HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2016 September 20.

Published in final edited form as:

Stat Med. 2015 September 20; 34(21): 2941–2957. doi:10.1002/sim.6526.

A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data

J. Wolfson[*], **S. Bandyopadhyay**[†], **M. Elidrisi**[†], **G. Vazquez-Benitez**[§], **D. M. Vock**[*], **D. Musgrove**[*], **G. Adomavicius**[‡], **P.E. Johnson**[‡], and **P.J. O'Connor**[§]

[*]Division of Biostatistics, University of Minnesota, Minneapolis, MN

[†]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN

[‡]Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN

[§]HealthPartners Institute for Education and Research, Minneapolis, MN

Abstract

Predicting an individual's risk of experiencing a future clinical outcome is a statistical task with important consequences for both practicing clinicians and public health experts. Modern observational databases such as electronic health records (EHRs) provide an alternative to the longitudinal cohort studies traditionally used to construct risk models, bringing with them both opportunities and challenges. Large sample sizes and detailed covariate histories enable the use of sophisticated machine learning techniques to uncover complex associations and interactions, but observational databases are often "messy," with high levels of missing data and incomplete patient follow-up. In this paper, we propose an adaptation of the well-known Naive Bayes (NB) machine learning approach to time-to-event outcomes subject to censoring. We compare the predictive performance of our method to the Cox proportional hazards model which is commonly used for risk prediction in healthcare populations, and illustrate its application to prediction of cardiovascular risk using an EHR dataset from a large Midwest integrated healthcare system.

Keywords

Naive Bayes; machine learning; risk prediction; survival analysis; electronic health records

1. Introduction

The task of predicting the occurrence of future clinical events is central to many areas of medical practice. "Risk calculators", which make personalized risk predictions based on individual characteristics, are now commonly employed for outcomes such as heart attack, stroke, and diabetes. A common feature of many of these risk calculators is that their predictions are based on data from cohort studies. For example, the Framingham Risk Score (FRS) [1, 2] for predicting the occurrence of cardiovascular disease (CVD) is based on data

[*]Correspondence to: Division of Biostatistics, University of Minnesota School of Public Health, A460 Mayo Memorial Building MMC 303, 420 Delaware St SE, Minneapolis, MN 55455.

from the Framingham Heart Study [3], a cohort study which has been ongoing since 1948. Many calculators, including the FRS, focus on predicting five- and ten-year cardiovascular (CV) risk, i.e., the probability of having a first CVD-defining event (coronary heart disease, stroke, or other CVD-related death) in the next five or ten years. Unfortunately, the cost of subject acquisition and long-term follow-up restricts the size of these studies; for example, total enrollment in the Framingham Heart Study across three generational cohorts is slightly less than 15,000 [4], and a recent study of a subset of 8,491 FHS participants [5] reported 1,174 incident CVD events over 12 years of follow-up. As a result, longitudinal cohort studies may be limited in their ability to describe unique features of contemporary patient populations, and in their capacity to predict risk over shorter time horizons (eg. 1 year, 5 years) when the event rate is low.

An alternative data source for risk prediction is electronic health data (EHD), which are increasingly available on entire populations under care within health insurance systems. Covering a fixed time period, typically several years, they often contain hundreds of thousands of records on enrolled subjects from multiple sources including electronic medical records, claims and prescription data, and state and national databases. Since the data are recorded longitudinally, they are well-suited to obtaining risk estimates, but their complexity and dimensionality challenge standard statistical methods. In particular, follow-up times are often subject to moderate to heavy censoring, as subjects may enroll in and/or disenroll from the health insurance plan during the data capture period.

Risk engines for a variety of clinical outcomes have traditionally been developed by fitting survival regression models such as the Cox proportional hazards model [6] or the accelerated failure time model [7] to censored time-to-event data from the aforementioned longitudinal studies. Because regression-based models yield estimates of the effects of each covariate in the model, they are useful not only as a predictive tool, but also for understanding the relative influence of covariates on risk. However, the drawback of fitting simple regression models is that non-linear covariate effects and interactions may not be fully captured unless they are pre-specified during the modeling process.

If predictive performance and not model interpretability is the central concern, then more flexible machine learning techniques may yield improved risk engines by capturing complex relationships and between covariates and risk [8, 9, 10, 11]. Fully supervised machine learning methods typically assume that the event indicator is known for all subjects, while in our setting the event indicator is undetermined for subjects who are censored, i.e., who do not experience an event and are not followed for the full time period over which one wants to make predictions. Simple approaches to dealing with this issue, such as discarding censored observations (see, e.g., [12, 13, 14]) or treating them as zeroes (non-events), are known to induce bias in the estimation of class probabilities [15], making typical fully supervised classification employing these approaches unsuitable; for example, Stajduhar et al. [16] 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 (non-censored) and unlabeled (censored) 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. Ishwaran et al. [17] describe random survival forests, and Lucas et al. [18] discuss the application of Bayesian networks to right-censored data. Zupan et al. [19] and Stajduhar and Dalbelo-Basic [20] propose approaches in which censored observations are repeated twice in the dataset, one 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 and the time the event status will be assessed. Stajduhar and Dalbelo-Basic [21] adopt 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.

In this paper, we present Censored Naive Bayes (CNB), a machine learning approach to predicting risk using time-to-event data subject to censoring. CNB is an extension of Naive Bayes (NB), a popular technique typically applied to classification problems with categorical outcomes [22, 23, 24, 25, 26]. CNB is nonparametric with respect to the underlying distribution of event times, and models the marginal covariate distributions in a flexible manner. However, while "standard" CNB performs well in terms of risk reclassification, it is poorly calibrated when covariates are not mutually independent. We therefore also introduce a modification of CNB, CNB-PC, which involves orthogonalizing the design matrix prior to analysis via a Singular Value Decomposition (SVD), resulting in a technique with much better calibration. We compare our proposed CNB techniques to a Cox proportional hazards (CPH) based approach in a variety of simulated scenarios, and illustrate the application of our methods to cardiovascular risk prediction using longitudinal electronic health record data from a large Midwest integrated healthcare system.

2. Naive Bayes for binary outcomes

Consider the task of estimating the probability of occurrence of an event E over a fixed time period $[0, \tau]$, based on individual characteristics $\boldsymbol{X} = (X_1, \ldots, X_p)$ which are measured at some well-defined baseline time $t = 0$. For the moment, we will assume that we have data on n subjects who have had $\boldsymbol{X}$ measured at $t = 0$ and been followed for τ time units to observe whether or not E occurred ($E = 1 \Rightarrow$ event occurrence). The target of estimation is $P(E = 1 \mid \boldsymbol{X})$.

Regression models for binary outcomes (e.g., logistic regression) require that the analyst specify the way in which $\boldsymbol{X}$ relates to the outcome by choosing appropriate main effects and interaction terms. An alternative is to proceed indirectly by using Bayes Rule, writing

$$P(E=1|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|E=1)P(E=1)}{P(\boldsymbol{X})} = \frac{P(\boldsymbol{X}|E=1)P(E=1)}{\sum_{e=0,1} P(\boldsymbol{X}|E=e)P(E=e)}. \quad (1)$$

In traditional classification problems, where ordering subjects with respect to the probability of the outcome is more important than estimating that probability precisely, the denominator can be ignored so that the focus is on estimation of the joint distribution of $\boldsymbol{X}|E = 1$ ($P(E = 1)$ is easily estimated using the sample proportion of subjects experiencing events). The

denominator must be retained to achieve properly calibrated predictions, and is typically computed via the law of total probability.

It may be difficult to model the joint distributions $P(\boldsymbol{X}|E = e)$, particularly if p is large. Researchers have therefore proposed to simplify estimation by assuming parsimonious dependence structures for $\boldsymbol{X} \mid E = e$. The *Bayesian network* approach [27] consists of partitioning $\boldsymbol{X}$ into low-dimensional, mutually independent subsets on the basis of a directed acyclic graph (DAG) encoding conditional independence relationships between covariates. The DAG structure may be pre-specified (e.g., based on the input of domain experts) or constructed from the available data.

