Cardiovascular risk prediction from Electronic Health Records using probabilistic graphical models.

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sunayan Bandyopadhyay

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Chad L. Myers, Paul E. Johnson

June, 2016

Acknowledgements

I want to thank:

Paul E. Johnson, Gedas Adomavicius, Julian Wolfson and David Vock for their support, constructive criticisms, and ideas that made this thesis possible;

Patrick O'Connor and Gabriela Vazquez-Benitez for their valuable insights into the world of medicine and their participation and feedback at presentations;

My committee, Chad Myers, Paul E. Johnson, Gedas Adomavicius, Julian Wolfson, David Vock and Rui Kuang for their support during this process, as well as their accommodation for scheduling meetings and defense;

And finally, Mohamed Elidrisi and Georg Meyer for their collaboration and help in developing my ideas and thoughts.

Abstract

Cardiovascular (CV) disease is one of the leading causes of death in the United States; therefore, it is of vital importance that it be managed and treated effectively. Such treatment requires information to determine optimal strategies for treating complex patients so as to minimize their risk of a CV event. Creating such information requires the availability of predictive models that can estimate the probability of a CV event occurring over a fixed time horizon.

Currently available predictive models are limited because they are constructed from carefully curated cohorts which may not be representative of the population currently under care. This limitation can largely be overcome by using more representative data. Electronic health records (EHR) provide us with such observations which are representative of the population currently being treated by physicians. They provide an attractive platform over which we can construct a predictive model. However, EHR data may have weaknesses, which include missing data and incomplete follow-up. As a result, it is not possible to apply unmodified traditional machine learning algorithms for constructing a predictive model. In this thesis we show how to adapt probabilistic graphical models (PGMs) to censored data with missing observations. In addition, we construct variants of adapted PGMs that allow us to take advantage of different types of historical observations available in the EHR to better predict the risk of CV events.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Prediction of risk of cardiovascular (CV) events over a fixed time horizon obtained from prediction models is of critical importance to physicians providing care to patients with chronic cardiovascular disease. These models are typically an important component of decision support system that are used by practitioners to evaluate treatment strategies for complex patients (i.e. Patients who have one or more uncontrolled risk factors such as hypertension or hyperlipidemia) by helping the physician choose a treatment strategy that minimizes the risk of CV event over a fixed time horizon.

There are several CV risk prediction model that are widely available most of which are relatively simple regression models. These models may be constructed using population cohorts which are not representative of the patient demographic that is being cared for by the physician resulting in risk predictions that are not optimal. While retraining these models can to a certain extent alleviate the "representativeness problem, most of these models fail to take into account nonlinearities and nonmonotonicities in response of CV risk to a risk factor. For example, the risk of CV event though largely decreases with decreasing lipid levels, but, beyond a certain lower limit, this trend reverses, however, one of the most popular CV risk prediction regression model has only a linear term representing lipids [37]. Further none of these models account for the historical state of a patients risk factor which is known to play a crucial role in determining risk of CV events.

Predictive model have typically been constructed using observations from patients followed up on a regular basis who make up carefully controlled epidemiological cohorts, which are logistically and financially expensive to maintain on a large scale. As a result such data are usually restrictive in terms of the population they represent and its size. EHR data, though not of as high quality as observations from epidemiological cohort are much more widely available because they are comprised of laboratory and vital measurements, physician diagnosis, procedures and prescriptions which are recorded by almost all health care providers. As a result, EHR data is much larger than data from cohorts, more heterogeneous and have a better coverage of the target demographic.

Though there are several machine learning techniques that can be used to construct risk predictive models, they cannot be used easily to learn from EHR data because EHR data often have significant missing observations which are not random, further the outcome of a significant fraction of the population is not observed, i.e. they are right censored, such a dataset can introduce significant biases in the prediction model if that are not properly accounted for. Further, most current techniques usually construct models using only a small fraction of possible risk indicators and usually do not take into account the historical observations of patients. Such models may not be accurate because the observations that are discarded are often good predictors of CV risk for different medically relevant subgroups of population. For example the history of blood pressure measurement is an important risk predictor of patients whose blood pressure fluctuates over time.

In this thesis, we demonstrate how to construct probabilistic graphical models that can make can make use of widely available EHR data. These models can handle censored data, missing data, use historical data to model legacy effect of risk factors on the risk of CV events. In addition to these we have also come up with a scheme that allows these models to make use of nontraditional risk indicators to construct a data driven probabilistic model.

1.1 Electronic health record data.

Although the scale and complexity of EHR data provide great opportunities to improve cardiovascular risk prediction for contemporary patient populations, these data also pose serious challenges to those seeking to mine it, especially in comparison to data collected from epidemiological cohorts.

EHR data are purely observational in the sense that they record individuals' encounters with the health care system, however regular or irregular those encounters may be. Without a structured schedule for clinic visits, laboratory measurements may be obtained sporadically or simply be unavailable. Further, even when individuals do make regular clinical visits, clinical practice guidelines may recommend against obtaining certain measurements within particularly low-risk populations. For instance, it is uncommon for individuals under the age of 40 to obtain cholesterol measurements as part of a routine checkup. Several statistical techniques exist for imputing missing data, but the scale and complexity of EMR data make many of these unappealing.

EHR data comprehensively describe diverse, contemporaneous patient populations. In contrast to a cohort study, patients do not have to satisfy eligibility criteria to appear in the database. Though in some cases it may make sense to restrict an EHR data-based study to a narrowly defined sub-population (e.g., individuals with diabetes), our goal is to develop methods which make maximal use of the available data. Hence, the advanced risk prediction methods should be capable of using the data to automatically identify non-linear relationships between risk factors and CV events as well as sub-populations where the effects of those factors differ.

The enormous sample size afforded by EHR data is simultaneously a blessing and a curse. While there are ample data to support flexible nonparametric risk models, the complexity of these models is restricted by the ability to train them on hundreds of thousands or millions of records in a reasonable amount of time.

Our motivating data set is drawn from 10 years of EHR data derived from an HMO. However, not all subjects were enrolled in the HMO for the full 10-year period, either because their first enrollment came after the beginning of this period or because they

terminated insurance coverage (e.g., due to a change in employer) during the period. While constructing predictive models, we seek to estimate the risk of cardiovascular events occurring over a fixed time horizon (5 years). But the occurrence of cardiovascular events is not recorded in the EMR or in claims data during times when subjects are not enrolled, and a substantial fraction of subjects (71% in our data) do not have enough follow-up data in the EMR to ascertain if they experienced a cardiovascular event over a 5-year period. In the language of statistical survival analysis, such subjects are said to be *right-censored*.

1.2 Machine learning techniques for EHR data

Many data mining techniques have been developed for the related tasks of classification and class probability estimation for a binary outcome. These methods have been shown to yield a performance that is superior to traditional regression-based techniques in healthcare-related classification problems [95, 27], and are more adept at handling the challenges of EHR data. However, they are not designed for use in circumstances where outcomes may be censored.

Fully supervised machine learning methods assume that the outcome is known for all subjects, but in our setting the binary outcome (whether or not a patient experiences a cardiovascular event within a fixed time horizon) is undetermined for subjects who are censored. Simplistic techniques to deal with this issue, such as discarding censored observations (see, e.g., [62, 93, 16]) or treating them as zeroes (non-events), are known to induce bias in the estimation of class probabilities [57], making typical fully supervised classification approaches unsuitable. For example, [98] 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. Furthermore, censored observations are not truly 'unlabeled' since they carry useful partial information about the outcome (i.e., that the event outcome did not occur before the subject was censored).

There has been an increasing interest to adapt machine learning techniques to this type of censored, time-to-event data [70]. Inspired by the growing need and opportunities for machine learning approaches to complex healthcare data-driven problems, we propose a general-purpose extension of Bayesian networks using inverse probability of censoring weights (IPCW). The resulting technique properly accounts for censoring while retaining the features which make this flexible machine learning approach appealing in the context of EHR data.

Bayesian networks are especially well-suited to handle the intricacies of risk prediction using EHD. Compared to support vector machines or neural networks, Bayesian networks have a clear edge in interpretability, which is important to the end-users of these prediction models in the healthcare domain (e.g., physicians and clinical researchers). The Bayesian network framework we adopt is computationally efficient and handles missing data naturally and efficiently, eliminating the need for electronic health data sets to be "pre-imputed" to produce complete data.

Bayesian networks can efficiently represent static systems, however for dynamic systems such as speech recognition systems, robotics, etc which change over time and produce sequential data, Dynamic Bayesian networks are a more optimal representation. A Dynamic Bayesian network consists of variables that indicate the state of the dynamic system at a given point of time. A DBN imposes restrictions on the dependency of the states, i.e., any given state depends only on the state preceding it. In medicine, Dynamic Bayesian networks (DBN) are frequently used to model conditions such as progression of ventilator-associated pneumonia [22], blood glucose variation over time to control insulin administration [6], outcome of a particular treatment for patients with congenital cardiac anomaly [78] or in prediction outbreaks of influenza based on historical records of influenza related hospitalizations among other conditions [90].We use DBNs to model the effect of a sequence of observations (of the risk factors) on the risk CV events. An overwhelming majority of CV risk prediction models have focused on a single measurements of risk factors [92], however, some recent studies have demonstrated that examining the trajectories of risk factors such as SBP leads to better prediction of CV events [103] because these trajectories are known to affect the risk

of CV events[3, 4, 75]. Dynamic Bayesian networks provides us a technique to model this dynamic data, and is yet, open to easy interpretation, can capture the potentially complex relationship between risk factor history and clinical risk, is able handle missing data in a principled fashion and, further be easily adapted for censored observation.

1.3 Dissertation organization

This thesis is composed of three parts. In the first part we demonstrate how a Bayesian network can be used to predict risk in a right censored data set. In the second part we use a Dynamic Bayesian network to incorporate historical patient information that is available in EHD into the predictive graphical model to significantly improve predictions of risk Finally we will develop a variable and structure selection scheme to choose covariates/structures that best predict risk of CV events and test if a data-driven approach to variable selection can lead to better predictions than variables chosen based on expert knowledge.

Part 1: Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. Traditional regression based predictive models are frequently constructed using carefully selected epidemiological cohorts which may not be representative of the population of interest, further more these techniques may not be able to resolve nonlinear relationships between covariates and CV risk. Constructing machine learning based predictive models from EHR data helps predict CV risk that is representative of the population of interest and, can efficiently model complex nonlinear relationships between covariates and CV risk. However, EHR data have several limitations such as missing observations, lack of followup (censoring) and observational biases. As a result, predictive models constructed using machine learning techniques which cannot handle missing observations requires data to be imputed, thus leaving the performance of the prediction model dependent of the imputation technique, and those, that cannot handle censored data (where the outcome of a given set of risk factors is not observed) produce models that have severe calibration errors. To overcome the challenges posed by censored data, we

adjust for right censoring using Inverse Probability of Censoring Weights (IPCW). This technique weighs observations by their probability of being censored which is determined from the duration the patient is enrolled. While this technique can be applied to most machine learning algorithms such as Bayesian networks, logistic regression and decision trees resulting in these techniques performing at least as well as regression based survival models; machine learning techniques such as logistic regression and decision trees still require imputation to make predictions using data with missing observations. Bayesian networks can natively work with missing data by marginalizing out the missing covariates, therefore its predictions are independent of any imputation technique, thus, Bayesian networks along with IPCW is an ideal candidate for constructing predictive models from EHR data. In our implementation of Bayesian Network for censored data we use mixtures of Gaussians to model the distribution of covariates. However, this approach is likely to over fitting if too many mixtures are used. To prevent over fitting we make use of a weighted model averaging technique which results in models with better calibration and discrimination compared to models with adhoc choices of number of mixtures, further such models retain enough flexibility to model non- monotonic relationship of covariates to risk; a feature difficult to reproduce in COX proportional hazards model without prior knowledge of the non-monotonicities.

