest is straightforward.

Most of the analyses reported in this paper were performed using implementations of these techniques in standard statistical software, but support for a "weights" argument is not universal across all machine learning packages. Even implementations which do allow for weights to be specified may not use them consistently, e.g., they will be used for training the model but not for tuning parameter selection. In ongoing work, we are developing more general resampling-based approaches which will allow IPCW analyses to be performed using machine learning software which lacks the capacity to handle user-specified weights.

REFERENCES

Firas Abdollah, R. Jeffrey Karnes, Nazareno Suardi, Cesare Cozzarini, Giorgio Gandaglia, Nicola Fossati, Damiano Vizziello, Maxine Sun, Pierre I. Karakiewicz, Mani Menon, Francesco Montorsi, and Alberto

Briganti. 2014. Impact of adjuvant radiotherapy on survival of patients with node-positive prostate cancer. *Journal of Clinical Oncology* (2014), in press.

U. Rajendra Acharya, S. Vinitha Sree, M. Muthu Rama Krishnan, Filippo Molinari, Luca Saba, Sin Yee Stella Ho, Anil T. Ahuja, Suzanne C. Ho, Andrew Nicolaides, and Jasjit S. Suri. 2012. Atherosclerotic risk stratification strategy for carotid arteries using texture-based features. *Ultrasound in Medicine & Biology* 38, 6 (2012), 899–915.

Alan Agresti. 2012. *Categorical Data Analysis* (third ed.). Wiley-Interscience.

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (Dec. 1974), 716–723.

Leila Amini, Reza Azarpazhouh, Mohammad Taghi Farzadfar, Sayed Ali Mousavi, Farahnaz Jazaieri, Fariborz Khorvash, Rasul Norouzi, and Nafiseh Toghianfar. 2013. Prediction and control of stroke by data mining. *International Journal of Preventive Medicine* 4, Suppl 2 (2013), S245.

Steen Andreassen, Christian Riekehr, Brian Kristensen, Henrik C Schønheyder, and Leonard Leibovici. 1999. Using probabilistic and decision-theoretic methods in treatment and prognosis modeling. *Artifical Intelligence in Medicine* 15, 2 (Feb. 1999), 121–134.

Muhammad Arif, Ijaz A. Malagore, and Fayyaz A. Afsar. 2012. Detection and localization of myocardial infarction using K-nearest neighbor classifier. *Journal of Medical Systems* 36, 1 (2012), 279–289.

Gerd Assmann, Paul Cullen, and Helmut Schulte. 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation* 105, 7 (19 Feb. 2002), 900–900.

Peter C. Austin, Jack V. Tu, Jennifer E. Ho, Daniel Levy, and Douglas S. Lee. 2013. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology* 66, 4 (April 2013), 398–407.

Sunayan Bandyopadhyay, Julian Wolfson, David M Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Mohamed Elidrisi, Paul E Johnson, and Patrick J O'Connor. 2014. Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery* (2014).

Heejung Bang and Anastasios A. Tsiatis. 2000. Estimating medical costs with censored data. *Biometrika* 87, 2 (June 2000), 329–343.

Heejung Bang and Anastasios A. Tsiatis. 2002. Median regression with censored cost data. *Biometrics* 58, 3 (Sep. 2002), 643–649.

Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. 1998. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine* 17, 10 (May 1998), 1169–1186.

Rosa Blanco, Iñaki Inza, Marisa Merino, Jorge Quiroga, and Pedro Larrañaga. 2005. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics* 38, 5 (Oct. 2005), 376–388.

Dennis D. Boos and L. A. Stefanski. 2013. *Essential Statistical Inference: Theory and Methods*. Springer, New York.

Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. CRC Press.

Jonathan Buckley and Ian James. 1979. Linear-Regression with Censored Data. *Biometrika* 66, 3 (Dec. 1979).

Qiongyu Chen, Guoliang Li, Tze-Yun Leong, and others. 2007. Predicting coronary artery disease with medical profile and gene polymorphisms data. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, 1219.