The most extreme version of this approach, termed *Naive Bayes*, makes the dramatic simplifying assumption that the covariates $X_1, X_2, \ldots, X_p$ constituting $\boldsymbol{X}$ are independent, given E, so that (1) can be rewritten as

$$P(E=1|\boldsymbol{X})=\frac{\prod_{j=1}^{p}P(X_j|E=1)P(E=1)}{\sum_{e=0,1}\prod_{j=1}^{p}P(X_j|E=e)P(E=e)} \quad (2)$$

A number of parametric and semi-parametric approaches to modeling the univariate covariate distributions are possible and have been described elsewhere [23, 28]; one common assumption in classification problems is that the individual covariate densities in the denominator of (2) can be represented using mixtures of Normal distributions. Estimation proceeds by computing the maximum likelihood parameter estimates of each covariate distribution (conditional on $E = 0, 1$), and plugging the resulting estimates into (2). $P(E = e)$ is estimated as the sample proportion of subjects with $E = e$. NB easily accommodates problems with large numbers of covariates, and can be computed quickly for large samples. The independence assumption also makes NB trivial to parallelize, allowing very fast implementations to be built for multiprocessor systems.

Even though the Naive Bayes assumption is often incorrect, NB has been found to perform quite well in a variety of classification tasks, including when covariates are highly correlated [22, 23, 29, 30, 31]. However, NB is also known to be poorly calibrated in some settings, i.e., conditional probability estimates are biased though they are highly correlated with the true probabilities so that classification performance remains acceptable [32]. In the section below, we describe a modification to NB using the singular value decomposition (SVD) which results in improved calibration when the goal is to obtain more accurate estimates of the conditional survival probability.

3. Censored Naive Bayes

In many studies, some participants will have incomplete data because of loss-to-follow-up or censoring. The dataset motivating this work consists of electronic health data captured between 2002 and 2012; over this time period, some patients who contributed data in 2002 disenrolled prior to 2012, and other patients enrolled and began contributing data after 2002. If E denotes the indicator that a subject experiences a cardiovascular event within 5 years of his/her “index time” (a term we define more formally in the next section), then subjects who

were followed for less than five years and who did not experience an event have E undefined. Ignoring or assigning a particular event status to these subjects may lead to biased probability estimates since their underlying failure times are right-truncated. We propose a more principled approach to handling these observations which we refer to as *Censored Naive Bayes* (CNB).

3.1. Notation

Consider a longitudinal study of τ years duration, from which data on n independent subjects are available. Subjects may enter the study at any time between the beginning of the study and year τ, and follow-up can end when an event occurs, when the study ends in year τ, or when a subject drops out of the study prior to year τ. For each subject i, define T_i as the time between some index time $t_i = 0$ and the occurrence of an event (failure time), and C_i as the time between $t_i = 0$ and the end of follow-up due to study termination or drop-out (censoring time). In our data application example, the index time $t_i = 0$ corresponds to the first clinic visit following the initial 12 months of continuous enrollment in the healthcare system. For all subjects $i = 1, \dots, n$, we observe data $(O_i, \delta_i, \boldsymbol{X}_i)$, where $O_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}[T_i \le C_i]$, and $\boldsymbol{X}_i = \{X_{i1}, \dots, X_{ip}\}$ is a p-vector of continuous-valued covariates available at $t_i = 0$. Let $Y_i(t) = \mathbb{1}[O_i \ge t]$ be the at-risk indicator at time t_k. We denote the marginal survivor function by $S(t) \equiv P(T \ge t)$, and the conditional survivor function by $S(\boldsymbol{x}, t) \equiv P(T \ge t \mid \boldsymbol{X} = \boldsymbol{x})$. We will use $f(A|B)$ as shorthand for the conditional pmf or density of A, given B, and write $f_j^{\geq}(x,t) = f(X_j | T \ge t)$ and $f_j^{<}(x,t) = f(X_j | T < t)$ where it is understood that, for fixed t, the densities are a function of x.

3.2. Estimating the conditional survivor function

By analogy with (2), we can write

$$S(\boldsymbol{x},t) = S(t)\left[\frac{\prod_{j=1}^{p} f_j^{\geq}(x;t)}{\prod_{j=1}^{p} f_j^{\geq}(x;t)S(t) + \prod_{j=1}^{p} f_j^{<}(x;t)(1-S(t))}\right] \equiv S(t)\Psi(\boldsymbol{x},t). \quad (3)$$

Our Censored Naive Bayes technique involves estimating the two components $S(t)$ and $\Psi(\boldsymbol{X}, t)$ separately and then combining them to yield risk estimates via

$$\hat{S}(\boldsymbol{x},t) = \hat{S}(t) \cdot \hat{\Psi}(\boldsymbol{x},t) \quad (4)$$

The following sections describe approaches to estimating the components $S(t)$ and $\Psi(\boldsymbol{X}, t)$.

3.3. Estimation of S(t)

We propose to estimate $S(t)$ non-parametrically via the Kaplan-Meier estimator [33]. Let the observed event times be $t_1 < t_2 < \cdots < t_K$. For $t_k < t \le t_{k+1}$, we estimate

$$\hat{S}_{KM}(t)=\prod_{m=1}^{k}\left(1-\frac{e_m}{r_m}\right)$$

where e_m and r_m respectively give the number of events and individuals at risk at time t_m,

$$e_m=\sum_{i=1}^{n}\delta_i \mathbb{1}\left[O_i=t_m\right], \quad r_m=\sum_{i=1}^{n}Y_i(t_m).$$

Since the Kaplan-Meier estimator makes no assumptions about the form of $S(t)$, it is ideally suited to a machine learning situation where little information may be available to guide parametric assumptions about the failure time distribution. In situations where the distribution of failure times are thought to follow a known distribution, a parametric approach to estimating the marginal survivor function may be chosen instead; however, in typical applications the large sample size will minimize the importance of any resulting efficiency gains. Consistent estimation of $S(t)$ via Kaplan-Meier requires that the censoring and event times be independent, which is implied by the assumptions in Theorem 1 below.

3.4. Estimation of Ψ(X, t)

If there are p total covariates, then the Naive Bayes assumption requires estimation of the densities $f_j^{\geq}(x,t)$ and $f_j^{<}(x,t)$ for $j = 1, \ldots, p$. Our method assumes that $X_j|T \geq t$ and $X_j|T < t$ follow Gaussian distributions with means and variances depending on t, i.e.,

$$f_j^{\geq}(x;t)=\phi\left(\frac{x-\mu_j(t)}{\sigma_j(t)}\right) \quad \text{and} \quad f_j^{<}(x;t)=\phi\left(\frac{x-\theta_j(t)}{\psi_j(t)}\right) \quad (5)$$

where φ is the standard Normal PDF. The assumption that changes in the covariate distributions with t are restricted to location-scale shifts within the Normal family is a strong one, but our simulations demonstrate that it need not hold exactly for NB to perform well. $\mu_j(t)$ and $\sigma_j^2(t)$ are estimated via:

$$\hat{\mu}_j(t)=\frac{\sum_{i=1}^{n}Y_i(t)X_{ij}}{\sum_{i=1}^{n}Y_i(t)} \text{ if} \sum_{i=1}^{n}Y_i(t) \neq 0, \text{ otherwise } \hat{\mu}_j(t)=0, \quad (6)$$

$$\hat{\sigma}_j^2(t)=\frac{\sum_{i=1}^{n}Y_i(t)X_{ij}^2}{\sum_{i=1}^{n}Y_i(t)}-\hat{\mu}_j^2(t) \text{ if} \sum_{i=1}^{n}Y_i(t) \neq 0, \text{ otherwise } \hat{\sigma}_j^2(t)=1, \quad (7)$$

All individuals with observation times $\geq t$ contribute to estimation of μ and σ^2, since $O_i \equiv \min(T_i, C_i) \geq t$ implies that $T_i \geq t$.