Part 2: A dynamic Bayesian network for modeling legacy and treatment effects on risk of cardiovascular events. EHR consists of sequence of observations; existing risk prediction algorithms make use of a snapshot of the data to predict risk of events. However, it is well documented that the effect of uncontrolled risk factors such as, blood pressure and blood sugar persists well after they have been brought under control either by medication or by life style changes (legacy effect), as a result, traditional predictive models using a snapshot of the patient state fail to capture the effect of previously elevated (or reduced) risk factors prior to the baseline period. A naive technique to make use of historical data to better predict CV risk consists of averaging the historical risk factors. While such an approach apparently reduces missingness at the baseline and produces better predictions compared to models using a snapshot at only the baseline, it is far from optimal use of the historical data available in EHR besides,

further, such naive techniques ignore variation of risk factors over time which are known to contribute to risk of CV events. A more principled approach for taking in to account the historical observations is, by using dynamic Bayesian networks. We partition the observation period (3years) into time slices each of which are represented by a step in the dynamic Bayesian network. The use of dynamic Bayesian network allows us to easily marginalize out unobserved variables thus making for efficient inferences, in addition, the structures used in a Dynamic Bayesian network follows naturally from the fact that risk factors are dependent on their respective previous state. In this study we show that dynamic Bayesian networks constructed using 3 years of historical observations substantially outperforms regular Bayesian models constructed with data from a single time slice and is significantly better then both the naive approach to including history and, other dynamic Bayesian models constructed with less than three years of history. The improvement is particularly evident in medically relevant cases where the risk indicators such as blood pressure or treatments vary substantially over time. In addition to these medically relevant groups, we try to discover risk based subgroups where the two prediction models vary significantly using a decision tree constructed using the risk covariates and the relative difference of predicted risks. Such risk based subgroups help us to discover situations where the legacy effects play an important role in the risk of CV events, additionally, such partitioning of the population helps us develop a model selection strategy where models more suitable for a particular subgroup is used to predict CV risk for that group, thus, allowing us to optimize risk predictions across the entire population

Part 3: A data driven approach to optimize Bayesian networks for EHR data The choice of risk factors used to construct the predictive models and the causal relationship that drive the structure of both the Bayesian and the dynamic Bayesian networks are based on either information from experts in the field or from the domain literature. However, such expert knowledge and literature may not exhaustive. Given that EHR data consists several measurements of patient state in addition to potentially numerous diagnoses and procedures performed on the patient prior to the baseline which may have an impact on the risk of CV events; it is important to develop a data driven

approach to select variables and to determine structure of the Bayesian network. In this study we will use a hill climbing strategy followed by simulated annealing to determine an optimal structure for the Bayesian networks that make use of all informative diagnoses and measurements in the EHR. In this study, we will examine if such a data-driven approach to construct a predictive model can outperform an expert driven approach.

Chapter 2

Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data

2.1 Overview

Models for predicting the risk of cardiovascular events based on individual patient characteristics are important tools for managing patient care. Most current and commonly used risk prediction models have been built from carefully selected epidemiological cohorts. However, the homogeneity and limited size of such cohorts restrict the predictive power and generalizability of these risk models to other populations. Electronic health data (EHD) from large health care systems provide access to data on large, heterogeneous, and contemporaneous patient populations. The unique features and challenges of EHD, including missing risk factor information, non-linear relationships between risk

factors and cardiovascular event outcomes, and differing effects from different patient subgroups, demand novel machine learning approaches to risk model development. In this paper, we present a machine learning approach based on Bayesian networks trained on EHD to predict the probability of having a cardiovascular event within five years. In such data, event status may be unknown for some individuals, as the event time is right-censored due to disenrollment and incomplete follow-up. Since many traditional data mining methods are not well-suited for such data, we describe how to modify both modeling and assessment techniques to account for censored observation times. We show that our approach can lead to better predictive performance than the Cox proportional hazards model (i.e., a regression-based approach commonly used for censored, time-to-event data) or a Bayesian network with *ad hoc* approaches to right-censoring. Our techniques are motivated by and illustrated on data from a large U.S. Midwestern health care system.

The work in this chapter is published in [8] and includes contributions from Julian Wolfson, David Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Mohamed Elidrisi, Paul Johnson, and Dr Patrick O'Connor. The data was provided to us by Patrick and Gabriela. They helped us understand the intricacies of the data and helped us process the data till it was ready to be consumed by the model. Patrick provided us with medical insights that helped us design the model. Julian and David helped provide ideas that lead to the development of this model. Julian, David and Gediminas also helped writing this manuscript. This work was supervised by Paul and Gediminas,

2.2 Background

In the United States, myocardial infarctions (MI) and strokes are the first and fourth leading causes of death, and, in addition to contributing significantly to mortality, these conditions account for substantial morbidity with treatment costing more than $300 billion annually [42]. Clinical risk prediction scores or algorithms remain important tools for managing patient care and improving outcomes in the population. Broadly speaking, risk prediction scores can raise awareness of the substantial burden of cardiovascular

disease (CVD) and risk factors associated with developing CVD [67]. In the clinical setting, accurate personalized cardiovascular risk prediction may identify patients at high risk for experiencing cardiovascular (CV) events (e.g., MI, stroke) so that clinicians may develop an appropriate intervention strategy and patients are motivated to remain adherent to that strategy.

Recent systematic reviews found that there are over 100 risk models produced between 1999 and 2009 [30, 31, 71] including the well-known Framingham [34], SCORE [28], ASSIGN-SCORE [113], QRISK1 [47, 46], QRISK2 [48], PROCAM [7], WHO/ISH, and Reynolds Risk Score [83, 84]. Although there are many risk prediction models for cardiovascular disease, nearly all have been estimated using data from carefully selected epidemiological cohorts. For example, the Framingham risk score is trained on a data set that excludes patients that have had a previous CV event, represents a predominantly Caucasian population, and includes patients from the late 1960s [34]. As a result of estimating the risk of CV events using data from these homogeneous cohorts, existing risk models are likely to only give accurate predictions for patients who are well represented in the training data sets. [25] provide an excellent illustration of the poor performance of the Framingham risk equations when applied to a contemporary population in the United Kingdom. Models constructed from a more diverse cohort are likely to produce more accurate estimates of CVD risk for a wider range of patients seen in the primary care clinic.

One source of clinical data is electronic health data (EHD) collected by a health maintenance organization (HMO). These data consist of electronic medical records (EMRs), insurance claims data, and mortality data obtained from the state government. EHD are increasingly available within the context of large health care systems and capture the characteristics of a heterogeneous population receiving care in a contemporary clinical setting. EHD databases typically include records on hundreds of thousands to millions of individual patients; therefore, a risk prediction model constructed from EHD could yield more accurate and generalizable risk predictions because even relatively specific sub-populations (e.g., patients with multiple comorbidities) are likely to be well-represented in such a large database.

2.2.1 Challenges of electronic health data

Although the scale and complexity of EHD provide great opportunities to improve cardiovascular risk prediction for contemporary patient populations, these data also pose serious challenges to those seeking to mine it, especially in comparison to data collected from epidemiological cohorts. Here, we briefly describe some of these challenges, all of which are present in the data set we describe in Section 2.4.

Missing data: EMR data are purely observational in the sense that they record individuals' encounters with the health care system, however regular or irregular those encounters may be. Without a structured schedule for clinic visits, laboratory measurements may be obtained sporadically or simply be unavailable. Further, even when individuals do make regular clinical visits, clinical practice guidelines may recommend against obtaining certain measurements within particularly low-risk populations. For instance, it is uncommon for individuals under the age of 40 to obtain cholesterol measurements as part of a routine checkup. Several statistical techniques exist for imputing missing data, but the scale and complexity of EMR data make many of these unappealing.

Subgroup heterogeneity: EHD comprehensively describe diverse, contemporaneous patient populations. In contrast to a cohort study, patients do not have to satisfy eligibility criteria to appear in the database. Though in some cases it may make sense to restrict an EHD-based study to a narrowly defined sub-population (e.g., individuals with diabetes), our goal is to develop methods which make maximal use of the available data. Hence, the advanced risk prediction methods should be capable of using the data to automatically identify non-linear relationships between risk factors and CV events as well as sub-populations where the effects of those factors differ.

Sample size: The enormous sample size afforded by EHD is simultaneously a blessing and a curse. While there are ample data to support flexible nonparametric risk models, the complexity of these models is restricted by the ability to train them on hundreds of thousands or millions of records in a reasonable amount of time.

Incomplete follow-up and censoring: Our motivating data set is drawn from 10 years of EHD derived from an HMO. However, not all subjects were enrolled in the HMO for the full 10-year period, either because their first enrollment came after the beginning of this period or because they terminated insurance coverage (e.g., due to a change in employer) during the period. In this paper, we seek to estimate the 5-year risk of cardiovascular events. But the occurrence of cardiovascular events is not recorded in the EMR or in claims data during times when subjects are not enrolled, and a substantial fraction of subjects (71% in our data) do not have enough follow-up data in the EMR to ascertain if they experienced a cardiovascular event over a 5-year period. In the language of statistical survival analysis, such subjects are said to be *right-censored*.

The standard statistical methods to model the relationship between risk factors and time-to-event outcomes are the Cox proportional hazards model [33] and, to a lesser extent, the accelerated failure time model [19]. Although these methods handle censored outcomes, they do not natively address the first three challenges encountered in EHD. In particular, the proportional hazards model assumes that the risk factors have a linear relationship with the log hazard of experiencing a CV event. If the analyst has *a priori* knowledge that this relationship is non-linear or differs in certain sub-groups, she may include non-linear transformations of predictors or interactions between predictors, but this is often based on trial and error. Secondly, standard software for proportional hazards models removes subjects with any missing features which is a large percentage for EHD, and, as noted above, imputation methods may be onerous given the scale and complexity of EHD. Finally, as noted by [57], the proportional hazards model will show improved fit when additional terms are added in the model, and there is no reliable indication when the model has been overfit.

The first three challenges (missing data, subgroup heterogeneity, and data dimensionality) motivate the use of a sophisticated, flexible, and scalable machine learning technique such as a Bayesian network to mine EHD. However, the fourth challenge (incomplete follow-up) is not directly addressed by usual machine learning approaches. In the next section, we discuss the implications of censoring on the application of machine

learning techniques.

2.2.2 Machine learning with censored outcome data

Many data mining techniques have been developed for the related tasks of classification and class probability estimation for a binary outcome. These methods have been shown to yield a performance that is superior to traditional regression-based techniques in healthcare-related classification problems [95, 27], and are more adept at handling the aforementioned challenges of EHD. However, they are not designed for use in circumstances where outcomes may be censored.

Fully supervised machine learning methods assume that the outcome is known for all subjects, but in our setting the binary outcome (whether or not a patient experiences a cardiovascular event within 5 years) is undetermined for subjects who are censored, i.e., who do not experience an event but do not have a full 5 years of follow-up. Simplistic techniques to deal with this issue, such as discarding censored observations (see, e.g., [62, 93, 16]) or treating them as zeroes (non-events), are known to induce bias in the estimation of class probabilities [57], making typical fully supervised classification approaches unsuitable. For example, [98] 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. Furthermore, censored observations are not truly 'unlabeled' since they carry useful partial information about the outcome (i.e., that the event outcome did not occur before the subject was censored).

There has been an increasing interest to adapt machine learning techniques to this type of censored, time-to-event data [70]. As one example, [50] make use of specialized random survival forests to analyze right-censored survival data, where a survival score is associated with every terminal node of the trees in the forest. Inspired by the growing need and opportunities for machine learning approaches to complex healthcare data-driven problems, we propose a general-purpose extension of Bayesian networks using inverse probability of censoring weights (IPCW). The resulting technique properly

accounts for censoring while retaining the features which make this flexible machine learning approach appealing in the context of EHD. We also show that traditional evaluation metrics for classification accuracy (e.g., receiver-operating curve, net reclassification improvement) may be misleading in the presence of censoring, and describe better alternatives.

2.3 Bayesian networks for electronic health data

2.3.1 Why Bayesian networks?

Bayesian networks are especially well-suited to handle the intricacies of risk prediction using EHD. Compared to support vector machines or neural networks, Bayesian networks have a clear edge in interpretability, which is important to the end-users of these prediction models in the healthcare domain (e.g., physicians and clinical researchers). The Bayesian network framework we adopt is computationally efficient and handles missing data naturally and efficiently, eliminating the need for electronic health data sets to be "pre-imputed" to produce complete data.

Because of their interpretability and their ability to aid in reasoning with uncertainty, Bayesian networks have been used extensively in biomedical applications (see [70] for a review). In particular, they have been: used to aid in understanding of disease prognosis and clinical prediction [5, 108, 66, 89, 110, 61]; used to guide the selection of the appropriate treatment [69, 58, 94, 114, 107]; and implemented as part of clinical decision support systems [68, 91].

In spite of the wide applicability in biomedical applications, there is a limited amount of previous work on the application of Bayesian networks to censored outcome data [70]. [115] and [96] 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 are assigned a weight based on the *marginal* probability of experiencing an event between the censoring time and τ, the time the event status will be assessed. This approach, although intuitive, is provably biased and inconsistent because the method to weight each of the replicated observations is based on the marginal probability and does

not properly account for the covariates. [97] 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. 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 [101, 57, 56]. However, these approaches require that the technique to mine the data permit a continuous outcome and are, therefore, not amenable to the Bayesian network approach considered here.