Feixiong Cheng and Zhongming Zhao. 2014. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association* (2014).

Gary S. Collins and Douglas G. Altman. 2009. An independent external validation and evaluation of QRISK cardiovascular risk prediction: A prospective open cohort study. *British Medical Journal* 339 (July 2009), b2584.

Isabelle Colombet, Alan Ruelland, Gilles Chatellier, François Gueyffier, P. Degoulet, and M. C. Jaulent. 2000. Models to predict cardiovascular risk: Comparison of CART, multilayer perceptron and logistic regression. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 156–160.

R. M. Conroy, K. Pyorala, A. P. Fitzgerald, S. Sans, A. Menotti, G. DeBacker, D. DeBacquer, P. Ducimetiere, P. Jousilahti, U. Keil, I. Njolstad, R. G. Oganov, T. Thomsen, H. Tunstall-Pedoe, A. Tverdal, H. Wedel, P. Whincup, L. Wilhelmsen, and I. M. Graham. 2003. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *European Heart Journal* 24, 11 (2003), 987–1003.

Marie Therese Cooney, Alexandra Dudina, Ralph D'Agostino, and Ian M. Graham. 2010. Cardiovascular Risk-Estimation Systems in Primary Prevention: Do They Differ? Do They Make a Difference? Can We See the Future? *Circulation* 122, 3 (2010), 300–310.

Marie Therese Cooney, Alexandra L. Dudina, and Ian M. Graham. 2009. Value and limitations of existing scores for the assessment of cardiovascular risk: A review for clinicians. *Journal of the American College Cardiology* 54, 14 (2009), 1209–1227.

D. R. Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society Series B* 34, 2 (1972), 187–220.

Ralph B. D'Agostino, Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. 2008. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation* 118, 4 (2008), E86–E86.

Ralph B D'Agostino Sr, Scott Grundy, Lisa M Sullivan, Peter Wilson, and others. 2001. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 286, 2 (2001), 180–187.

Pedro Domingos and M Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29 (1997), 103–130. http://link.springer.com/article/10.1023/A:1007413511361

Klaus Eichler, Milo A Puhan, Johann Steurer, and Lucas M Bachmann. 2007. Prediction of first coronary events with the Framingham score: A systematic review. *American Heart Journal* 153, 5 (2007), 722–731.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* 28, 2 (2000), 337–407.

Yair Goldberg and Michael R. Kosorok. 2012. Support vector regression for right censored data. *arXiv preprint arXiv:1202.5130* (2012).

Major Greenwood. 1926. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects. Ministry of Health* 33 (1926).

Fikret Gürgen and Nurgül Gürgen. 2003. Intelligent data analysis to interpret major risk factors for diabetic patients with and without ischemic stroke in a small population. *Biomedical Engineering Online* 2, 1 (2003), 5.

Frank E. Harrell. 2001. *Regression Modeling Strategies.* Springer-Verlag, New York.

Mark Hartney, Yazhuo Liu, Vic Velanovich, Peter Fabri, Jorge Marcet, Michael Grieco, Shuai Huang, and Jose Zayas-Castro. 2014. Bounceback branchpoints: Using conditional inference trees to analyze readmissions. *Surgery* 156, 4 (2014), 842–848.

Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, and P. Brindle. 2008. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: A validation study. *Heart* 94, 1 (2008), 34–39.

Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *British Medical Journal* 335, 7611 (2007), 136–141.

Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Rubin Minhas, Aziz Sheikh, and Peter Brindle. 2008. Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *British Medical Journa;* 336, 7659 (2008), 1475–1489.

David W Hosmer and Stanley Lemesbow. 1980. Goodness of Fit Tests for the Multiple Logistic Regression-Model. *Communications in Statistics-Theory and Methods* 9, 10 (1980), 1043–1069.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. 2004. Bagging survival trees. *Statistics in Medicine* 23, 1 (2004), 77–91.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. Random survival forests. *The Annals of Applied Statistics* (2008), 841–860.