θ_j and ψ^2 are estimated using inverse probability of censoring weighting (IPCW; [34]). Let $G(t) = P(C_i > t)$, the probability that the censoring time is greater than t. We define inverse probability of censoring weights via

$$\omega_i = \begin{cases} \frac{1}{G(\min(T_i,t))} & \text{if } \min(T_i,t) < C_i \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The weights are computed by first estimating G via Kaplan-Meier, using the censoring indicators $\delta_i^* = 1-\delta_i$. Subjects censored prior to t have $\hat{\omega_i} = 0$, otherwise $\hat{\omega_i} = 1/\hat{G}(\min(O_i, t))$. θ and ψ are estimated using the weights $\hat{\omega_i}$ as:

$$\hat{\theta}_j(t) = \frac{\sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t)) X_{ij}}{\sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t))} \text{ if } \sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t)) \neq 0, \text{ otherwise } \hat{\theta}_j(t) = 0, \quad (9)$$

$$\hat{\psi}_j^2(t) = \frac{\sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t)) X_{ij}^2}{\sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t))} - \hat{\theta}_j^2(t) \text{ if } \sum_{i=1}^n \hat{\omega}_i \delta_i (1-Y_i(t)) \neq 0, \text{ otherwise } \hat{\psi}_j^2(t) = 1. \quad (10)$$

The densities in (5) are evaluated by plugging in the resulting parameter estimates.

In Appendix A, we provide a proof of the following theorem:

Theorem 1—Suppose that the following assumptions hold:

> ***Assumption 1:*** *$C \perp T \mid \boldsymbol{X}$, i.e., the censoring and failure time are independent, given the covariates.*
>
> ***Assumption 2:*** *$C \perp \boldsymbol{X}$, i.e., the censoring time does not depend on covariates.*

Then, at a fixed time t for which $\sum_{i=1}^n Y_i(t) \nrightarrow_p 0$ and , $\sum_{i=1}^n \delta_i (1-Y_i(t)) \nrightarrow_p 0$,

$$(\hat{\mu}_j(t), \hat{\sigma}_j^2(t), \hat{\theta}_j(t), \hat{\psi}_j^2(t)) \rightarrow_p (\mu_j(t), \sigma_j^2(t), \theta_j(t), \psi_j^2(t))$$

Assumption 1 (a.k.a. “random censoring”) is fairly standard in many survival analysis settings. The assumption can be slightly weakened if there is a known set of discrete-valued variables $\boldsymbol{Y}$ such that the failure and censoring times are conditionally independent given $\boldsymbol{Y}$, in which case one could consider estimating the survivor function $P(T \geq t) = \int_{\boldsymbol{y}} P(T \geq t \mid \boldsymbol{R} = \boldsymbol{r}) dF_{\boldsymbol{R}}(\boldsymbol{r})$, where $P(T \geq t \mid \boldsymbol{R} = \boldsymbol{r})$ can be consistently estimated within the subset of individuals with $\boldsymbol{R} = \boldsymbol{r}$. Assumption 2 is a less common and fairly strong assumption which implies that censoring time does not depend on individual characteristics. While in some settings this assumption may be unrealistic, it is relatively reasonable in our EHD setting where most censoring is induced either by the end of the data capture window or by subjects disenrolling from the healthcare plan, often due to an employer switching plans. Note that Assumptions 1 and 2, taken together, are equivalent to the assumption that $C \perp (T, \boldsymbol{X})$.

An immediate corollary of Theorem 1 is that the density estimates obtained by plugging $(\hat{\mu_j}, \hat{\sigma_j})$ into $f_j^{\geq}$ and $(\hat{\theta_j}, \hat{\psi_j})$ into $f_j^{<}$ are consistent for the true densities provided $X_j|T \geq t$ and $X_j|T > t$ follow the assumed Normal distributions. We note, however, that the consistency of the density estimates does not guarantee consistency of $\hat{S}(\boldsymbol{x}, t)$ if the Naive Bayes assumption fails. Indeed, as we show in the simulations, while the classification performance of CNB is not dramatically affected by correlation between predictors, conditional survival probability estimates are severely biased. In the next section, we develop a modification to the CNB technique which may improve calibration in situations where the Naive Bayes assumption of predictor independence is violated.

3.5. CNB-PC

The Naive Bayes assumption treats the available covariates as independent for the purposes of estimating their joint distribution. Hence, it will perform best when the design matrix $\mathcal{X}$ is close to orthogonal. Therefore, we propose to orthogonalize the design matrix via singular value decomposition (SVD) prior to applying the estimation approaches described in the previous sections.

The matrix $\mathcal{X}^T \mathcal{X}$, which is proportional to the empirical covariance matrix of $\mathcal{X}$, can be written using SVD as

$$\mathcal{X}^T \mathcal{X} = \mathcal{W} \boldsymbol{\Lambda} \mathcal{W}^T$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of and $\mathcal{W}$ the eigenvectors of $\mathcal{X}^T \mathcal{X}$. By convention, we often rearrange $\boldsymbol{\Lambda}$ and $\mathcal{W}$ such that the eigenvalues making up the diagonal of $\boldsymbol{\Lambda}$ are in decreasing order. Let

$$\mathcal{D} = \mathcal{X} \mathcal{W}$$

The matrix $\mathcal{D}$ is called the *principal component decomposition* of $\mathcal{X}$. The key property of $\mathcal{D}$ is that its columns are orthogonal, since

$$\mathcal{D}^T \mathcal{D} = \mathcal{W}^T \mathcal{W} \boldsymbol{\Lambda} \mathcal{W}^T \mathcal{W} = \boldsymbol{\Lambda}$$

as eigenvectors corresponding to unique eigenvalues are orthogonal. Exploiting this fact, we suggest the following technique, which we refer to as *Censored Naive Bayes - Principal Components* (CNB-PC):

1. Compute the principal component decomposition $\mathcal{D}$ of the (standardized) design matrix $\mathcal{X}$, yielding the eigenvector matrix $\mathcal{W}$.
2. Given a fixed time t and a $p \times 1$ vector of covariates $\boldsymbol{x}$ for which one wishes to estimate $S(\boldsymbol{x}, t)$,

a. Let $\boldsymbol{d} = \mathcal{W}^T\boldsymbol{x}$, and apply the CNB method as previously described to estimate $\Psi(\boldsymbol{d}, t)$ and hence $S(\boldsymbol{d}, t)$.

b. Set $\hat{S}(\boldsymbol{x}, t) = \hat{S}(\boldsymbol{d}, t)$.

While the principal component decomposition results in a set of covariates likely to satisfy the marginal relationship $f(\boldsymbol{D}) \approx \prod_{j=1}^{p} f(D_j)$, it does not imply that this relationship is satisfied conditional on $T \geq t$ or $T < t$. However, based on results from our simulation studies and real data analysis, the technique appears to perform quite well, possibly because the marginal and conditional correlations between predictors are often similar.

Above, we proposed to use the full matrix of eigenvectors $\mathcal{W}$ to obtain $\mathbf{d} = \mathcal{W}^T\boldsymbol{x}$. However, principal component decomposition is widely applied as a dimension reduction technique by truncating $\mathcal{W}$ to $\mathcal{W}_{-K}$, where $\mathcal{W}_{-K}$ contains the rows of $\mathcal{W}$ corresponding to the $p - K$ largest eigenvalues. This approach could easily be used as part of CNB-PC by setting $\mathbf{d}_{-K} = \mathcal{W}_{-K}^T\mathbf{x}$ and $\hat{S}(\boldsymbol{x},t) = \hat{S}(\mathbf{d}\_k,t)$. Truncating $\mathcal{W}$ in this way could substantially speed up computation if p is large and a small number of eigenvectors account for most of the variance in $\mathbf{X}$, so that $\prod_{j=1}^{p} f_j^{\geq}(x;t) \approx \prod_{j=1}^{p-K} f_j^{\geq}(\mathbf{d}_{-K};t)$ and $\prod_{j=1}^{p} f_j^{<}(x;t) \approx \prod_{j=1}^{p-K} f_j^{<}(\mathbf{d}_{-K};t)$.

3.6. Implementation

Software implementing the Censored Naive Bayes techniques described here is available on the website of the first author (Wolfson). The algorithms are fast; in our simulation studies, unoptimized R code for obtaining predictions from CNB (and CNB-PC) had faster running time than extracting predictions from a coxph model object using the survival package. In our data application, fitting the CNB and CNB-PC models to the training dataset with > 60,000 observations and obtaining predictions on the > 20, 000 observations in the test set took less than two seconds on a desktop computer. The assumption of covariate independence allows the CNB/CNB-PC algorithm to be easily parallelized, which is not true of procedures for maximizing the Cox partial likelihood.