In this paper, we develop an extension of Bayesian networks which accounts for right-censored event indicators using inverse probability of censoring weights (IPCW) [85, 9, 10, 87, 105]. We begin by reviewing a classical Bayesian network approach, operating on a binary indicator of whether or not each subject experienced an event during the follow-up period. Clinical knowledge is used to construct a graphical model relating the relevant risk factors to the probability of an event. We then describe a model-averaging strategy to control model complexity and stabilize risk predictions for small patient subgroups. The application of IPCW to handle censored outcomes in Bayesian networks is discussed before we describe how to extend traditional predictive performance evaluation metrics for the censored outcome setting.

2.3.2 Bayesian networks

Let the features recorded on a patient be represented by a p-dimensional vector $\mathbf{X} = (X_1 \cdots X_p)$ where X_i is the i^{th} risk factor (values for some of the factors could be missing for certain patients). Let $E = 1$ indicate that an event (e.g., a CV event) occurred for a given patient within τ years of the beginning of the follow-up period, and $E = 0$ indicate the absence of such an event in that time frame. Though our ultimate goal is to handle the case where E is unknown for some patients, for now we assume that at least τ years of follow-up is available on each patient so that E is fully observed.

The target of estimation is $P_{E|\mathbf{X}}(e|\mathbf{x})$, the conditional probability that $E = e$ given

the features $\mathbf{x}$ of a particular patient. Using Bayes theorem, one can rewrite this conditional probability of an event as

$$P_{E|\mathbf{X}}(e=1|\mathbf{x}) = \frac{P_{\mathbf{X}|E}(\mathbf{x}|e=1)P_E(e=1)}{\sum_{e\in\{0,1\}} P_{\mathbf{X}|E}(\mathbf{x}|e)P_E(e)}, \tag{2.1}$$

so that the focus is shifted to estimation of the conditional density/probability $P_{\mathbf{X}|E}(\mathbf{x}|e)$ and the probability $P_E(e)$ for $e = 0, 1$. To maintain notational brevity, we use $P_Y(y)$ to denote either the probability that the random variable Y equals y if Y is discrete or the probability density of Y evaluated at y if Y is a continuous random variable. Similarly, $P_{Y|Z}(y|z)$ is the conditional probability/density of the random variable Y evaluated at y given $Z = z$. In general, the dimensionality of the feature space p may be too large to make joint modeling of $P_{\mathbf{X}|E}(\mathbf{x}|e=1)$ feasible. To simplify the joint modeling task, one can represent the joint distributions of $\mathbf{X}|E = e$ using a directed acyclic graph (DAG), i.e., a Bayesian network. 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 [100]:

$$P_{\boldsymbol{X}|E}(\boldsymbol{x}|e) = \prod_{i=1}^{p} P_{X_i|\mathrm{Pa}(X_i),E}\{x_i|\mathrm{Pa}(x_i), e\} \tag{2.2}$$

where $\mathrm{Pa}(X_i)$ are the parents of X_i.

One advantage of the Bayesian network approach is that clinical knowledge can be used to suggest and refine DAG structures. While methods exist to infer the DAG structure from data, in our application we used parsimonious DAGs based on information from the medical literature as well as clinical judgment from the medical experts who collaborated on this research project. The DAG structure used in our predictive models is shown in Figure 2.1 (see the caption of the figure for a brief description) and is explained in greater detail in Section 2.4. The graphical model in Figure 2.1 contains both continuous-valued nodes (which are elliptical in the figure) and discrete-valued nodes (which are rectangular). The network is, therefore, a *hybrid Bayesian network* [73].

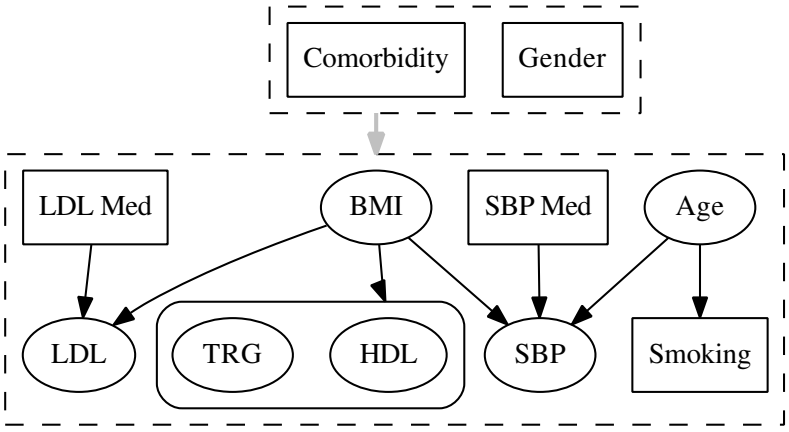


Figure 2.1: The graphical model for our Bayesian network for CV risk prediction. The figure includes the structure of risk factors, conditioned on the CV event status. In particular, nodes represent input variables and edges represent conditional dependencies between the variables. The reader should also assume that our outcome variable (CV Event) is connected to every node in the graph (omitted in the picture for parsimony). Continuous and discrete variables are indicated by elliptical and rectangular nodes, respectively. Nodes in boxes with rounded corners indicate that they are modeled jointly. The nodes are grouped into subgraphs indicated by the dashed boxes. The grey edge between subgraphs indicates an edge from every node in the source subgraph to every node in the destination subgraph or node. The full description of each of the features appears in Section 2.4. *Comorbidity*: Whether or not patient has pre-existing CVD or another comorbidity related to CVD; *Smoke*: current smoking status of patient; *BMI*: body mass index of patient; *SBP*: systolic blood pressure; *SBP Med*: Number of blood pressure medication classes currently prescribed to patient; *TRG*: triglycerides; *HDL*: high density lipoprotein; *LDL*: low density lipoprotein (note that TRG, HDL, and LDL are components of a patient's cholesterol measurement and that HDL and TRG are modeled jointly); *LDL Med*: indicator for whether the patient is on LDL lowering medication.

2.3.3 Learning Bayesian network parameters

As noted above, to estimate $P_{E|\mathbf{X}}(e = 1|\mathbf{x})$ we choose to develop models for $P_{\mathbf{X}|E}(\mathbf{x}|e)$ and $P_E(e)$. When E is observed on all subjects, the maximum likelihood estimate of $P_E(e)$ is straightforward:

$$\hat{P}_E(e) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}[E_j = e], \tag{2.3}$$

where E_j is the CV event status for the j^{th} person, n is the number of people in the training set, and $\mathbb{I}[\cdot]$ is the indicator function.

To evaluate the joint distributions $P_{\boldsymbol{X}|E}(\boldsymbol{x}|e)$, for each feature X_i (or group of features modeled jointly, such as the lipid measures in our DAG) we construct vectors $\boldsymbol{G}_i$, such that $\boldsymbol{G}_i = \{X_i, \mathrm{Pa}(X_i)\}$, and learn the joint density $P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i|e)$ of each of the groups. The conditional density used in Equation (2.2), $P_{X_i|\mathrm{Pa}(X_i),E}\{x_i|\mathrm{Pa}(x_i), e\}$, can be derived from these joint distributions, as described in Section 2.3.6. This approach differs slightly from the conventional strategy, where a regression model is constructed to link the values of the continuous parent nodes of X_i to the mean of X_i. Typically, we would assume that the mean of X_i is linearly related to the values of the parent nodes and we must learn or estimate the parameters in this linear regression model. Our approach provides more modeling flexibility (because we do not have to specify a regression model) without a substantial increase in computation due to our DAG structure, where continuous nodes are arranged in relatively small, mutually independent groupings.

The components in $\boldsymbol{G}_i$ are partitioned into $(\mathbf{Y}_i, \mathbf{Z}_i)$, where $\boldsymbol{Y}_i$ represents the $C_{\boldsymbol{Y}_i}$ continuous risk factors and $\boldsymbol{Z}_i$ the $C_{\boldsymbol{Z}_i}$ discrete risk factors, where $C_{\boldsymbol{G}_i} = C_{\boldsymbol{Y}_i} + C_{\boldsymbol{Z}_i}$. The joint distribution of $\boldsymbol{G}_i$ given E is

$$P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i|e) = P_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i, e) \times P_{\boldsymbol{Z}_i|E}(\boldsymbol{z}_i|e) \tag{2.4}$$

where $P_{\boldsymbol{Z}_i|E}(\boldsymbol{z}_i|e)$ is simply a discrete probability distribution. We can easily estimate this distribution by computing the proportion of observations in each unique state of $\boldsymbol{Z}_i$ separately for $E = 0$ or $E = 1$, which is the (non-parametric) maximum likelihood

estimator of this discrete distribution. That is,

$$\hat{P}_{\boldsymbol{Z}_i|E}(\boldsymbol{z}_i|e) = \frac{\hat{P}_{\boldsymbol{Z}_i,E}(\boldsymbol{z}_i,e)}{\hat{P}_E(e)} = \frac{\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}[\boldsymbol{Z}_{ij}=\boldsymbol{z}_i, E_j=e]}{\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}[E_j=e]}, \tag{2.5}$$

where again j indexes the subject.

For the continuous components, we model the conditional density of $\boldsymbol{Y}_i$ given $\boldsymbol{Z}_i$ and E as a mixture of M multivariate normal densities conditional on the levels of $\mathbf{Z}_i$ and E:

$$P_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i,e) = \sum_{m=1}^{M}\rho_{i,m,z_i,e}\phi_{C_{\boldsymbol{Y}_i}}\left\{\Sigma_{i,m,z_i,e}^{-1/2}(y_i-\mu_{i,m,z_i,e})\right\}, \tag{2.6}$$

where $\phi_k(\cdot)$ is the density function of a k-variate standard normal random variable, $\mu_{i,m,z_i,e}$ and $\Sigma_{i,m,z_i,e}$ are the mean and variance matrix of $\boldsymbol{Y}_i$ given $\boldsymbol{Z}_i=\boldsymbol{z}_i$ and $E=e$ for the m^{th} multivariate normal density in the mixture, and $\rho_{i,m,z_i,e}$ are the mixing parameters where $\sum_{m=1}^{M}\rho_{i,m,z_i,e}=1$. Here we allow $\mu_{i,m,z_i,e}$ and $\Sigma_{i,m,z_i,e}$ to differ for different levels of $\boldsymbol{Z}_i$ and E. These parameters are estimated using the Expectation Maximization algorithm to solve for the maximum likelihood estimates. If each component Z_i^k of $\mathbf{Z}_i$ has Q_i^k possible states, then $|\boldsymbol{Z}_i| = \prod_{k=1}^{C_{\boldsymbol{z}_i}} Q_i^k$. The resulting number of parameters to estimate is $2\times M\times|\boldsymbol{Z}_i|\times\{C_{\boldsymbol{Y}_i}(2+(C_{\boldsymbol{Y}_i}+1)/2)-1\}$.

For a fixed number of mixing components M and given $E=e$ and $\boldsymbol{Z}_i=\boldsymbol{z}_i$, a standard expectation maximization (EM) algorithm [35] is used to solve for the maximum likelihood estimators of the mean, variance, and mixing parameters. The mixing indicators $I_{i,m,z_i,e}$ $(m=1,\ldots,M)$ are viewed as missing data with probabilities/expectations $\rho_{i,m,z_i,e}$. We note that the complete data log-likelihood for the model of $P_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i,e)$ can be written as

$$\begin{aligned}\log L_C(\theta|\boldsymbol{Y}_i,\boldsymbol{Z}_i,E) &= \sum_{j=1}^{n}\sum_{m=1}^{M} I_{i,m,z_i,e}[\log\phi_{C_{\boldsymbol{Y}_i}}\left\{\Sigma_{i,m,z_i,e}^{-1/2}(\boldsymbol{Y}_{ij}-\mu_{i,m,z_i,e})\right\} \\ &+ \log\rho_{i,m,z_i,e}]\mathbb{I}[\boldsymbol{Z}_{ij}=z_i, E_j=e]\end{aligned} \tag{2.7}$$

where θ is the unique parameters of $\mu_{i,m,z_i,e}$, $\Sigma_{i,m,z_i,e}$, and $\rho_{i,m,z_i,e}$, for $m=1,\ldots,M$.