George H John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence.*

John D. Kalbfleisch and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data.* Wiley, Hoboken, NJ.

Michael W. Kattan. 2003. Comparison of Cox regression with other methods for determining prediction models and nomograms. *The Journal of Urology* 170, 6 (2003), S6–S9.

Michael W Kattan, Kenneth R Hess, and J Robert Beck. 1998. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. *Computers and Biomedical Research* 31, 5 (1998), 363–373.

Emily Kawaler, Alexander Cobian, Peggy Peissig, Deanna Cross, Steve Yale, and Mark Craven. 2012. Learning to predict post-hospitalization VTE risk from EHR data. In *AMIA Annual Symposium Proceedings*, Vol. 2012. American Medical Informatics Association, 436.

Joanna Kazmierska and Julian Malicki. 2008. Application of the Naive Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology* 86, 2 (Feb. 2008), 211–216.

Edward H. Kennedy, Wyndy L. Wiitala, Rodney A. Hayward, and Jeremy B. Sussman. 2013. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical Care* 51, 3 (2013), 251–258.

Faisal M. Khan and Valentina Bayer Zubek. 2008. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In *Eighth IEEE International Conference on Data Mining (ICDM 2008)*. IEEE, 863–868.

Max Kuhn and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer.

Martijn Lappenschaar, Arjen Hommersom, Peter J. F. Lucas, Joep Lagro, and Stefan Visscher. 2013. Multilevel Bayesian networks for the analysis of hierarchical health care data. *Artificial Intelligence in Medicine* 57, 3 (Mar. 2013), 171–183.

Pedro Larranaga, Basilio Sierra, Miren J. Gallego, Maria J. Michelena, and Juan M. Picaza. 1997. Learning Bayesian networks by genetic algorithms: A case study in the prediction of survival in malignant skin melanoma. *Artifical Intelligence in Medicine* 1211 (1997), 261–272.

E.L. Lehmann and George Casella. 1998. *Theory of Point Estimation*. Vol. 31. Springer.

Stanley Lemeshow and David W. Hosmer. 1982. A review of goodness of fit statistics for use in the development of logistic-regression models. *American Journal of Epidemiology* 115, 1 (1982), 92–106.

Yu-Kai Lin, Hsinchun Chen, Randall A Brown, Shu-Hsing Li, and Hung-Jen Yang. 2014. Predictive Analytics for Chronic Care: A Time-to-Event Modeling Framework Using Electronic Health Records. *Available at SSRN 2444025* (2014).

Ari M. Lipsky and Roger J. Lewis. 2005. Placing the Bayesian network approach to patient diagnosis in perspective. *Annals of Emergency Medicine* 45, 3 (MAR 2005), 291–294.

Dan Liu, Daniel Y.T. Fong, Albert C.Y. Chan, Ronnie T.P. Poon, and Pek-Lan Khong. 2014. Hepatocellular carcinoma: Surveillance CT schedule after hepatectomy based on risk stratification. *Radiology* (2014).

Wei-Yin Loh. 2008. Classification and regression tree methods. *Encyclopedia of Statistics in Quality and Reliability* (2008).

Peter Lucas, Henk Boot, and Babs Taal. 1998. Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine* 37, 3 (SEP 1998), 206–219.

Peter J. F. Lucas, Nicolette C. de Bruijn, Karin Schurink, and Andy Hoepelman. 2000. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in Medicine* 19, 3 (JUL 2000), 251–279.

Peter J. F. Lucas, Linda C. van der Gaag, and Ameen Abu-Hanna. 2004. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine* 30, 3 (MAR 2004), 201–214.

Thomas Lumley. 2004. Analysis of complex survey samples. *Journal of Statistical Software* 9, 1 (2004), 1–19.

Yohann Mansiaux and Fabrice Carrat. 2014. Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. *BMC Medical Research Methodology* 14, 1 (2014), 99.