4. Simulations

One theoretical advantage of the proposed Bayesian network approach is that it does not depend on the assumption that hazards are proportional. Hence, NB might be better suited to modeling data which do not follow a proportional hazards model, or for which the functional relationship between covariates and the hazard is misspecified. In this section, we compare the performance of the CNB models with the Cox proportional hazards model (CPH) when 1) failure times follow a proportional hazards model, 2) failures times follow a log-logistic accelerated failure time (AFT) model such that the hazards are non-proportional, and 3) the manner in which covariates influence the failure time is misspecified in the regression model. In all cases, censoring times were generated as min(10, *Unif*(0, 20)), thereby guaranteeing the conditional and unconditional independence of censoring and failure times. Observed event times O and censoring indicators δ were defined in the usual way: $O = \min(T, C)$ and $\delta = \mathbb{1}[T \leq C]$. As in our real data example (Section 5), we consider predicting

the probability of surviving beyond 5 years, $S(\boldsymbol{x}, 5) = P(T \geq 5 \mid \boldsymbol{x})$. All tables are based on 5000 simulations.

We report on both the model calibration and discrimination of the CNB and CPH techniques. Since the true (i.e., simulated) event times and survival probabilities are fully known for the simulated test data, it is straightforward to assess these characteristics. Model calibration was assessed by considering bias and mean squared error (MSE) of the predicted probabilities with respect to the true event probabilities. We report conditional bias and MSE within specific population subgroups, as well as marginal measures which average across the entire simulated population. Model discrimination was assessed via the area under the ROC curve (C-index/AUC). In the real data analyzed in Section 5, the event times in the test data are subject to censoring, hence we use modified versions of the C-index which account for censoring; further details appear in Section 5.3.

4.1. Proportional hazards

Table 1 summarizes the performance of the CPH and NB models when data are generated from a proportional hazards model using the method suggested in [35]. Covariates $\boldsymbol{X}_c = (X_1, X_2, \ldots, X_p)$ were generated from a p-variate Normal distribution with mean zero, variance one, and exchangeable correlation with pairwise correlation ρ. Different values of ρ were considered in the various simulation settings. The final design matrix was defined by $\boldsymbol{X} = \mathbf{[1}\ \boldsymbol{X}_c\mathbf{]}$, where $\mathbf{1}$ is an n-vector of ones. Because each covariate was marginally standard normal, no covariate scaling was performed before applying the computational methods. Given observed covariates $\boldsymbol{x}_i$ for subject i, a failure time was generated as

$$T_i = -\left(\frac{\log(U_i)}{\lambda \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}\right)^{\nu}$$

where $U_i \sim Unif\,(0, 1)$, which corresponds to generating failure times from a Weibull distribution with scale parameter $\lambda(\boldsymbol{x}_i) = \lambda \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)$, so that the hazard for subject i is proportional to the baseline hazard with proportionality factor $\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)$. For this set of simulations, we set $\lambda = 0.01$, $\nu = 2$, and

$$\boldsymbol{\beta} = (\beta_0, 0.5, 0, 0, 0, 0)$$

β_0 was set to 0, −1, and −2 across simulations, yielding data with true event rates of 24%, 10%, and 4%, and censoring rates of 80%, 92%, and 97%, respectively. Across all scenarios, about 24% of subjects were censored before accruing less than five years of follow-up.

When covariates are independent ($\rho = 0$), CNB, CNB-PC have small bias and MSE, while CPH has better discrimination as measured by AUC, particularly when the sample size is $n = 1000$ and when the event rate is smaller ($\beta_0 = -1$ and -2). When covariates are strongly correlated ($\rho = 0.7$), CNB has large bias and MSE, while CNB-PC retains similar performance to the zero-correlation scenario. Interestingly, the bias of CNB does not

degrade the discrimination ability of CNB as measured by the AUC; in fact, in the scenarios where $n = 1000$ and $\beta_0 \in \{-1,-2\}$, AUC of the CNB is greater for the $\rho = 0.7$ case where bias and MSE are substantial. This finding is not entirely surprising; it has been shown that Naive Bayes tends to drive predicted probabilities towards either 0 or 1 [32] so that calibration may poor while an approximately correct ranking of observations with respect to their risk is maintained.

4.2. Nonproportional hazards

Table 2 compares the performance of the NB with the CPH model when data follow a log-logistic accelerated failure time (AFT) model, and covariates are generated as described above. Given observed covariate $\boldsymbol{x}_i$, failure times were generated from a Log-logistic distribution with shape parameter $\exp(\boldsymbol{\beta}\boldsymbol{x}_i)$ and scale parameter $\varphi = 20$. Censoring times were generated as $\min(10, \mathit{Unif}(0, 20))$. $\boldsymbol{\beta}$ was set as

$$\boldsymbol{\beta}=(\beta_0, 0.5, 0.1, -0.1, 0, 0)$$

β_0 was set to 1, 0, and −1 across simulations, yielding data with true event rates of 36%, 20%, and 5%, and censoring rates of 65%, 82%, and 96%, respectively. The proportion of subjects censored with less than five years of follow-up ranged from 18–24%.

As in the proportional hazards case, bias and MSE are small for all three methods when covariates are independent, but is higher for CNB when covariates were correlated while CNB-PC retains acceptable calibration in this case. CNB-PC has both smaller bias and MSE than the proportional hazards model while retaining comparable AUC in scenarios where $n = 5000$, showing the advantage conferred by this more flexible model.

4.3. Regression model misspecification

Table 3 summarizes the performance of CNB, CNB-PC, and CPH models when data are generated from a Log-logistic model and the CPH linear predictor is misspecified. The simulation model emulates a scenario where the effect of age on risk is non-linear: subjects at middle ages are at highest risk, while those at low and high age are at lower risk. In our data, such a pattern is observed when modeling the probability of certain nonfatal CV events; young subjects are at low risk of CV events, and among older subjects these events are more likely to be fatal.

For this set of simulations, the total sample size was fixed at $n = 5000$. We started by generating an ‘age’ covariate from a Log-logistic distribution with shape parameter 10, scale parameter 50, and rate parameter 0.4; this yielded a distribution of ages with a median of 50 and an interquartile range of 10. A small fraction (<0.1%) of simulated ages were > 100; these ages were set to 100. Given age= a, systolic blood pressure values were simulated from a Normal distribution with mean $\text{mean } 130+sign(a-50)\sqrt{|a-50|}$ and standard deviation 15. Based on the simulated ages and SBPs, failure times were simulated from a Log-logistic distribution with scale parameter $\varphi = 20$ and shape parameter

$$\lambda(age, sbp) = \exp\left[-0.2 \cdot age - 0.6 \cdot SBP + 0.2 \cdot (age > 60) - 0.2 \cdot age \times SBP + 0.4 \cdot age \times (age > 60)\right] \quad (11)$$

where (*age* > 60) is a binary indicator, *age* × *sbp* is an age-by-SBP interaction, and *age* × (*age* > 60) is an interaction between continuous age and binary indicator of age > 60. All columns of the design matrix implied by this linear predictor were standardized to have mean zero and variance one before estimation was performed, so that the coefficients represent standardized effect sizes for each covariate. The left panel of Figure 1 shows a set of 5-year event probabilities simulated from this model and plotted vs. age; the right panel plots these event probabilities vs. SBP.

We consider the performance of CNB, CNB-PC and CPH models for different specifications of the linear predictor in the Cox model. There is no explicit regression model associated with CNB and CNB-PC, and hence it is provided with (standardized) vectors of ages and SBPs. We consider varying degrees of misspecification of the regression model, from a model incorporating only linear terms for age and SBP (omitting the (age > 60) indicator and higher-order interactions) to the full model including all the covariate terms from (11).

Table 3 and Figure 1 illustrate how the additional flexibility of the CNB models allows more accurate predictions to be obtained when relevant covariates are omitted from the CPH model. Note that the performance of the CNB/CNB-PC models are identical across the different CPH regression model specifications as the CNB models are always provided with the same set of information (standardized vectors of age and SBP). The CNB models have smaller MSE when compared to a CPH model which includes only main effects for age (or log(Age)) and SBP. Figure 1 displays predicted survival probabilities from a typical simulation run using the CNB-PC and the main effects only CPH model. The overlaid smoothers show clearly that CNB/CNB-PC capture the non-linear effects of age and SBP observed in the true probabilities (left panel) better than CPH. Returning to Table 3, adding an age × SBP interaction term does not markedly improve the performance of CPH. Incorporating a quadratic effect for age, a standard strategy when faced with apparent non-linearity in the effect of a covariate, yields some improvement in CPH but CNB/CNB-PC retain an advantage in MSE. When the (Age> 60) indicator term is included so that the regression model is well specified, CPH outperforms CNB/CNB-PC both in terms of MSE and AUC.