For a current iteration of the parameter estimate $\theta^{(\nu)}$, we can (i) take the expectation of the complete data log-likelihood given the observed data assuming that $\theta^{(\nu)}$ is the

true value of the parameter (E-step), and then (ii) maximize the conditional expectation with respect to θ (M-step) to obtain $\theta^{(\nu+1)}$. In this well-studied Gaussian mixture problem, it is possible to derive explicit update formulas for both mixing and distributional parameters so that the "E-step" and "M-step" are performed simultaneously. In particular,

$$\begin{aligned}
\mu_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} p_{j,m}^{(\nu)} \boldsymbol{Y}_{ij}}{\sum_{j=1}^{n} p_j^{(\nu)}} \\
\Sigma_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} p_{j,m}^{(\nu)} (\boldsymbol{Y}_{ij} - \mu_{i,m,z_i,e}^{(\nu+1)})(\boldsymbol{Y}_{ij} - \mu_{i,m,z_i,e}^{(\nu+1)})^T}{\sum_{j=1}^{n} p_j^{(\nu)}} \\
\rho_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} p_{j,m}^{(\nu)}}{\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_{ij} = z_i, E_j = e]},
\end{aligned} \tag{2.8}$$

where $p_j^{(\nu)} = \mathrm{E}_{\theta^{(\nu)}}(I_{i,m,z_i,e} | \boldsymbol{Y}_{ij}, E_j, \boldsymbol{Z}_{ij}) \times \mathbb{I}[\boldsymbol{Z}_{ij} = z_i, E_j = e]$. Additional details of the algorithm can be found in Bilmes [14].

The preceding discussion has assumed a fixed M. To control overfitting, one could consider M to be a tunable parameter and select the number of mixture components using the Bayes Information Criteria (BIC) or some other goodness-of-fit measure. Alternatively, in the following section, we describe a model averaging procedure to combine results across multiple values of M.

2.3.4 Model averaging

Increasing the number of mixture components M in our models can lead to overfitting. This is even more likely when we try to model distributions that are represented by small subgroups of patients. For example, in our proposed Bayesian network (see Figure 2.1), we must estimate the conditional distribution of age given CV event status, comorbidity status, and gender. However, the number of females with pre-existing conditions and experiencing a CV event is small relative to the size of the cohort; therefore, a highly flexible model for the conditional age distribution may lead to unreliable estimates of this distribution. We propose to reduce overfitting by implementing a model averaging approach using a weighted average of conditional probability densities from models with

varying number of mixing parameters. The weights are derived from the BIC, which penalizes model complexity and is an approximation to the posterior probability of the model [51]. Although model averaging has been used to average over different network structures in Bayesian networks [102], to the best of our knowledge it has not been previously used to average over models of varying complexity within a single network structure.

To implement the model averaging, we take $T = 40$ bootstrap samples from our training data. For each bootstrap sample, we fit models as described in Section 2.3.3 with $M = 1, \ldots, Q$ mixture components. That is, for a given bootstrap sample t and number of mixture components M, we estimate the parameters from Equation (2.6) using the standard EM algorithm described above. The weights $w_{i,z_i,e}^{(M)}$ for each level of model complexity M for modeling the conditional distribution $\boldsymbol{Y}_i$ given $\boldsymbol{Z}_i = \boldsymbol{z}_i$ and $E = e$ are computed as follows:

$$w_{i,z_i,e}^{(M)} = \frac{\exp(-0.5B_{i,z_i,e}^{(M)})}{\sum_{M=1}^{Q} \exp(-0.5B_{i,z_i,e}^{(M)})}, \tag{2.9}$$

where $B_{i,z_i,e}^{(M)} = \frac{1}{T}\sum_{j=1}^{T} B_{i,z_i,e,j}^{(M)}$, and $B_{i,z_i,e,j}^{(M)}$ is the BIC of the model for the conditional distribution $\boldsymbol{Y}_i$ given $\boldsymbol{Z}_i = \boldsymbol{z}_i$ and $E = e$ fit on the j^{th} bootstrap sample with M mixtures. The BIC is computed as $-2\log(\mathbf{L}) + k\log(n)$ where $\boldsymbol{L}$ is the log-likelihood evaluated at the maximum likelihood parameter estimates, n is the number of subjects included in the training set, and k the number of (free) parameters included in the model.

The probability density of an ensemble is estimated by finding the weighted mean of the probability of each mixture model in the ensemble. For example,

$$\hat{P}_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i, e) = \sum_{1\leq M\leq Q} w_{i,z_i,e}^{(M)} \left(\sum_{1\leq m\leq M} \left[\frac{1}{T} \sum_{1\leq t\leq T} \hat{\Psi}_{i,m,z_i,e}^{t} \right] \right) \tag{2.10}$$

where $\hat{\Psi}_{i,m,z_i,e}^{t}$ is the summand of Equation (2.6) with parameters estimated from bootstrap sample t. Generally, the maximum number of mixing parameters Q should be large enough so that the largest values of M ($M = 1, \ldots, Q$) do not receive high weight

$w_{i,z_i,e}^{(M)}$. In our real data analysis of cardiovascular risk, we take $Q = 4$.

2.3.5 Extension to censored time-to-event data

Our development thus far has assumed that the presence or absence of a CV event E is fully observed for all patients in the data set, but this is unlikely to be true when using information from real-world EHD. In our application, once a patient leaves the health system or the study ends, their health state (i.e., features of the risk prediction model) and event history are no longer recorded in the EMR. If the patient's follow-up ends prior to τ, then their event status at τ is unknown and their event indicator is said to be *right-censored*. In this section, we show how it is possible to use a Bayesian network to predict the risk of a CV event in τ years when the event status at τ years is right-censored. To establish notation (which is standard in the statistical literature), define T as the time between the beginning of the follow-up period and a CVD event, and define C as the time between the beginning of the follow-up period and disenrollment or the end of the data capture period. We observe $V = \min(T, C)$ and $\delta = \mathbb{I}(T < C)$, the indicator for whether or not a CV event occurs. If $\delta = 0$, the subject's event time is right-censored. We can only ascertain the value of E either if $\delta = 1$, or if $\delta = 0$ and $V > \tau$; in other words, the value of E is only known if $\min(T, \tau) < C$.

As mentioned earlier, one naive approach to handling the subjects for whom we cannot ascertain the value of E would be to exclude them from our training data set or to set $E = 0$, but both approaches would lead to biased estimators of the CV risk. Instead, we propose to adjust for right-censoring using an inverse probability of censoring weighting (IPCW) approach. This approach to handling censored event times assumes that the censoring time C is independent of the CV event time T and all features $\boldsymbol{X}$. In our study, most patients are censored due to the end of the study or because they disenroll from the HMO due to a change in employment, reasons unrelated to their health status (i.e., $\boldsymbol{X}$ and T). Let $G(t) = P(C > t)$ be the probability that the censoring time is greater than t. We can estimate $G(t) = P(C > t)$, using the Kaplan-Meier estimator of the survival distribution (i.e., 1 minus the cumulative distribution function) of the censoring times. The Kaplan-Meier estimator [53] of the censoring

process is given by

$$\hat{G}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i^*}{n_i} \right) \tag{2.11}$$

where d_i^* is the number of subjects who were censored at time t_i, and n_i is the number of subjects "at risk" for censoring (i.e., not previously censored or experiencing a CV event) at time t_i. Unlike other *ad hoc* approaches to handling censored observations, the Kaplan-Meier estimator is a consistent estimator of G [53]. 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 for computing the Kaplan-Meier estimator of events times can be used to estimate G by setting the "event" indicators to $\delta_i^* = 1 - \delta_i$.

Having computed $\hat{G}$, for each patient j we define a inverse probability of censoring weight (IPCW):

$$\omega_j = \begin{cases} \frac{1}{\hat{G}(\min(V_j,\tau))} & \text{if } \min(T_j,\tau) < C_j \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

To fit the Bayesian network using IPCW, estimation of the parameters in $P_E(e)$ and $P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i, |e)$ is carried out using weighted maximum likelihood, where the contribution of the j^{th} subject to the likelihood is weighted by ω_j. In particular, Equations (2.3) and (2.5) become

$$\hat{P}_E(e) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}[E_j = e]\omega_j \tag{2.13}$$

$$\hat{P}_{\boldsymbol{Z}_i|E}(\boldsymbol{z}_i|e) = \frac{\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e]\omega_j}{\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}[E_j = e]\omega_j}, \tag{2.14}$$

respectively. One can show that the IPCW estimator of $P_E(e = 1)$ is the Kaplan-Meier event probability (i.e., the Kaplan-Meier estimator is an IPCW estimator).

To estimate the parameters in $P_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i, e)$ we use a weighted EM algorithm where the contribution of each subject j is weighted by ω_j. In particular, the complete data log-likelihood for the EM algorithm becomes

$$\log L_C(\theta|\boldsymbol{Y}_i, \boldsymbol{Z}_i, E) = \sum_{j=1}^{n} \omega_j \left(\sum_{m=1}^{M} I_{i,m,z_i,e} [\log \phi_{C_{\boldsymbol{Y}_i}} \left\{ \Sigma_{i,m,z_i,e}^{-1/2} (Y_{ij} - \mu_{i,m,z_i,e}) \right\} + \log \rho_{i,m,z_i,e}] \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e] \right). \quad (2.15)$$

The update formulas for the parameter estimates previously given in Equation (2.8) now become

$$\begin{aligned} \mu_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} \omega_j p_{j,m}^{(\nu)} \boldsymbol{Y}_{ij}}{\sum_{j=1}^{n} p_j^{(\nu)} \omega_j} \\ \Sigma_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} \omega_j p_{j,m}^{(\nu)} (\boldsymbol{Y}_{ij} - \mu_{i,m,z_i,e}^{(\nu+1)})(\boldsymbol{Y}_{ij} - \mu_{i,m,z_i,e}^{(\nu+1)})^T}{\sum_{j=1}^{n} p_j^{(\nu)} \omega_j} \\ \rho_{i,m,z_i,e}^{(\nu+1)} &= \frac{\sum_{j=1}^{n} \omega_j p_{j,m}^{(\nu)}}{\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e] \omega_j}, \end{aligned} \quad (2.16)$$

where $p_j^{(\nu)} = E_{\theta^{(\nu)}}(I_{i,m,z_i,e}|\boldsymbol{Y}_{ij}, E_j, \boldsymbol{Z}_{ij}) \times \mathbb{I}[\boldsymbol{Z}_{ij} = z_i, E_j = e]$, as before. Further details of the weighted EM algorithm are provided in Fraley et al. [38]. In the case that we use model averaging with bootstrap subsamples rather than a fixed model complexity, the weights are computed on the full training sample, and are not re-estimated for each bootstrap sample. Many software packages implementing the EM algorithm (e.g., Matlab, R) allow weights to be provided as arguments to the EM function, making the IPCW Bayesian network approach straightforward to implement.

In the IPCW approach, only those patients for whom we can determine E contribute to the analysis, but they are reweighted to accurately "represent" the patients who were censored prior to τ and were, therefore, omitted from the analysis. For example, patients that have a longer time to event are more likely to be censored (G is smaller) and, hence, receive larger weights. Note that for subjects with $E = 0$ (and $V > \tau$) the weights for all individuals are $1/\hat{G}(\tau)$, so the maximum likelihood estimators for $P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i|e = 0)$ are the same as in the unweighted analysis.

We briefly discuss why inverse probability of censoring weighting results in consistent estimators (i.e., unbiased in large samples and converging in probability to the true

parameter) for the parameters in the Bayesian network using an illustrative example. In particular, if we fully observed the event status E on all subjects and wished to estimate $\mu_{i,z_i,e}$, a reasonable estimator would be the sample average of $\boldsymbol{Y}_i$ among all subjects in the training set with $\boldsymbol{Z}_i = \boldsymbol{z}_i$ and $E = e$:

$$\hat{\mu}_{i,z_i,e} = \frac{\sum_{j=1}^{n} \boldsymbol{Y}_{ij} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e]}{\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e]} = \frac{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{Y}_{ij} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e]}{\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_{ij} = \boldsymbol{z}_i, E_j = e]} \tag{2.17}$$

Since $E(\boldsymbol{Y}_{ij}\mathbb{I}[\boldsymbol{Z_{ij}} = \boldsymbol{z_i}, E_j = e]) = \mu_{i,z_i,e} P_{\boldsymbol{Z_i},E}(\boldsymbol{z_i}, e)$ and $E(\mathbb{I}[\boldsymbol{Z_{ij}} = \boldsymbol{z_i}, E_j = e]) = P_{\boldsymbol{Z_i},E}(\boldsymbol{z_i}, e)$, by the weak law of large numbers $\hat{\mu}_{i,z_i,e}$ converges in probability to $\mu_{i,z_i,e}$.