Judith C. Maro, Richard Platt, John H. Holmes, Brian L. Strom, Sean Hennessy, Ross Lazarus, and Jeffrey S. Brown. 2009. Design of a national distributed health data network. *Annals of Internal Medicine* 151, 5 (2009), 341–344.

Michael Matheny, Melissa L. McPheeters, Allison Glasser, Nate Mercaldo, Rachel B. Weaver, Rebecca N. Jerome, Rachel Walden, J Nikki McKoy, Jason Pritchett, and Chris Tsai. 2011. *Systematic review of cardiovascular disease risk assessment tools*. Technical Report. Agency for Healthcare Research and Quality (US).

Jose A. Muñoz-Moreno, Núria Pérez-Álvarez, Amalia Muñoz-Murillo, Anna Prats, Maite Garolera, M. Àngels Jurado, Carmina R. Fumaz, Eugènia Negredo, Maria J. Ferrer, and Bonaventura Clotet. 2014. Classification models for neurocognitive impairment in HIV infection based on demographic and clinical variables.

PloS One 9, 9 (2014), e107625.

RM Parry, W Jones, TH Stokes, JH Phan, RA Moffitt, H Fang, L Shi, A Oberthuer, M Fischer, W Tong, and others. 2010. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal* 10, 4 (2010), 292–309.

Michael J. Pencina, Ralph B. D'Agostino, and Ewout W. Steyerberg. 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* 30, 1 (2011), 11–21.

Michael J. Pencina, Ralph B. D'Agostino Sr, Ralph B. D'Agostino Jr, and Ramachandran S. Vasan. 2008. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 27, 2 (2008), 157–172.

Margaret S. Pepe. 2011. Problems with risk reclassification methods for evaluating prediction models. *American Journal Epidemiology* 173, 11 (2011), 1327–35.

Richard Platt, Robert Davis, Jonathan Finkelstein, Alan S. Go, Jerry H. Gurwitz, Douglas Roblin, Stephen Soumerai, Dennis Ross-Degnan, Susan Andrade, Michael J. Goodman, and others. 2001. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiology and Drug Safety* 10, 5 (2001), 373–377.

J. Ross Quinlan. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco, CA, USA.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Paul M. Ridker, Nader Buring, Julie E.and Rifai, and Nancy R Cook. 2007. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score. *JAMA: Journal of the American Medical Association* 297, 6 (2007), 611–619.

Paul M. Ridker, Nina P. Paynter, Nader Rifai, J. Michael Gaziano, and Nancy R. Cook. 2008. C-Reactive protein and parental history improve global cardiovascular risk prediction: The Reynolds risk score for men. *Circulation* 118, 18 (2008), S1145–S1145.

Brian D. Ripley and Ruth M. Ripley. 2001. Neural networks as statistical methods in survival analysis. *Clinical Applications of Artificial Neural Networks* (2001), 237–255.

James M. Robins and Dianne M. Finkelstein. 2000. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56, 3 (2000), 779–788.

Andrea G. Rotnitzky and James M. Robins. 2004. Inverse probability weighted estimation in survival analysis. In *The Encyclopedia of Biostatistics* (second ed.), Peter Armitage and Theodore Colton (Eds.). John Wiley & Sons.

Stuart Russell and Peter Norvig. 2003. *Artificial intelligence: A modern approach*. Vol. 2. Prentice Hall, Upper Saddle River, New Jersey. 495 pages.

Shantanu Sarkar and Jodi Koehler. 2013. A dynamic risk score to identify increased risk for heart failure decompensation. *Biomedical Engineering, IEEE Transactions on* 60, 1 (JAN 2013), 147–150.

Mark Robert Segal. 1988. Regression trees for censored data. *Biometrics* (1988), 35–47.

Joe V. Selby. 1997. Linking automated databases for research in managed care settings. *Annals of Internal Medicine* 127, 8_Part_2 (1997), 719–724.