5. Data application

5.1. Study population

We illustrate our Censored Naive Bayes method by applying it to the task of predicting the risk of cardiovascular events from longitudinal electronic health record 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 [36, 37, 38]. The VDW stores data including insurance enrollment, demographics, pharmaceutical dispensing, utilization, vital signs, laboratory, census, and death records. This health care 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.

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) 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; we ignored gaps in enrollment less than 90 days and considered a patient-member continuously enrolled over this period; 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, with drug coverage at the end of the baseline period; 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. 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.

5.2. Predictors and outcomes

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. We also 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%, and approximately 50% of subjects were censored with < 5 years of follow-up. Risk factors used as predictors in our models included age, gender, systolic blood pressure (SBP), cholesterol markers (HDL) and body mass index (BMI). Missing risk factor values were filled in prior to model fitting using imputation by chained equations [39] to create a single dataset with no missing values.

5.3. Model evaluation metrics for censored data

In our simulation study, it was straightforward to compute the prediction and classification error of our models via MSE and AUC as the true underlying event times and status were known for all subjects. However, with real-life data, we must use calibration and reclassification metrics that account for censoring. Throughout this section, we will use the shorthand $\hat{\pi_i}$ to denote the predicted *t*-year event probability for subject *i*, i.e., $\hat{\pi_i} = 1 - \hat{S}(\boldsymbol{x}_i, t)$.

5.3.1. Calibration—To accommodate censoring, we consider a calibration statistic which compares the average predicted risk to the empirical risk within a set of bins $\beta_1, \beta_2, \ldots, \beta_m$ defined by clinically relevant cut points. Since the event times are censored, the empirical risk (and its variance) within each bin is calculated via Kaplan-Meier. The calibration statistic *K* takes the form:

$$K=\sum_{k=1}^{m}\frac{(\overline{\pi}_k-\hat{p}_k^{KM})^2}{\hat{var}(\hat{p}_k^{KM})} \quad (12)$$

where *m* is the number of bins, $\hat{\pi_k} = \Sigma_{i\in\beta_k}\hat{\pi_i}/|\beta_k|$ is the average of predicted probabilities in bin *k*, and $\hat{p}_k^{KM}$ is the Kaplan-Meier estimate of experiencing an event before *t* for subjects in bin *k*. $\hat{var}\{\hat{p}_k^{KM}\}$ is the sampling variance of the Kaplan-Meier estimator of the event rate calculated using Greenwood's formula [40] applied to the data in bin *k*:

$$\hat{var}(\hat{p}_k^{KM})=(1-\hat{p}_k^{KM})^2\sum_{t_i<t}\frac{d_{ik}}{n_{ik}-d_{ik}} \quad (13)$$

where d_{sk} is the number of events occurring at time t_s in bin *k*, and n_{sk} are the number of subjects at risk for an event at time t_s (i.e., not censored and not experiencing an event before time t_i). *K* is analogous to the χ^2 statistic with $m - 2$ degrees of freedom for assessing the calibration of logistic models suggested by Hosmer and Lemeshow [41, 42].

5.3.2. Concordance index—When the outcome is fully observed on all subjects (as in the simulations earlier), the AUC 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, but 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 [43]. 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. The C-index adapted for censoring is therefore given by

$$C_{cens}=\frac{\sum_{i\neq j}\delta_i\mathbb{1}[O_i<O_j]\mathbb{1}[\hat{\pi}_i<\hat{\pi}_j]}{\sum_{i\neq j}\delta_i\,\mathbb{1}\,[O_i<O_j]} \quad (14)$$

where $\mathbb{1}[\cdot]$ is the indicator function.

5.4. Results

The cohort was randomly split into a training dataset consisting of 75% of the observations (65,522 subjects), and a test dataset consisting of the remaining 25% (21,841 subjects). After standardizing the covariate values to have zero mean and unit variance, we applied the CNB, CNB-PC, and CPH methods to the training data and obtained predictions for the test data. The CPH model regression equation included linear terms for each of the risk factors (age, blood pressure, LDL cholesterol, and BMI). Risk factor values were log-transformed, consistent with existing techniques such as the Framingham risk equations.

Table 4 summarizes the calibration and reclassification performance of CPH, CNB, and CNB-PC on the test data. CNB is rather poorly calibrated, overpredicting the risk in four out of five risk categories by 15–35%. Both CPH and CNB-PC are well calibrated, though CNB-PC is slightly better. Under the null hypothesis of perfect calibration, the CNB-PC calibration statistic has a χ^2_3 distribution and $P(\chi^2_3 \geq 8.8) = 0.032$ so the null hypothesis of good calibration for CNB-PC is barely rejected at the $\alpha = 0.05$ level. Classification performance of all three models–including the poorly-calibrated CNB–is similar, with a C-index ranging from 0.787 to 0.789. One interesting observation is that CPH and CNB-PC achieve near-identical calibration and reclassification performance despite disagreeing on the risk predictions for a non-trivial fraction of subjects: approximately 12.5% of subjects are placed in different risk bins (e.g., 0–5% vs. 5–10%) by the two methods. The percentage of discordant predictions for CPH vs. CNB and CNB vs. CNB-PC are 16.5% and 17%, respectively.

While the improvement offered by CNB-PC over CPH in this application is modest at best, it is important to note that the proportional hazards model is a very worthy competitor for the task of cardiovascular risk prediction: the well-established Framingham risk model as well as a recently-developed model designed to replace it [44] are both based on CPH regression. The success of these relatively simple CPH models suggests that, in datasets used for cardiovascular risk prediction, the proportional hazards assumption is not severely violated and the relationships between predictors and risk are relatively linear. The main benefit of CNB/CNB-PC is its flexibility, which may allow it to achieve greater gains over CPH on datasets in settings where the proportional hazards assumption is clearly violated, or when predictor-outcome relationships are highly non-linear.

6. Discussion

We have presented Censored Naive Bayes, an extension of the Naive Bayes technique to accommodate time-to-event data subject to censoring, and illustrated its application to predicting cardiovascular risk using electronic health record data. Our method, like many machine learning techniques, is ‘model-agnostic’ and hence can capture complex relationships between demographics, biomarkers, and clinical risk without requiring *a priori* specification of a regression model. While it is true that a more complex Cox regression model or a set of stratified Kaplan-Meier estimates might capture observed non-linearities in risk, considering multiple candidate regression models or covariate strata increases the risk

of overfitting, even when an independent test set is used. In contrast, our relatively simple Censored Naive Bayes approach captures non-linear effects without requiring any model tuning, scoring, or comparisons.

CNB assumes that covariates are independent and normally distributed conditional on both $T \geq t$ and $T < t$ for all t. This is unlikely to hold in practice, though Naive Bayes has been shown to be effective in other settings where distributional assumptions regarding the covariates are not satisfied. Indeed, while most covariates in our simulation study were generated from marginal normal distributions, no constraint was placed on their conditional distributions. We found in our simulations that predictions from CNB were biased when between-covariate correlation was high, though this bias was largely corrected by applying the CNB-PC technique. However, many of the poorly calibrated CNB models had acceptable risk reclassification performance. In clinical practice, the ability to discriminate between low- and high-risk patients may be more important than obtaining a precise risk estimate, which may recommend the use of Censored Naive Bayes in such settings.

The assumption of conditional normal distributions (given $T \leq t$ and $T > t$) is somewhat limiting in that it restricts CNB/CNB-PC to cases where covariates are continuous and have distributions which are approximately Normal. In practice, separate CNB models could be fit at each level of categorical risk factors (e.g., smokers and non-smokers), but this could be time-consuming for factors with more than two levels and inefficient for ordered factors (e.g., tumor staging). Furthermore, many risk factors of clinical disease have skewed distributions either marginally or among subjects with survival times beyond a certain threshold. In ongoing work, we are developing techniques which accommodate discrete and skewed covariates in the CNB framework.