When E is not observed on all subjects, an IPCW estimator is given by

$$\hat{\mu}^{IPCW}_{i,z_i,e} = \frac{\sum_{j=1}^{n} \boldsymbol{Y}_{ij} \mathbb{I}[\boldsymbol{Z}_i j = \boldsymbol{z}_i, E_j = e]\omega_j}{\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_i j = \boldsymbol{z}_i, E_j = e]\omega_j} = \frac{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{Y}_{ij} \mathbb{I}[\boldsymbol{Z}_i j = \boldsymbol{z}_i, E_j = e]\omega_j}{\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}[\boldsymbol{Z}_i j = \boldsymbol{z}_i, E_j = e]\omega_j} \tag{2.18}$$

Note that we can rewrite ω_j as $\mathbb{I}[\min(T_j, \tau) < C_j]/\hat{G}\{\min(T_j, \tau)\}$ which approaches $\mathbb{I}[\min(T_j, \tau) < C_j]/G\{\min(T_j, \tau)\}$ for large samples. Combining this fact with the assumption that C is independent of T and $\boldsymbol{X}$, it can be shown that the numerator and denominator in Equation (2.18) converge in probability to the same quantity as in the non-IPCW case, so that $\hat{\mu}^{IPCW}_{i,z_i,e}$ converges in proability to $\mu_{i,z_i,e}$. The reader is encouraged to consult the cited references above, especially Bang and Tsiatis [9], for a rigorous theoretical treatment. The IPCW parameter estimators and parameter updates in the EM algorithm, i.e., Equations (2.13), (2.14), and (2.16), take the form of weighted sample averages. Therefore, the results of the illustrative example considered here are generalizeable to all IPCW estimators.

2.3.6 Obtaining predictions

To obtain the prediction $P_{E|\boldsymbol{X}}(e = 1|\boldsymbol{x})$, we need to evaluate the right-hand side of Equation (2.1), where $P_{\boldsymbol{X}|E}(\boldsymbol{x}|e)$ is given by the product in Equation (2.2) resulting in:

$$P_{E|\boldsymbol{X}}(e = 1|\boldsymbol{x}) = \frac{\prod_{i=1}^{p} P_{X_i|\mathrm{Pa}(X_i),E}(x_i|\mathrm{Pa}(x_i), e = 1) P_E(e = 1)}{\sum_{e\in\{0,1\}} \prod_{i=1}^{p} P_{X_i|\mathrm{Pa}(X_i),E}(x_i|\mathrm{Pa}(x_i), e) P_E(e)} \tag{2.19}$$

As noted above, the joint modeling described in Section 2.3.3 yields conditional distributions via the expression

$$\begin{aligned} P_{X_i|\mathrm{Pa}(X_i),E}(x_i|\mathrm{Pa}(x_i),e) &= \frac{P_{X_i,\mathrm{Pa}(X_i)|E}(x_i,\mathrm{Pa}(x_i)|e)}{P_{\mathrm{Pa}(X_i)|E}(\mathrm{Pa}(x_i)|e)} \\ &= \frac{P_{X_i,\mathrm{Pa}(X_i)|E}(x_i,\mathrm{Pa}(x_i)|e)}{\int P_{X_i,\mathrm{Pa}(X_i)|E}(x_i,\mathrm{Pa}(x_i)|e)dx_i} \\ &= \frac{P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i|e)}{\int P_{\boldsymbol{G}_i|E}(\boldsymbol{g}_i|e)dx_i}. \end{aligned} \tag{2.20}$$

Estimates for $P_{X_i,\mathrm{Pa}(X_i)|E}(x_i,\mathrm{Pa}(x_i)|e)$, which we discussed in the preceding sections, may be substituted in Equation (2.20) to obtain an estimate for $P_{X_i|\mathrm{Pa}(X_i),E}(x_i|\mathrm{Pa}(x_i),e)$ which in turn may be substituted into Equation (2.19) to obtain $P_{E|\boldsymbol{X}}(e=1|\boldsymbol{x})$.

Missing features

Within EHD, it is relatively common for attributes to be unmeasured on certain patients; for example, cholesterol measurements are available on only a small fraction of patients under the age of 40. One of the advantages of Bayesian networks is that we can still obtain predictions for subjects in the validation set with incomplete features; we can also use information on subjects in the training set to learn parameters in the Bayesian network without having to impute the missing covariate values. To obtain an estimate of $P_{E|\boldsymbol{X}}(e=1|x)$ if X_i is missing, the corresponding product term from Equation (2.1) is dropped. When any attribute from $\mathrm{Pa}(X_i)$ is missing (say A_i), we substitute $P_{X_i|Pa(X_i)\setminus A_i,E}$ for the corresponding term in Equation (2.1). That is, we implicitly marginalize over A_i by computing $P_{\boldsymbol{G}_i\setminus A_i|E}(\boldsymbol{g}_i \setminus a_i|e) = \int_a P_{\boldsymbol{G}_i|A_i,E}(\boldsymbol{g}_i|a_i = a,e)dP_{A_i}(a|E)$ where $P_{A_i}(a|E)$ is estimated from subjects with A_i observed and then derive the conditional probability from the joint distribution. In our data, information regarding medications, gender, smoking, and comorbidity is never missing; therefore, we are concerned only about marginalizing the continuous part of the joint distribution.

2.3.7 Summary of the modeling process

In the previous sections we provided a detailed discussion of the components of the proposed approach. To explain how all of these components fit together, Algorithm 1 provides a high-level overview of the entire modeling process.

Algorithm 1 High-level overview of the proposed approach

Input:

Graphical structure of the probabilistic relationships (edges) between input features (nodes)

Training dataset (each record consists of input values, follow-up time, and an event indicator)

Output:

Function for estimating conditional probability of the CV event given the input values

1: Estimate survival distribution of the censoring times using Kaplan-Meier estimator // Equation (2.11)
2: **For each** subject (each record in a dataset):
3: Compute the inverse probability of censoring weight using the distribution from Line 1 // Equation (2.12)
4: **For each** node X_i in the graphical model: // Figure (2.1)
5: Identify the set of parent nodes of X_i, i.e., $\text{Pa}(X_i)$
6: Model the conditional joint distribution of X_i and $\text{Pa}(X_i)$ given event status as follows:
7: Let $\boldsymbol{G}_i = \{X_i, \text{Pa}(X_i)\}$
8: Partition $\boldsymbol{G}_i$ into continuous and discrete features
9: For the set of discrete features of $\boldsymbol{G}_i$:
10: Compute the IPCW estimates of the conditional probability given event status
using the IPCW weights from Line 3 // Equation (2.14)
11: For the set of continuous features of $\boldsymbol{G}_i$:

12: **For each** distinct state of discrete features of $\boldsymbol{G}_i$ and event status:
13: **For each** number of multivariate Gaussian mixtures (vary from 1 to 4):
14: Create 40 bootstrap samples of the training data
15: **For each** bootstrap sample:
16: Use IPCW EM algorithm to estimate parameters in multivariate normal mixture distribution using the IPCW weights from Line 3 // Equations (2.15-2.16)
17: Calculate the BIC values for the current model // Section (2.3.4)
18: Compute the average model BIC value across 40 bootstrap samples from individual BIC values from Line 17 // Section (2.3.4)
19: Find the model complexity weight for each number of mixtures using the average model BIC values from Line 18 // Equation (2.9)
20: Use the model complexity weights from Line 19 and parameter estimates from bootstrap samples from Line 16 to find conditional joint density of continuous features given discrete states of $\boldsymbol{G}_i$ and event status // Equation (2.10)
21: Obtain conditional joint distribution of $\boldsymbol{G}_i$ given event status by multiplying discrete probability estimates from Line 10 and conditional density of continuous features from Line 20 // Equation (2.4)
22: Derive conditional distribution of node X_i given parents $\text{Pa}(X_i)$ and event status using the conditional joint distribution of $\boldsymbol{G}_i$ from Line 21 // Equation (2.20)
23: Estimate the marginal probability of CV event using IPCWs from Line 3 // Equation (2.13)
24: Derive conditional probability estimation function of CV event given input features using (a) conditional distribution of nodes given their parents and event status from Line 22 and (b) the marginal probability of CV event from Line 23 // Equation (2.19)

2.3.8 Stability, scalability, and model complexity

The computational complexity of our model is $O(T \times Q \times \max_i S_{\boldsymbol{G_i}} \times R \times |\boldsymbol{X}|)$, where T is the number of bootstrap samples, Q is the maximum number of mixtures that we try to fit, $S_{\boldsymbol{G}_i}$ is the number of possible discrete states in the set of variables $\boldsymbol{G}_i$, $|\boldsymbol{X}|$ is the number of variables in the graphical network, and R is the worst-case time-complexity of the EM algorithm. T and Q are tuning parameters which do not depend on sample size. Under the assumption that a fixed constant number of iterations will be sufficient for the convergence of EM algorithm (a common assumption in many EM software implementations), R is linear in the sample size in our case. $S_{\boldsymbol{G_i}}$ and $|\boldsymbol{X}|$ are determined by the network structure, which in our case is fixed in advance. Hence, these terms do not depend on sample size but may dominate the complexity; in particular, $S_{\boldsymbol{G_i}}$ is the product of the number of different states for each discrete variable appearing in the joint distribution. The graphical model used in this paper involves a fairly small number of discrete features (gender, comorbidity, blood pressure medications, LDL medications, and smoking) each with a relatively small number of states, and $\max_i S_{\boldsymbol{G_i}} = 16$.

In terms of practical runtime considerations, an implementation of this algorithm using Matlab on a standard desktop machine required approximately 25 minutes to train the model on a dataset comprised of around 130,000 records. Obtaining the predictions of approximately 42,000 records required around 10 seconds. This further demonstrates the ability of the proposed approach to scale to large real-world healthcare applications.

The proposed method takes several steps to ensure stability. In general, the stability of the EM algorithm used to compute the model parameters depends on the data. There could be scenarios where the size of the training set is insufficient to learn the parameters using the EM algorithm for certain combinations of discrete features. We circumvent this problem by dynamically adjusting the number of multivariate Gaussian mixtures that we use to learn the distribution using model averaging. That way, more complex models which may have difficulty converging in small training sets and are likely to overfit the data receive little weight in the model averaging. In addition, the bootstraping component of the model averaging technique minimizes the effect of the

EM not converging (within the fixed number of iterations) for some of the samples.

2.3.9 Performance evaluation metrics for censored data

The challenge of machine learning for censored data extends beyond improving existing methods to handle censoring. As we discuss in this section, the usual performance metrics applied to classification and prediction problems can be misleading when outcomes are subject to censoring. Here we present generalizations 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. For the statistics described in this section, closed-form expressions for the asymptotic variance are usually not available, and hence standard errors for the calculation of confidence intervals and p-values are obtained by bootstrap resampling.

Calibration

For standard binary classification problems, calibration is commonly assessed by ranking the predicted class probabilities for the test set, binning the ranked predictions (e.g., by decile), and comparing the mean (or median) predicted class probability in each bin to the empirical class probability of the instances in that bin. We adopt a similar approach here, except that the empirical probability of experiencing an event prior to time τ within each bin is computed via the Kaplan-Meier estimator to properly account for censoring. Calibration plots compare predicted and Kaplan-Meier probabilities of experiencing an event before τ within bins defined by ranges of predicted probabilities. We compute the calibration statistic

$$K = \sum_{j=1}^{B} \frac{(\bar{p}_j - p_j^{KM})^2}{var(p_j^{KM})} \tag{2.21}$$

$$var(p_j^{KM}) = var(S_j(\tau)) = S_j(\tau)^2 \sum_{t_i<\tau} \frac{d_{ij}}{n_{ij} - d_{ij}} \tag{2.22}$$

where B is the number of bins, $\bar{p}_j$ is the average of predicted probabilities in bin j, p_j^{KM} is the Kaplan-Meier estimate of experiencing an event before τ, $S_j(\tau)$ is its corresponding survival rate ($= 1 - p_j^{KM}$), $var(p_j^{KM})$ is its variance calculated using Greenwood's formula [44] applied to the data in bin j, d_{ij} is the number of events occurring at time t_i in bin j, and n_{ij} are the number of people "at risk" for an event at time t_i (i.e., not censored and not experiencing an event before time t_i). K is analogous to the χ^2 statistic for assessing the calibration of logistic models suggested by [49, 64].

Concordance index

The area under the ROC curve (AUC) is a widely used summary measure of predictive model performance. However, for the same reasons that standard classification techniques fail on censored outcomes, the AUC can also be misleading. Hence, we employ a generalization of the AUC, the concordance index (C-index) for censored data.