M. Berkan Sesen, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. 2013. Bayesian networks for clinical decision support in lung cancer care. *PLoS One* 8, 12 (DEC 2013), e82349.

Jooyong Shim and Changha Hwang. 2009. Support vector censored quantile regression under random censoring. *Computational Statistics & Data Analysis* 53, 4 (2009), 912–919.

Pannagadatta K. Shivaswamy, Wei Chu, and Martin Jansche. 2007. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007.* IEEE, 655–660.

Basilio Sierra and Pedro Larranaga. 1998. Predicting survival in malignant skill melanoma using Bayesian networks automatically induced by genetic algorithms: An empirical comparison between different approaches. *Artificial Intelligence in Medicine* 14, 1-2 (1998), 215–230.

Wade P Smith, Jason Doctor, Jürgen Meyer, Ira J. Kalet, and Mark H. Phillips. 2009. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model. *Artificial Intelligence in Medicine* 46, 2 (JUN 2009), 119–130.

Xiaowei Song, Arnold Mitnitski, Jafna Cox, and Kenneth Rockwood. 2004. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Medinfo* 11, 1 (2004), 736–40.

Ivan Štajduhar and Bojana Dalbelo-Bašić. 2010. Learning Bayesian networks from survival data using weighting censored instances. *Journal of Biomedical Informatics* 43, 4 (AUG 2010), 613–622.

Ivan Štajduhar and Bojana Dalbelo-Bašić. 2012. Uncensoring censored data for machine learning: A likelihood-based approach. *Expert Systems with Applications* 39, 8 (JUN 15 2012), 7226–7234.

Ivan Štajduhar, Bojana Dalbelo-Bašić, and Nikola Bogunović. 2009. Impact of censoring on learning Bayesian networks in survival modelling. *Artificial Intelligence in Medicine* 47, 3 (2009), 199–217.

Walter F. Stewart, Jason Roy, Jimeng Sun, and Shahram Ebadollahi. 2014. Clinical utility of machine learning and longitudinal EHR data. In *Machine Learning in Healthcare Informatics*. Springer, 209–227.

Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Edabollahi, Steven E Steinhubl, Zahra Daar, and Walter F Stewart. 2012. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings*, Vol. 2012. American Medical Informatics Association, 901.

C.E. Sundberg, T. Aulin, and N. Rydbeck. 1972. The rate of convergence of k-NN regression estimates and classification rules. *IEEE Transactions on Information Theory* 20 (1972), 429–435.

Terry M. Therneau, Patricia M. Grambsch, and Thomas R. Fleming. 1990. Martingale-based residuals for survival models. *Biometrika* 77, 1 (1990), 147–160.

Anastasios A. Tsiatis. 2006. *Semiparametric Theory and Missing Data*. Springer, New York.

Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan A. K. Suykens. 2011. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine* 53, 2 (2011), 107–118.

Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45, 3 (2011), 1–67. http://www.jstatsoft.org/v45/i03/

Marina Velikova, Josien Terwisscha van Scheltinga, Peter JF Lucas, and Marc Spaanderman. 2014. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning* 55, 1 (JAN 2014), 59–73.

Marion Verduijn, Niels Peek, Peter M. J. Rosseel, Evert de Jonge, and Bas A. J. M. de Mol. 2007. Prognostic Bayesian networks I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics* 40, 6 (DEC 2007), 609–618.

Joan Vila-Francés, Juan Sanchis, Emilio Soria-Olivas, Antonio José Serrano, Marcelino Martínez-Sober, Clara Bonanad, and Silvia Ventura. 2013. Expert system for predicting unstable angina based on Bayesian networks. *Expert Systems With Applications* 40, 12 (SEP 15 2013), 5004–5010.

Xiang Wang, Fei Wang, Jun Wang, Buyue Qian, and Jianying Hu. 2013. Exploring patient risk groups with incomplete knowledge. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1223–1228.

Simon Wood. 2006. *Generalized Additive Models: An Introduction with R*. CRC Press.