Acknowledgments

The authors wish to thank the reviewers for their comments, which undoubtedly improved the manuscript. This work was partially supported by NHLBI grant #R01HL102144-01 and AHRQ grant #R21HS017622-01.

References

1. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. American Heart Journal. Jan; 1991 121(1 Pt 2):293–8. http://www.ncbi.nlm.nih.gov/pubmed/1985385. [PubMed: 1985385]
2. D'Agostino RB, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. Journal of the American Medical Association. Jul; 2001 286(2):180–7. http://www.ncbi.nlm.nih.gov/pubmed/11448281. [PubMed: 11448281]
3. Kagan A, Dawber TR, Kannel WB, Revotskie N. The Framingham study: a prospective study of coronary heart disease. Federation proceedings. 1962; 21 (4 Pt 2):52–7. http://www.ncbi.nlm.nih.gov/pubmed/14453051. [PubMed: 14453051]
4. Arruda, H. About the Framingham Heart Study Participants. http://www.framinghamheartstudy.org/participants/index.html
5. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. Feb; 2008 117(6):743–53. http://www.ncbi.nlm.nih.gov/pubmed/18212285. 10.1161/CIRCULATIONAHA.107.699579 [PubMed: 18212285]

6. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972; 34(2):187–220. http://links.jstor.org/sici?sici=0035-9246(1972)34:2<187:RMAL>2.0.CO;2-6.
7. Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66(3):429–436.
8. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care. Jun; 2010 48(6 Suppl):S106–13. http://www.ncbi.nlm.nih.gov/pubmed/20473190. 10.1097/MLR.0b013e3181de9e17 [PubMed: 20473190]
9. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Medinfo. 2004; 11(1):736–740.
10. Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F.; Degoulet, P.; Jaulent, MC. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000. Models to predict cardiovascular risk: Comparison of CART, multilayer perceptron and logistic regression; p. 156-160.
11. Chia CC, Rubinfeld I, Scirica BM, McMillan S, Gurm HS, Syed Z. Looking Beyond Historical Patient Outcomes to Improve Clinical Models. Science Translational Medicine. Apr.2012 4(131): 131ra49. http://stm.sciencemag.org/content/4/131/131ra49.abstracthttp://www.ncbi.nlm.nih.gov/pubmed/22539773. 10.1126/scitranslmed.3003561
12. Larranaga P, Sierra B, Gallego MJ, Michelena MJ, Picaza JM. Learning Bayesian networks by genetic algorithms: A case study in the prediction of survival in malignant skin melanoma. Artif Intell Med. 1997; 1211:261–272. http://www.springerlink.com/index/w31m361t2v5q1357.pdf.
13. Sierra B, Larranaga P. Predicting survival in malignant skill melanoma using Bayesian networks automatically induced by genetic algorithms: An empirical comparison between different approaches. Artif Intell Med. 1998; 14(1–2):215–230. [PubMed: 9779891]
14. Blanco R, Inza M, Merino M, Quiroga J, Larranaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with *{*TIPS*}*. J Biomed Inform. 2005; 38(5): 376–388. [PubMed: 15967731]
15. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. Comput Biomed Res. 1998; 31(5):363–373. [PubMed: 9790741]
16. Stajduhar I, Dalbelo-Basic B, Bogunovic N. Impact of censoring on learning Bayesian networks in survival modelling. Artif Intell Med. 2009; 47(3):199–217. [PubMed: 19833488]
17. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. 2008:841–860.
18. Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med. Mar; 2004 30(3):201–214. [PubMed: 15081072]
19. Zupan B, Demsar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: A case study on recurrence of prostate cancer. Artif Intell Med. 1999; 1620:346–355.
20. Stajduhar I, Dalbelo-Basić B. Learning Bayesian networks from survival data using weighting censored instances. J Biomed Inform. Aug; 2010 43 (4):613–622. http://dx.doi.org/10.1016/j.jbi.2010.03.005. 10.1016/j.jbi.2010.03.005 [PubMed: 20332035]
21. Stajduhar I, Dalbelo-Basić B. Uncensoring censored data for machine learning: A likelihood-based approach. Expert Syst Appl. Jun; 2012 39(8):7226–7234. http://dx.doi.org/10.1016/j.eswa.2012.01.054. 10.1016/j.eswa.2012.01.054
22. Domingos, P.; Pazzani, M. Machine Learning. Morgan Kaufmann; 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier; p. 105-112.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.4134http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.1930
23. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1997; 29:103–130. http://link.springer.com/article/10.1023/A:1007413511361.
24. Hand DJ, Yu K. Idiot's Bayes: Not So Stupid After All? International Statistical Review. Dec; 2001 69(3):385–398. http://doi.wiley.com/10.1111/j.1751-5823.2001.tb00465.x. 10.1111/j.1751-5823.2001.tb00465.x
25. Singh, S. Using Naive Bayes for Data Clustering. University of New Brunswick; Canada: 2005. http://books.google.com/books?id=LuO576rEXCIC&pgis=1

26. Maimon, OZ.; Rokach, L. Theory And Applications. World Scientific; 2005. Decomposition Methodology For Knowledge Discovery And Data Mining. http://books.google.com/books?id=RARx4rjp_7AC&pgis=1
27. Pearl J. Fusion, propagation, and structuring in belief networks. Artificial Intelligence. Sep; 1986 29(3):241–288. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.8016http://www.sciencedirect.com/science/article/pii/000437028690072X. 10.1016/0004-3702(86)90072-X
28. John, GH.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence; 1995;
29. Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Gelpke GJ. Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. Journal of the Royal Statistical Society Series A (General). 1981; 144(2):145. http://www.citeulike.org/user/wkretzsch/article/9414776. 10.2307/2981918
30. Mani, S.; Pazzani, MJ.; West, J. Artificial Intelligence in Medicine. Springer; Berlin/Heidelberg: 1997. Knowledge Discovery from a Breast Cancer Database; p. 130-133.chap. Knowledgehttp://www.researchgate.net/publication/221450627_Knowledge_Discovery_from_a_Breast_Cancer_Database
31. Russek E, Kronmal RA, Fisher LD. The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. Computers and biomedical research. Dec; 1983 16(6):537–52. http://www.ncbi.nlm.nih.gov/pubmed/6653089. [PubMed: 6653089]
32. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. ICML. 2001; 1:609–616. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.3039&rep=rep1&type=pdf.
33. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Assocation. Nov; 1958 53(282):457–481. http://www.jstor.org/stable/10.2307/2281868http://www.jstor.org/pss/2281868.
34. van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer Science & Business Media; 2003. http://books.google.com/books?hl=en&lr=&id=z4_-dXslTyYC&pgis=1
35. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine. Jun; 2005 24 (11):1713–23. http://www.ncbi.nlm.nih.gov/pubmed/16680804. 10.1002/sim.2059 [PubMed: 15724232]
36. Selby JV. Linking automated databases for research in managed care settings. Annals of internal medicine. Oct; 1997 127(8 Pt 2):719–24. http://www.ncbi.nlm.nih.gov/pubmed/9382386. [PubMed: 9382386]
37. Platt R, Davis R, Finkelstein J, Go AS, Gurwitz JH, Roblin D, Soumerai S, Ross-Degnan D, Andrade S, Goodman MJ, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiology and drug safety. 2001; 10(5):373–7. http://www.ncbi.nlm.nih.gov/pubmed/11802579. 10.1002/pds.607 [PubMed: 11802579]
38. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Annals of internal medicine. Sep; 2009 151(5):341–4. http://www.ncbi.nlm.nih.gov/pubmed/19638403. [PubMed: 19638403]
39. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011; 45(3):1–67. http://www.jstatsoft.org/v45/i03/.
40. Greenwood M. A Report on the Natural Duration of Cancer. Reports on Public Health and Medical Subjects. Ministry of Health. 1926; (33)
41. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods. 1980; 9(10):1043–1069. http://www.tandfonline.com/doi/abs/10.1080/03610928008827941. 10.1080/03610928008827941
42. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology. Jan; 1982 115(1):92–106. http://aje.oxfordjournals.org/content/115/1/92.short. [PubMed: 7055134]

43. Harrell, FE. With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer Science & Business Media; 2001. Regression Modeling Strategies. http://books.google.com/books?hl=en&lr=&id=kfHrF-bVcvQC&pgis=1

44. Goff, DC.; Lloyd-Jones, DM.; Bennett, G.; Coady, S.; D'Agostino, RB.; Gibbons, R.; Greenland, P.; Lackland, DT.; Levy, D.; O'Donnell, CJ., et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. Nov. 2013 http://www.ncbi.nlm.nih.gov/pubmed/24222018

Appendix A - Proof of Theorem 1

Theorem 1

Suppose that the following assumptions hold:

Assumption 1: $C \perp T \mid \boldsymbol{X}$, i.e., the censoring and failure time are independent, given the covariates.