As described in [45], the C-index adapted for censoring considers 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 a CV event, or one subject is observed to experience a CV event before the other subject is censored. Pairs in which both subjects are censored or in which the censoring time of one precedes the failure of the other do not contribute to this metric. Let $\hat{P}_{E_j|\boldsymbol{X}_j}(e_j = 1|\boldsymbol{x}_j)$ be the estimated probability that the j^{th} subject experiences an event within τ years. Then the C-index adapted for censoring is given by

$$C_{cens}(\tau) = \frac{\sum_{k \neq j} \delta_k \mathbb{I}[V_k < V_j] \mathbb{I}[\hat{P}_{E_k|\boldsymbol{X}_k}(e_k = 1|\boldsymbol{x}_k) < \hat{P}_{E_j|\boldsymbol{X}_j}(e_j = 1|\boldsymbol{x}_j)]}{\sum_{k \neq j} \delta_k \mathbb{I}[V_k < V_j]} \tag{2.23}$$

Note that the only pairs which contribute to $C_{cens}(\tau)$ are those where one subject experiences an event prior to τ and the other is known not to have experienced an event before the first subject. In the absence of censoring, the C-index and AUC coincide.

Net Reclassification Improvement

The C-index may be inadequate to distinguish between models that differ in relatively modest but clinically important ways [79]. Therefore, in addition to the C-index, we

also evaluate the performance of our models using the Net Reclassification Improvement (NRI) metric [79], which allows the incorporation of relevant domain knowledge into the performance evaluation process. As with the other metrics presented in this section, the NRI must be adapted to handle censored outcomes.

The NRI compares the number of "wins" for two models among discordant predictions. It has been argued that NRI is a particularly relevant measure of comparison between models in the clinical domain, where it is often more important to discriminate between lower and higher risk patients than to estimate their risk precisely. Briefly, 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:

$$\text{NRI}(M_1, M_2) = \frac{E^{\uparrow}_{M_1} - E^{\uparrow}_{M_2}}{n_E} + \frac{\bar{E}^{\downarrow}_{M_1} - \bar{E}^{\downarrow}_{M_2}}{n_{\bar{E}}} \tag{2.24}$$

Here $E^{\uparrow}_{M_1}$ is the number of individuals 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). Also, 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. While NRIs can theoretically range from -1 to 1, in the risk prediction setting they do not typically exceed the range of -0.25 to 0.25. For example, [29] calculated the effects of omitting various risk factors from the Reynolds Risk Score model for prediction of 10-year cardiovascular risk. The estimated NRIs ranged from -0.195 (omitting age) to -0.032 (omitting total cholesterol or parental history of MI), and all were statistically significant at the 0.05 level. In our application, we defined three risk strata based on clinically relevant cutoffs for the risk of experiencing a cardiovascular event within 5

years: 0-5% (low risk), 5-10% (moderate risk), and $> 10\%$ (high risk); risk predictions for an individual were considered discordant between two models if the predictions fell in different ranges.

The predictions that are reclassified from one risk category to another can be represented in two tables, one for people having events (Table 2.1b) and one without (Table 2.1a), where the risk has been categorized into 3 levels: "high", "medium", and "low" using the aforementioned risk cutoffs. The entry in Table 2.1a, Nn_{ij}, represents the number of people without events who were classified as belonging to risk category i by model M_2 that were reclassified as belonging to risk category j by model M_1. Similarly, the entry in Table 2.1b, Ne_{ij}, represents the number of people with events who were classified as belonging to risk category i by model M_2 that were reclassified as belonging to risk category j by model M_1.

For a person with events, a "win" for model M_1 over M_2 means M_1 predicts a higher risk category for the person than M_2. Based on our reclassification table, $E^{\uparrow}_{M_1} = Ne_{mh} + Ne_{lh} + Ne_{lm}$. Similarly, the number of "wins" M_2 has over M_1 for people with events is: $E^{\uparrow}_{M_2} = Ne_{hm} + Ne_{hl} + Ne_{ml}$. For people without events, we can write the write the corresponding "wins" from Table 2.1a as follows: $\bar{E}^{\downarrow}_{M_1} = Nn_{hm} + Nn_{hl} + Nn_{ml}$ and $\bar{E}^{\downarrow}_{M_2} = Nn_{mh} + Nn_{lh} + Nn_{lm}$. n_E and $n_{\bar{E}}$ are the sum of the entries in Tables 2.1b and 2.1a respectively. From the values obtained using the reclassification tables, the Net Reclassification improvement is evaluated using Equation (2.24).

Table 2.1: Net Reclassification Improvement computation tables for comparing performance of model M_1 vs. model M_2.

(a) Reclassification table for people who do not have CV events.

		Model M_1		
	Risk	high(h)	medium(m)	low(l)
Model M_2	high(h)	Nn_{hh}	Nn_{hm}	Nn_{hl}
	medium(m)	Nn_{mh}	Nn_{mm}	Nn_{ml}
	low(l)	Nn_{lh}	Nn_{lm}	Nn_{ll}

(b) Reclassification table for people who have CV events.

		Model M_1		
	Risk	high(h)	medium(m)	low(l)
Model M_2	high(h)	Ne_{hh}	Ne_{hm}	Ne_{hl}
	medium(m)	Ne_{mh}	Ne_{mm}	Ne_{ml}
	low(l)	Ne_{lh}	Ne_{lm}	Ne_{ll}

The NRI statistic cannot be applied directly to data where the outcome status of some subjects is unknown. In our setting, omitting subjects with less than five years of follow-up (or treating them as non-events) will result in biased estimates of the NRI. To evaluate risk reclassification on our test data which are subject to censoring, we use a "censoring-adjusted" NRI (cNRI) due to [80] which takes the form:

$$\text{cNRI}(M_1, M_2) = \frac{E^{*,\uparrow}_{M_1} - E^{*,\uparrow}_{M_2}}{n^*_E} + \frac{\bar{E}^{*,\downarrow}_{M_1} - \bar{E}^{*,\downarrow}_{M_2}}{n^*_{\bar{E}}} \tag{2.25}$$

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 (2.24), but correspond to expected number of subjects in each category, with the expectations computed using the Kaplan-Meier estimator to account for censoring. As with the usual NRI, a positive cNRI(M_1, M_2) means better reclassification performance for model M_1, while a negative cNRI(M_1, M_2) favors model M_2.

2.4 Electronic health data and preprocessing

Our study was conducted utilizing the HMO Research Network Virtual Data Warehouse (HMORN VDW) from a healthcare system from the Midwestern United States. The VDW stores data in standardized data structures including insurance enrollment, demographics, pharmaceutical dispensing, utilization, vital signs, laboratory, census and death records. These data are obtained from both the EMR and insurance claims. 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 EMR of this system. Furthermore, once the patient-member disenrolls from the HMO, the patient no longer has any information 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. This initial selection identified 448,306 subjects.

The goal of our analysis is to develop accurate prediction models of CV risk for patients seen in the primary care clinics of the HMO. Such models will help inform physicians and patients of appropriate courses of action to take in order to reduce the patient's likelihood of experiencing a CV event.

In many epidemiological cohorts, patients are screened at a single or small number of visits, where measurements on all risk factors are collected, and then followed from the final screening visit to determine if and when they experienced a CV event. However, in clinical practice, patients may have relevant risk factors recorded over a series of visits to the physician. Therefore, we divided the EHD on any given patient-member into: (i) a baseline period, where we ascertained the risk factors, 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. This approach balances the competing goals of having a long baseline period so that we capture data on as many features as possible and a long follow-up period to reduce censoring.

The follow-up period for a patient begins at the end of the baseline period and continues until either the patient experiences a CV event (defined below), or the patient disenrolls from the HMO for more than 90 days, or the study ends (2011), whichever comes first.

2.4.1 Inclusion and exclusion criteria

To ensure that we had sufficient time to collect baseline risk factors on subjects, we restricted the analysis 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 HMO).

We further restrict our study population to include 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, and greater than 18 years of age at the end of the baseline period. These inclusion criteria were implemented because we wanted to predict CV risk among those patients 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 counselled about their CV risk) were not of interest in this analysis.

Finally, in this study we included only non-diabetic patients. Diabetic patients represent a highly specialized population and would benefit from a specialized risk prediction model that is targeted specifically to them, such as the UKPDS model [24], which was beyond the scope for this specific paper. The inclusion and exclusion criteria

are summarized in Figure 2.2, and the distribution of the follow-up periods for the resulting analysis cohort is given in Figure 2.3.

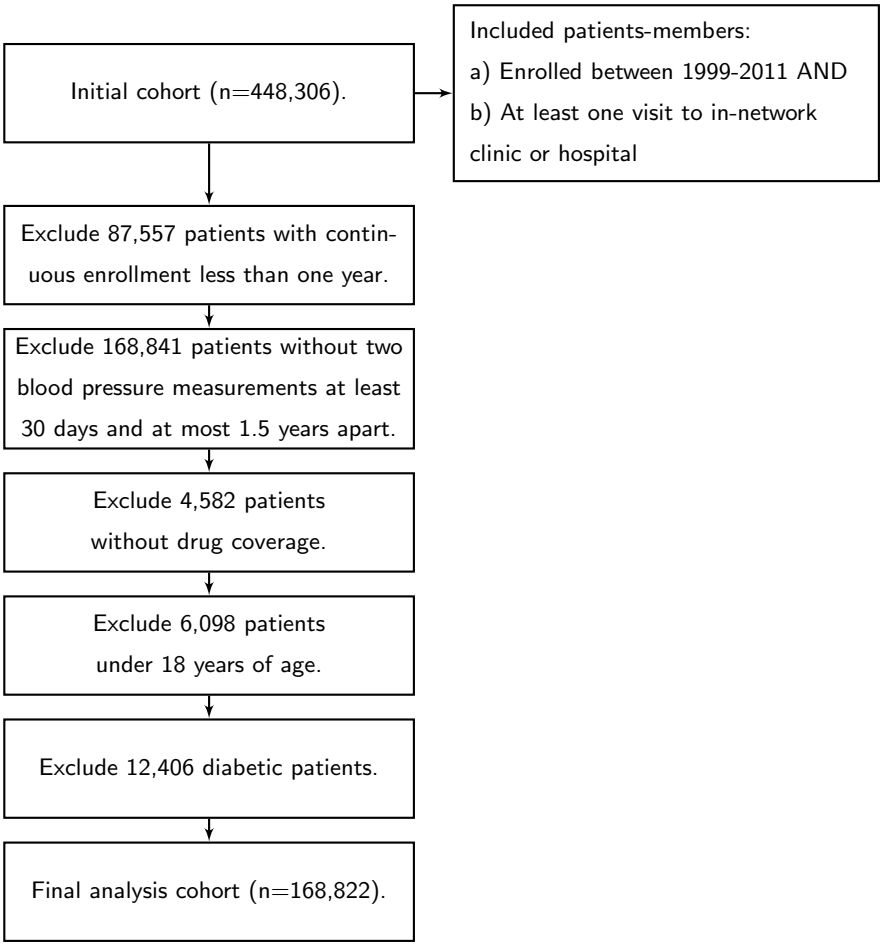


Figure 2.2: Flowchart of inclusion and exclusion criteria for analysis

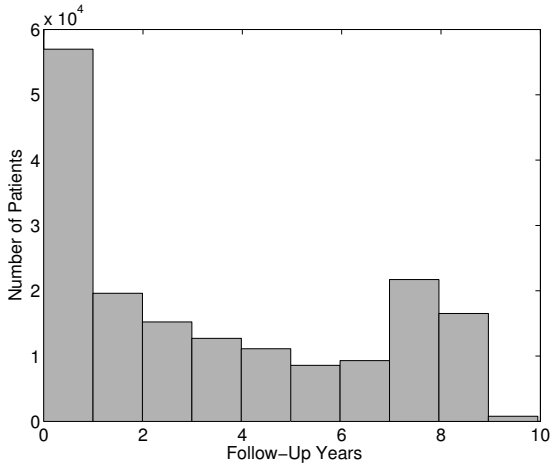


Figure 2.3: Distribution of patient follow-up times, i.e., time from the end of the baseline period until the patient experiences a CV event, the patient disenrolls from the HMO for more than 90 days, or the study ends, in our entire cohort after applying inclusion and exclusion criteria.

2.4.2 Risk factor ascertainment

Risk factors incorporated in the Bayesian network include age, gender, systolic blood pressure (SBP), smoking status, body mass index (BMI), cholesterol-related measurement values (LDL, HDL, TRG), blood pressure and cholesterol medications, as well as indicators of pre-existing cardiovascular disease and other diseases related to CV events (i.e., pre-existing related diagnoses or procedures). The summary statistics for the risk factors are given in Table 2.2. We also provide a brief discussion on how each of the features and the outcome, CV event, was defined based on information in the EMR and claims data.