Mark Woodward, Peter Brindle, and Hugh Tunstall-Pedoe. 2007. Adding social deprivation and family history to cardiovascular risk assessment: The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 93, 2 (2007), 172–176.

Jionglin Wu, Jason Roy, and Walter F Stewart. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care* 48, 6 (2010), S106–S113.

Barbaros Yet, Kaveh Bastani, Hendry Raharjo, Svante Lifvergren, William Marsh, and Bo Bergman. 2013. Decision support system for Warfarin therapy management using Bayesian networks. *Decision Support Systems* 55, 2 (MAY 2013), 488–498.

Ruoqing Zhu and Michael R Kosorok. 2012. Recursively imputed survival trees. *J. Amer. Statist. Assoc.* 107, 497 (2012), 331–340.

Blaž Zupan, Janez DemšAr, Michael W Kattan, J Robert Beck, and Ivan Bratko. 1999. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine* 1620 (1999), 346–355.

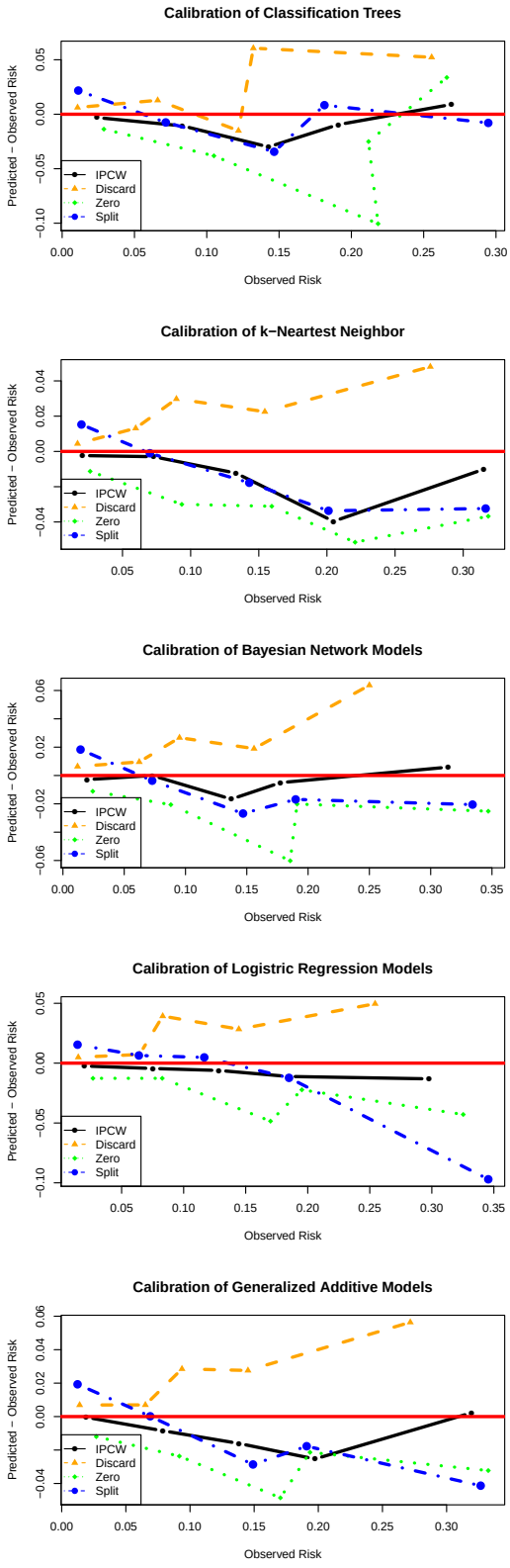

Fig. 4. Predicted CV risk minus empirical or observed CV risk across bins defined by the predicted risk. The predicted risk bins were based on clinically relevant cutoffs for the risk of experiencing a cardiovascular event within 5 years: 0-5%, 5-10%, 10-15%, 15-20% and $>$ 20%.