Assumption 2: $C \perp \boldsymbol{X}$, i.e., the censoring time does not depend on covariates.

Then, at a fixed time t for which $\sum_{i=1}^{n} Y_i(t) \nrightarrow_p 0$ and $\sum_{i=1}^{n} \delta_i(1-Y_i(t)) \nrightarrow_p 0$,

$$\hat{\mu}_j(t), \hat{\sigma}_j^2(t), \hat{\theta}_j(t), \hat{\psi}_j^2(t) \rightarrow_p \mu_j(t), \sigma_j^2(t), \theta_j(t), \psi_j^2(t)$$

Proof

Let t be an event time such that, for sufficiently large n, $\sum_{i=1}^{n} \delta_i(1-Y_i(t))>0$ and $\sum_{i=1}^{n} Y_i(t)>0$. We start by consider $\hat{\mu_j}(t)$. By writing

$$\hat{\mu}_j(t)=\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(t)X_{ij}}{\sum_{i=1}^{n} Y_i(t)}$$

it is clear by the Weak Law of Large Numbers that the numerator and denominator respectively converge in probability to $E(Y(t)X_j)$ and $E(Y(t))$. Now,

$$\begin{gathered} E(Y(t)X_j)=E[E(Y(t)X_j|Y(t))] \\ =E(Y(t)X_j|Y(t)=0)P(Y(t)=0)+E(Y(t)X_j|Y(t)=1)P(Y(t)=1) \\ =E(X_j|O \geq t)E(Y(t)) \end{gathered}$$

which implies that

$$\hat{\mu}_j(t) \rightarrow_p \frac{E(X_j|O \geq t)E(Y(t))}{E(Y(t))}=E(X_j|O \geq t)$$

A similar argument yields the result

$$\hat{\sigma}_j^2(t) \to_p E(X_j^2|O \geq t) - [E(X_j|O \geq t)]^2 = Var(X_j|O \geq t)$$

Hence to obtain the result for $\hat{\mu_j}$ and $\hat{\sigma}_j^2$ it suffices to show that $E(X_j|O \geq t) = E(X_j|T \geq t) \equiv \mu_j(t)$ and similarly $E(X_j^2|O \geq t) = E(X_j^2|T \geq t)$.

Using f_j as shorthand to denote the density function of X_j (which we assume exists), we have

$$\begin{aligned} f_j(x|O \geq t) &= f_j(x|\min(T, C) \geq t) \\ &= f_j(x|T \geq t, C \geq t) \\ &= \frac{P(T \geq t, C \geq t|X_j = x) f_j(x)}{\int_r P(T \geq t, C \geq t|X_j = r) f_j(r) dr} \\ &= \frac{P(T \geq t|X_j = x) P(C \geq t|X_j = x) f_j(x)}{\int_r P(T \geq t|X_j = r) P(C \geq t|X_j = r) f_j(r) dr} \text{ by Assumption 1} \\ &= \frac{P(T \geq t|X_j = x) f_j(x)}{\int_r P(T \geq t|X_j = r) f_j(r) dr} \frac{P(C \geq t)}{P(C \geq t)} \text{ by Assumption 2} \\ &= f_j(x|T \geq t) \end{aligned}$$

and hence all functionals of $f_j(\cdot|O \geq t)$ are equal to the corresponding functionals of $f_j(\cdot|T \geq t)$.

The results for $\hat{\theta_j}$ and $\hat{\psi}_j^2$ follow standard arguments from the inverse probability of censoring weighting literature. Let $\delta_i(t) = \delta_i \times (1 - Y_i(t)) = \mathbb{1}[T_i < C_i] \cdot \mathbb{1}[T_i < t]$ be the indicator that a subject experienced an event prior to time t. Then

$$\hat{\theta}_j(t) = \frac{\frac{1}{n}\sum_{i=1}^n \omega_i \delta_i(t) X_{ij}}{\frac{1}{n}\sum_{i=1}^n \omega_i \delta_i(t)} \quad \text{and} \quad \hat{\psi}_j^2(t) = \frac{\frac{1}{n}\sum_{i=1}^n \omega_i \delta_i(t) X_{ij}^2}{\frac{1}{n}\sum_{i=1}^n \omega_i \delta_i(t)} - \hat{\theta}_j^2(t)$$

We first focus on $\hat{\theta_j}(t)$. We show that

$$\hat{\theta}_j(t) = \frac{\sum_{i=1}^n \hat{\omega}_i \delta_i (1 - Y_i(t)) X_{ij}}{\sum_{i=1}^n \hat{\omega}_i \delta_i (1 - Y_i(t))} \to_p \frac{E(X_j|T<t)P(T<t)}{P(T<t)} = E(X_j|T<t)$$

By the WLLN, we have $\frac{1}{n}\sum_{i=1}^n \omega_i X_{ij} \delta_i(t) \to_p E(\omega X_j \delta(t))$ and

$$
\begin{aligned}
E(\omega X_j \delta(t)) &= E\left(\tfrac{[\min(T,t)<C]}{G(\min(T,t))} X_j \delta(t)\right) \\
&= E\left[E\left(\tfrac{[\min(T,t)<C]}{G(\min(T,t))} X_j \delta(t) | X_j, T\right)\right] \\
&= E\left[\tfrac{X_j}{G(\min(T,t))} E\left(\mathbb{1}[\min(T,t)<C]\delta(t) | X_j, T\right)\right] \\
&= E\left[\tfrac{X_j}{G(\min(T,t))} E\left(\mathbb{1}[\min(T,t)<c] \cdot \mathbb{1}[T<t] | X_j, T\right)\right] \\
&= E\left[\tfrac{X_j}{G(\min(T,t))} \mathbb{1}[T<t] E(\mathbb{1}[\min(T,t)<C] | X_j, T)\right] \\
&= E\left[\tfrac{X_j}{G(\min(T,t))} \mathbb{1}[T<t] P(C>\min(T,t))\right] \quad \text{by Assumption 2.} \\
&\equiv E\left[\tfrac{X_j}{G(\min(T,t))} \mathbb{1}[T<t] G(\min(T,t))\right] \\
&= E(X_j \mathbb{1}[T<t]) \\
&= E(X_j | T<t) P(T<t)
\end{aligned}
$$

A similar argument yields $E(\omega\delta(t)) = P(T < t)$, giving the final result for $\hat{\theta_j}$. The calculation for $\hat{\psi}_j^2(t)$ is identical with the exception of replacing X_{ij} by X_{ij}^2 and using the consistency of $\hat{\theta_j}(t)$ just shown.

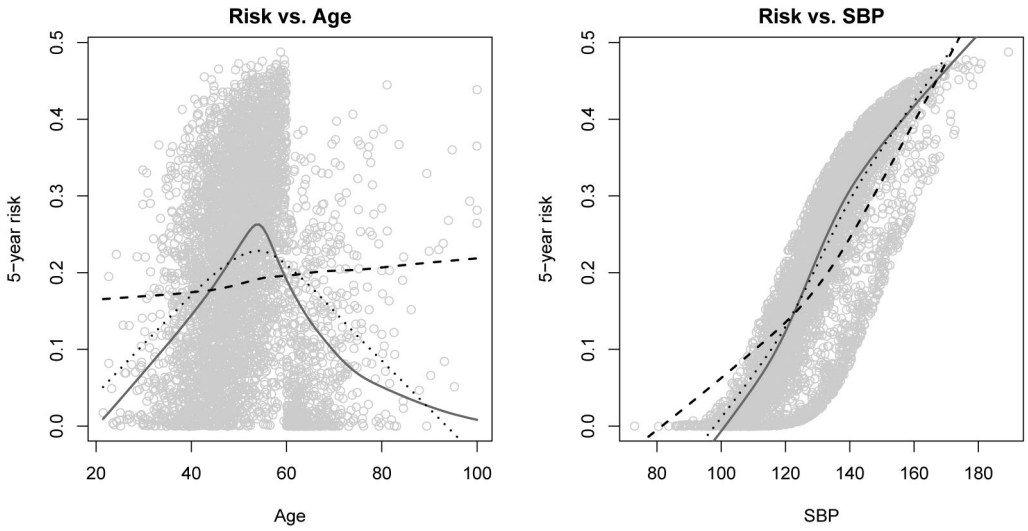

Figure 1.