Table 2.2: Summary measures of the risk factors included in our prediction models in the entire study cohort.

Feature Name	Median (IQR) or N (%)	% Missing	Description
Gender			
Female	102,754 (60.9)	0	
Male	66,068 (39.1)	0	
Age (Years)	43 (31 - 56)	0	Age at the end of the baseline period
SBP (mm Hg)	118 (110 - 127)	0	Mean systolic blood pressure during baseline period
BMI (kg/m^2)	26.7 (23.5 - 31.4)	14	Body mass index
LDL (mg/dL)	115 (94 - 138)	66	Final low density lipoprotein during baseline period
HDL (mg/dL)	47 (39 - 56)	55	Final high density lipoprotein during baseline period
TRG (mg/dL)	106 (76 - 153)	66	Final triglyceride during baseline period
Smoking			Smoking status in EMR
Never or Passive	124,580 (73.8)	0	
Quit	16,655 (9.9)	0	
Current	27,587 (16.3)	0	
Comorbidity			Presence of comborbidities related to cardiovascular disease
Yes	18,031 (10.7)	0	
No	150,791 (89.3)	0	
SBP Meds			Number of SBP medication classes filled during baseline period
0	113,478 (67.2)	0	
1	20,593 (12.2)	0	
2	14,187 (8.4)	0	
3+	20,564 (12.2)	0	
LDL Meds			Number of LDL medication classes filled during baseline period
0	150,049 (88.9)	0	
1+	18,773 (11.1)	0	

Systolic blood pressure (SBP): Calculated as an average of all the blood pressure measurements taken during the baseline period. Blood pressure readings obtained during emergency department visits, urgent care visits, hospital admission, and during procedures (e.g., surgeries) were excluded from consideration because they may be influenced by acute conditions.

Body mass index (BMI): Calculated as a function of patient's height and weight. The height of an individual is the average height measured at any encounter (possibly outside of the baseline period). Because all subjects in the analysis dataset are over 18 years of age, we expect height to remain relatively constant over the follow-up period. The weight is calculated as an average of all weight measurements taken during the baseline period.

Low density lipoprotein (LDL), high density lipoprotein (HDL), and triglycerides (TRG): The most recent laboratory measurements before the end of the baseline period is used for these lipid measures including low density lipoprotein, high density lipoprotein, and triglycerides.

Smoking status: Information about smoking is complicated by the fact that many individual's responses vary considerably over time. In our dataset, there are four categories of smoking history, never smoked, smoking, quit smoking, and passive (i.e., second-hand) smoking. In our analysis, a person is considered to have never smoked only if they consistently recorded "no smoking" throughout their association with the insurance provider. A person who has recorded at least two "smoking" responses is considered currently smoking. For the purpose of constructing the model we combine the "passive smoking" and "no smoking" categories.

SBP and LDL medications: In our model, SBP medications are represented as the number of different medication categories a person is prescribed at the end of the baseline period. In particular, SBP medication categories included: alpha-blockers, beta-blockers, calcium-blockers, ace-inhibitors, angiotensin, vasodialator, and diuretics. LDL medication represents an indicator for whether or not a patient is taking any LDL lowering medications, such as statins and fibrates, at the end of the baseline period. For our analysis, we ignore information regarding the specific drug dosages because it is difficult to make comparisons between doses from different variants of the same drug.

Comorbidities: Comorbidities represent serious pre-existing conditions (diseases) and previously occurred CV events or procedures (surgeries). The existence of comorbidities significantly elevate the risk of having a CV event in the future. In our study, we included the presence of any of the following diagnoses (including a diagnosis of a

"history" of these conditions) at any point before the end of the baseline period: chronic kidney disease, coronary heart disease, cardiovascular disease, peripheral artery disease, atrial fibrillation, congestive heart failure, myocardial infarction (MI), and stroke. As we discuss below, the diagnosis may be part of the EMR or contained as part of an insurance claim. The Bayesian model that we consider treats comorbidities as a binary variable.

CV Event (outcome variable): Events are 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. This information is 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 consider a patient to have experienced a CV event if the cause of death listed on the death certificate included MI or stroke.

We note that the diagnosis and procedure codes used to define CV events and comorbidities may be part of the EMR or part of claims data (to justify the insurance claim). The implication of this is that patients do not have to seek care at an in-network hospital following a CV event to infer that the patient had a CV event. The total number of first CV events recorded within the follow-up period is 5,410; the Kaplan-Meier 5-year event rate for the entire analysis cohort is 4.53%.

2.4.3 Determining the structure of the Bayesian network

Figure 2.1 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. For example, SBP is known to depend on age, BMI, and the number of blood pressure medication classes prescribed [86]; therefore, age, BMI, and blood pressure medication nodes are represented as parents of SBP nodes in the DAG. People with higher BMI are known to have higher LDL, triglycerides (TRG) and lower HDL [60]. In addition, lipid medications (statins) reduce LDL levels [63]. Therefore, in this case, BMI is represented as parent of LDL,

HDL, and triglycerides; also, LDL medication is a parent of LDL. Moreover, smoking incidence in the United States is known to vary as a function of age [2]; therefore, age is represented as a parent of smoking.

In some cases, it may be preferable to consider several network structures reflecting plausible relationships between predictor variables. In the process of developing our approach, we considered several candidate network structures; within a set of similar candidates, the particular structure did not have a large impact on prediction performance.

2.5 Models and evaluation metrics

Our goal is to build risk models from EHD and, in the process, address issues typically encountered with EHD, such as right-censored outcomes, non-linear and non-monotonic effects of risk factors on the risk of events, and overfitting in subgroups that are represented by relatively few patients. In the context of estimating the five-year risk of a cardiovascular event using the data described in Section 2.4, we sought to compare our approach which trains the Bayesian network in Figure 2.1 using inverse probability of censoring (Section 2.3.5) and using model averaging (Section 2.3.4) to other more traditional Bayesian network approaches as well as other techniques for modeling censored outcomes. Our approach, which we refer to as Bayes-AC (as it includes both model **a**veraging and **c**ensoring weights), allows for non-linear relationships between the risk factors and outcome, addresses censored outcomes, and protects against overfitting in small subpopulations. In addition to our main approach, we also considered more basic Bayesian models using a single multivariate normal distribution (Bayes-1C) and a mixture of three multivariate normal densities (Bayes-3C) to model $P_{\boldsymbol{Y}_i|\boldsymbol{Z}_i,E}(\boldsymbol{y}_i|\boldsymbol{z}_i,e)$, i.e., no model averaging, but still accounting for censored outcomes.

For comparison, we also train the Bayesian network using an *ad hoc* approach for censored observations in which we excluded patients who did not experience an event and did not have 5-years of follow-up. This *ad hoc* approach to censoring was used both when the number of mixture components was fixed (Bayes-1 and Bayes-3) and with

model averaging (Bayes-A).

In addition to the different variations of the Bayesian network models, we considered a Cox proportional hazards model (COX) because this model is well-known in the medical community and well-suited to work with censored data. A drawback of the proportional hazards model is that it does not automatically allow for non-linear relationships between the risk factors and the log hazard. We chose to parameterize the proportional hazards model using the same parameters as was included in the Framingham risk model [34]. However, [34] do not include patients with prior comorbidities in the training set, so we include an indicator variable for comorbidity status. Additionally, the Framingham model includes an interaction between the log of SBP and blood pressure medication but does not include a main effect for blood pressure medication. We have included the blood pressure main effect in the model. One of the limitations in using the Cox proportional hazards model is that it requires complete predictor data to fit the model and to predict risks; therefore, we have to impute the missing data. We construct linear regression models stratified by gender and comorbidity with higher order terms (squared and cubic terms of the independent variables) to impute the missing data. In our case, age and SBP (which are known for all patients) are used to construct such a linear regression model for imputing BMI. Age, SBP, and the imputed BMI are in turn used to construct 3 different models for imputing LDL, HDL, and triglycerides, respectively.

For completeness, we also considered a regression-based approach which does not account for censored outcomes. Specifically, we considered a logistic regression model of the 5-year CV event status with all the same predictors as the Cox proportional hazards model described above. The same imputation method was used for missing features as described above. However, following the approach used for the Bayes-1, Bayes-3, and Bayes-A models, we excluded patients from the training set who did not experience an event and did not have 5-years of follow-up.

Finally, given the relatively large cohort size, we wanted to investigate whether we could achieve adequate performance using a subset of patients with only complete features and, therefore, avoid employing sophisticated data imputation techniques. We

considered training the Bayes-AC and COX models using only subjects with all covariates measured (i.e., subjects with **n**on-**m**issing data), which we refer to as the Bayes-AC-NM and COX-NM models.

Brief descriptions of the models that we have used in our analysis are given in Table 2.3. The training dataset consisting of 129,428 patients (75% of the entire analysis population) was drawn at random from the cohort. However, as noted above, some of the modeling approaches use only a subset of this training dataset. In particular, those models which exclude patients who did not experience a CV event and did not have 5-years of follow-up were trained on 48,300 subjects. Those models which exclude patients that did not have complete data on all the features were trained on 42,523 subjects. The performance of all models is evaluated based on the risk predictions of the remaining 43,143 patients not included in any training set. The models are compared based on their calibration (as described in Section 2.3.9) and discrimination metrics, i.e., the C-index and cNRI (as described in Sections 2.3.9 and 2.3.9). In addition to the calibration and the discrimination metrics, we also plot the difference between average risk and observed risk across different groups of predicted risk. That is, the training set is partitioned into bins with predicted CV risk between 0-2.5%, 2.5-5%, 5-7.5%, 7.5-10%, 10-15%, 15-20%, and >20%. These bins were based on clinically relevant risk categories suggested by our clinical collaborators (similar categories have been used in prior literature). The observed risk within each bin was computed using the Kaplan-Meier estimator. While the calibration statistic K helps us compare the models using a single numeric metric, these plots provide us with more information regarding the calibration of a model for different risk ranges and the direction of deviation (over- or under-prediction) of the model from the observed risk. A perfectly calibrated model would be indicated by a horizontal line at 0 in these plots.

Table 2.3: Overview of the models considered in the analysis of 5-year cardiovascular event risk.

Model Name	Description
Bayesian network models without accounting for censoring:	
Bayes-1	Basic model, using 1 normal distribution for all continuous variables
Bayes-3	More complex model, using mixture of 3 normal distributions for all continuous variables
Bayes-A	Ensemble approach, using model averaging for models with mixtures of different sizes (1-4 normals)
Standard regression model without accounting for censoring:	
Logistic	Logistic regression model
Bayesian network models accounting for censoring:	
Bayes-1C	Bayes-1 with inverse probability of censoring weights
Bayes-3C	Bayes-3 with inverse probability of censoring weights
Bayes-AC	Bayes-A with inverse probability of censoring weights
Standard regression-based baseline approach for censored data:	
COX	Cox proportional hazards model
Models trained on cohort with complete data	
Bayes-AC-NM	Bayes-AC trained on the subset of the population with complete features (i.e., no missing values)
COX-NM	COX trained on the subset of the population with complete features (i.e., no missing values)

2.6 Results

The calibration and discrimination of the models, evaluated on the hold-out test set, is summarized in Table 2.4. Specific comparisons between our approach, which incorporated model averaging and IPCW to train a Bayesian network, and other more standard modeling approaches are described in detail below.

Table 2.4: Calibration and discrimination of the models described in Table 2.3, 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 (lower values indicate better calibration); *C-index*: concordance index (higher values indicate better discrimination; standard errors for all models were approximately 0.0005); *cNRI*: net reclassification improvement for censored outcomes compared to COX model (positive values indicate an improvement over COX).