True and predicted risks of experiencing an event within 5 years based on a Log-logistic model with shape parameter defined by (11), plotted versus age (left panel) and SBP (right panel). On each plot, the points show (for a typical simulation run) true simulated risks and the solid line is a lowess scatterplot smooth showing the trend in risk across age and SBP respectively. The dashed lines in each panel show the trend in risk predictions obtained from the Cox Proportional Hazards model incorporating linear main effects for age and SBP (first row of Table 3), while the dotted lines show the trend in predictions by Naive Bayes.

Table 1

Performance of CNB, CNB-PC, and CPH models on data simulated from Weibull (proportional hazards) model. Bias, mean squared error (MSE) and area under the curve (AUC) are computed over the entire validation set; bias and MSE are multiplied by 100.

			Bias			MSE			AUC		
n	β_0	ρ	***CNB***	***CNB-PC***	***CPH***	***CNB***	***CNB-PC***	***CPH***	***CNB***	***CNB-PC***	***CPH***
1000	0	0.0	0.035	0.076	0.825	0.23	0.24	0.08	0.643	0.642	0.652
1000	0	0.7	−5.056	0.040	0.809	3.86	0.23	0.08	0.638	0.643	0.652
1000	−1	0.0	0.034	0.043	0.161	0.13	0.13	0.04	0.620	0.620	0.639
1000	−1	0.7	−5.081	0.030	0.154	2.54	0.13	0.04	0.628	0.620	0.640
1000	−2	0.0	0.029	0.031	0.038	0.06	0.06	0.01	0.593	0.592	0.629
1000	−2	0.7	−3.277	0.025	0.031	1.34	0.06	0.01	0.617	0.592	0.629
5000	0	0.0	0.004	0.047	0.761	0.05	0.05	0.02	0.652	0.652	0.654
5000	0	0.7	−5.072	0.028	0.758	3.80	0.05	0.02	0.639	0.652	0.654
5000	−1	0.0	0.010	0.023	0.141	0.03	0.03	0.01	0.639	0.639	0.643
5000	−1	0.7	−5.174	0.020	0.143	2.46	0.03	0.01	0.629	0.639	0.643
5000	−2	0.0	0.005	0.007	0.024	0.01	0.01	0.00	0.627	0.627	0.638
5000	−2	0.7	−3.283	0.007	0.025	1.13	0.01	0.00	0.626	0.627	0.638

Table 2

Performance of CPH and NB models on data simulated from Log-logistic (non-proportional hazards) model. Bias, mean squared error (MSE) and area under the curve (AUC) are computed over the entire validation set; bias and MSE are multiplied by 100.

			Bias			MSE			AUC		
n	β_0	ρ	CNB	CNB-PC	CPH	CNB	CNB-PC	CPH	CNB	CNB-PC	CPH
1000	−1	0.0	−0.057	−0.061	0.451	0.29	0.30	0.19	0.564	0.563	0.571
1000	−1	0.7	−1.917	−0.059	0.437	1.93	0.30	0.18	0.566	0.561	0.569
1000	0	0.0	0.031	0.052	0.420	0.22	0.24	0.19	0.664	0.663	0.668
1000	0	0.7	−7.353	0.043	0.406	4.27	0.22	0.19	0.653	0.660	0.665
1000	1	0.0	0.106	0.164	0.123	0.11	0.11	0.07	0.798	0.791	0.810
1000	1	0.7	−9.561	0.130	0.107	5.61	0.11	0.06	0.784	0.791	0.807
5000	−1	0.0	0.005	0.004	0.378	0.06	0.08	0.08	0.575	0.574	0.577
5000	−1	0.7	−1.918	0.001	0.359	1.72	0.07	0.08	0.568	0.572	0.575
5000	0	0.0	0.044	0.074	0.386	0.05	0.08	0.12	0.674	0.673	0.675
5000	0	0.7	−7.481	0.057	0.371	4.26	0.07	0.12	0.654	0.670	0.672
5000	1	0.0	0.040	0.108	0.092	0.03	0.03	0.04	0.815	0.813	0.818
5000	1	0.7	−9.818	0.077	0.089	5.78	0.03	0.04	0.785	0.811	0.816

Table 3

Performance of CPH and CNB models for data simulated from a Log-logistic regression model including the variables from (11). The second column of the table specifies which terms are included in the CPH regression model; CNB and CNB-PC are provided with standardized age and SBP values only. The final row for each sample size ("Full Model") shows the results when the regression model is correctly specified. Bias and mean squared errors are multiplied by 100.

		Bias			MSE			AUC		
n	**CPH regression terms**	**CNB**	**CNB-PC**	**CPH**	**CNB**	**CNB-PC**	**CPH**	**CNB**	**CNB-PC**	**CPH**
1000	Age, SBP	−0.108	0.006	0.592	0.41	0.40	0.65	0.709	0.709	0.698
1000	log(Age), SBP	−0.108	0.006	0.587	0.41	0.40	0.66	0.709	0.709	0.698
1000	Age, SBP, Age x SBP	−0.108	0.006	0.615	0.41	0.40	0.65	0.709	0.709	0.699
1000	Age, SBP, Age^2	−0.108	0.006	0.661	0.41	0.40	0.52	0.709	0.709	0.710
1000	Age, SBP, (Age>60)	−0.108	0.006	0.817	0.41	0.40	0.29	0.709	0.709	0.731
1000	Full Model	−0.108	0.006	0.849	0.41	0.40	0.31	0.709	0.709	0.729
5000	Age, SBP	−0.083	0.013	0.570	0.33	0.33	0.61	0.712	0.712	0.699
5000	log(Age), SBP	−0.083	0.013	0.564	0.33	0.33	0.62	0.712	0.712	0.698
5000	Age, SBP, Age x SBP	−0.083	0.013	0.577	0.33	0.33	0.60	0.712	0.712	0.700
5000	Age, SBP, Age^2	−0.083	0.013	0.627	0.33	0.33	0.48	0.712	0.712	0.712
5000	Age, SBP, (Age>60)	−0.083	0.013	0.787	0.33	0.33	0.24	0.712	0.712	0.733
5000	Full Model	−0.083	0.013	0.791	0.33	0.33	0.24	0.712	0.712	0.733

Table 4

Calibration and reclassification performance of the CPH, CNB, and CNB-PC models on the cardiovascular risk test data. The top portion of the table illustrates the calibration within bands defined by the predicted risks from each model; for instance, the first row shows the number of predictions (n), average prediction $\bar{\pi}$ and Kaplan-Meier estimate of the risk $\hat{p_{KM}}$ for subjects predicted to have a 5-year CV risk between 0 and 5% by each of the three methods. The bottom rows give the estimated calibration statistic and C-index for each model.

	CPH			CNB			CNB-PC		
Predicted risk	**n**	$\bar{\pi}$	$\hat{p_{KM}}$	**n**	$\bar{\pi}$	$\hat{p_{KM}}$	**n**	$\bar{\pi}$	$\hat{p_{KM}}$
0–5%	14338	0.023	0.023	14722	0.019	0.023	14353	0.022	0.023
5–10%	3910	0.070	0.056	3139	0.070	0.057	3976	0.070	0.057
10–15%	1482	0.122	0.104	1351	0.123	0.094	1414	0.122	0.121
15–20%	773	0.173	0.180	768	0.173	0.150	734	0.173	0.163
>20%	1338	0.284	0.306	1861	0.341	0.265	1364	0.294	0.299
Calibration statistic		15.0			67.1			8.8	
C-index		0.787			0.788			0.789	