	Predicted event rate (%) (Observed rate: 4.53%)	Calibration statistic K (Standard error)	C-index (SE: 0.0005)	cNRI (%) (compared to COX)
Bayes-1	6.05	137.1 (2.6)	0.8843	9.40
Bayes-3	6.05	163.1 (2.8)	0.8790	9.08
Bayes-A	6.29	215.6 (4.1)	0.8835	10.17
Logistic	6.87	427.1 (5.8)	0.8752	8.66
Bayes-1C	4.40	7.3 (0.5)	0.8845	5.50
Bayes-3C	4.88	29.5 (1.0)	0.8773	3.06
Bayes-AC	4.46	3.8 (0.5)	0.8859	5.23
COX	4.61	11.7 (0.9)	0.8795	-
Bayes-AC-NM	3.94	57.6 (1.3)	0.8710	-3.81
COX-NM	3.95	39.2 (1.1)	0.8788	-4.61

2.6.1 Censoring-unaware predictive models

The Bayesian network models that do not incorporate inverse probability of censoring weighting and the logistic regression model discard subjects who do not have a follow-up time of at least 5 years and do not experience an event, but still include subjects who have events even if they have a follow-up time of less than five years. As a result, all of these models (Bayes-1, Bayes-3, Bayes-A, and Logistic) over-predict risk and, hence, are poorly calibrated (see Table 2.4 and also Figure 2.4). Although the observed 5-year event rate for the test set was 4.53%, the average risk predicted by Bayes-1, Bayes-3, Bayes-A, and Logistic was 6.05%, 6.05%, 6.29%, and 6.87%, respectively.

While the censoring-unaware Bayesian network models are poorly calibrated, their

discrimination capability (0.8843, 0.8790, and 0.8835) is at least as high as that of the COX model (0.8795) and statistically significantly higher than that of the logistic regression model (0.8752). The higher discrimination of the Bayesian network models compared to the regression model can be attributed to the fact that these models can capture dependencies and correlations between risk factors (such as SBP and age) and can fit the nonlinearities in the data better than the regression models without higher order terms and interaction effects.

We note briefly that, while these models show a significant improvement in the discrimination compared to the standard Cox proportional hazards models using cNRI, it is well known that the NRI statistic can be misleading when one of the models is badly calibrated, such as the models that do not account for censoring [81]. The cNRI statistic that we consider weights the percentage of "wins" for those experiencing an event equally to those not experiencing an event. The Bayesian network models that do not account for censoring are heavily biased toward predicting higher risk; therefore, it produces a significant number of "wins" for people who have events. Therefore, it is most meaningful to use the cNRI metric only when both models being compared are reasonably well-calibrated.

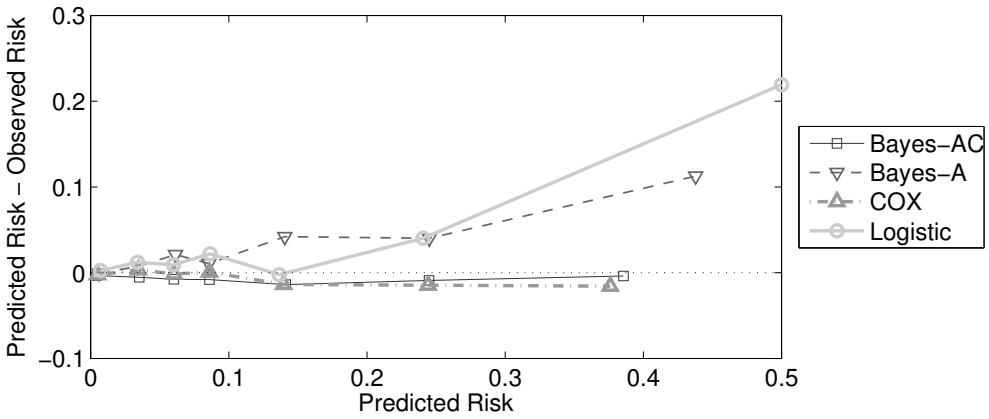


Figure 2.4: Calibration of Bayesian network models both with and without IPCW, COX, and logistic regression model on the hold-out test set.

2.6.2 Bayesian networks with model averaging

The Bayes-3C model, in which all continuous features are modeled as a mixture of three multivariate normal distributions, consistently performs worse than the modeling approach with less flexibility (Bayes-1C) across all measures of calibration and discrimination. In contrast, considering a model-averaged estimate of CV risk, which averages over model complexity using a data-driven approach, led to significant improvement in both calibration and discrimination (see Figure 2.5). Specifically, the averaged model (Bayes-AC) predicts the risk of a CV event better in groups where few events occur. For example, in the sub-population of patients that are not on blood pressure medications which has an event rate of only 0.81% (as compared to an average event rate of 4.53%), the Bayes-AC model has C-index of 0.81 which is significantly higher than the C-indices 0.78 and 0.77 for the Bayes-1C and Bayes-3C models, respectively. In summary, complex model strategies are able to extract more structure from the data but must be implemented intelligently. The benefit of model averaging is that it allows the analyst to consider more complex models, but these more complex models (with greater number of mixture components) are only given substantial weight in Equation (2.9) if there is enough improvement in the model fit to justify models with more parameters. As mentioned earlier, we use the BIC as the metric to balance model complexity and parsimony.

2.6.3 Bayesian networks accounting for censoring using IPCW

Comparing the parsimonious Bayesian networks and model-averaged Bayesian networks that do and do not incorporate inverse probability of censoring weights (Bayes-1 versus Bayes-1C and Bayes-A versus Bayes-AC), there is a dramatic improvement in the calibration of those models which properly account for censoring (Figure 2.4). In these models, the average predicted event rates 4.40% (Bayes-1C) and 4.46% (Bayes-AC) is much closer to the observed event rate in the test set than the equivalent models that are trained ignoring censoring, as shown in Table 2.4. Their calibration statistic is also much closer to zero, the value we would expect for a perfectly calibrated model. Both

Bayes-1C and Bayes-AC also show slightly improved discrimination compared to the respective models that do not account for censoring (C-index 0.8845 versus 0.8843 and 0.8859 versus 0.8835, respectively). Considering the cNRI, both of these models are significantly better in terms of discrimination than the standard Cox proportional hazards model but, unlike the Bayes-1 and Bayes-A models, do not sacrifice calibration to improve net reclassification.

Overall, the Bayes-AC model, which uses inverse probability of censoring weights and model averaging, results in the best performance of all our Bayesian network models. Unlike the Bayes-1 model (a more traditional Bayesian network model), our approach properly accounts for censoring leading to improved calibration and properly protects against over-fitting leading to improved discrimination.

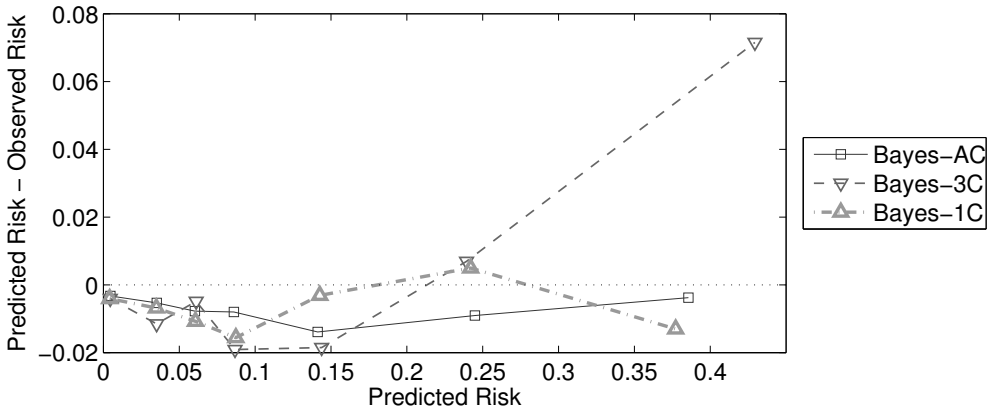


Figure 2.5: Calibration of censoring-aware Bayesian network models of different model complexities on the hold-out test set.

2.6.4 Censoring-aware Bayesian networks versus traditional survival analysis

As noted previously, the Cox proportional hazards model (COX) is the standard approach for data in which the outcome may be right-censored. As compared to this

model, our Bayesian network model (Bayes-AC) provides improvement across all calibration and discrimination metrics considered for this predictive task. Overall, the calibration statistic for the Bayes-AC model (3.84) was much closer to 0, (i.e., the statistic for a perfectly calibrated model) in the test set than the one demonstrated by COX (11.66). The C-index of Bayes-AC is 0.8859 (versus 0.8795 for COX), and the cNRI of Bayes-AC compared to COX is 5.88%, both of which reflect significant improvement in the discrimination. To put the reclassification performance in context, this is more than the improvement that can be obtained from adding the total cholesterol into the COX risk prediction model [29]. Our improvement in cNRI can be attributed to the fact that the Bayes-AC model classifies substantially more people with events into a higher risk category (Tables 2.5b and 2.5c) as well as larger number of people without events into a lower risk category (Tables 2.5a and 2.5c) than COX. Overall, Bayes-AC reclassifies a higher fraction of people in the correct direction.

Table 2.5: Censoring-adjusted Net Reclassification Improvement computation tables for comparing performance of Bayes-AC vs. COX on the hold-out test set.

(a) Reclassification table for patients without CV events.

		Bayes-AC		
	Risk	>15%	5-15%	0-5%
COX	>15%	2,130	540	67
	5-15%	542	2,132	2,299
	0-5%	95	1,299	31,165

(b) Reclassification table for patients with CV events.

		Bayes-AC		
	Risk	>15%	5-15%	0-5%
COX	>15%	1,013	81	5
	5-15%	133	245	88
	0-5%	9	88	298

(c) cNRI computation based on the reclassification tables.

Censoring-adjusted NRI calculation	**Quantity**
Number of patients without events predicted in a lower risk category by Bayes-AC ($\bar{E}^{*,\downarrow}_{M_1}$)	2,906
Number of patients without events predicted in a lower risk category by COX ($\bar{E}^{*,\downarrow}_{M_2}$)	1,936
Total number of patients without events ($n^*_{\bar{E}}$)	36,983
Number of patients with events predicted in a higher risk category by Bayes-AC ($E^{*,\uparrow}_{M_1}$)	230
Number of patients with events predicted in a higher risk category by COX ($E^{*,\uparrow}_{M_2}$)	174
Total number of patients with events (n^*_E)	1,962
Risk reclassification improvement for patients without events	0.0282
Risk reclassification improvement for patients with events	0.0241
Total net risk reclassification improvement	0.0523

Although the overall improvement in calibration was not very large (Figure 2.4),

the Bayesian network model allows for greater flexibility to model CV risk in certain subgroups. For example, it is generally understood that the CV risk rises with increased blood pressure [99]. In most cases, our data supports this assertion. However for males who are already on a blood pressure medication, the relationship between the risk and blood pressure is not strictly increasing, as we see in Figure 2.6. The risk increases with decreasing SBP for SBP below 130 mmHg. This observation is interesting because physicians typically treat SBP down to 130 mmHg for people whose blood pressure is not controlled. Blood pressure being treated below 130 mmHg may indicate an underlying disease which is probably evident to a physician but is not captured by the risk factors that we have trained our model on. The underlying disease increases the risk for this group of patients.

The proportional hazards model that we are using forces a linear relationship between the log hazard and the log of SBP and, thus, ends up modeling the risk well only for people whose blood pressure is relatively elevated. This is evident in Figure 2.6 which shows the Bayes-AC model with better fit with respect to the observed SBP and CV risk relationship. In addition, the C-index of the Bayes-AC model (0.611) is significantly higher than that of the COX model (0.517) for predictions in this subgroup.

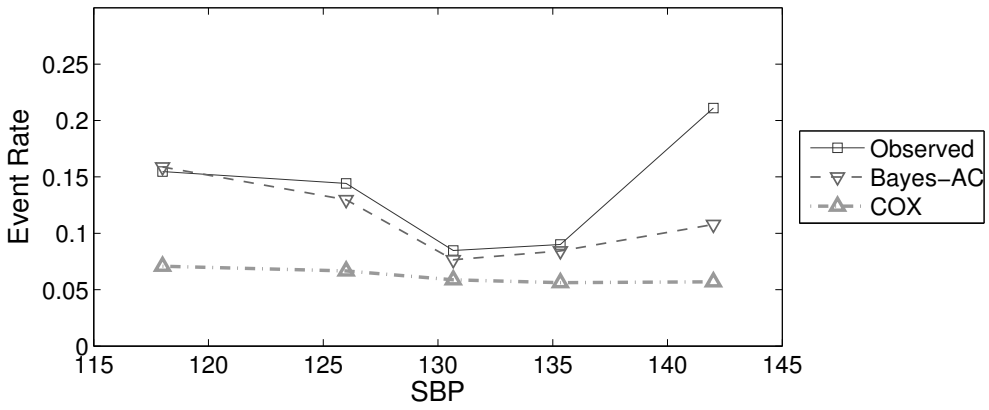


Figure 2.6: Relationship between systolic blood pressure and CV risk for a subgroup of males between ages 40 and 55 who are on SBP medication.

2.6.5 Including patients with missing data

Of the 168,822 patients included in our study, 56,698 patients have observations for all the risk factors. Given that our model uses of only 11 risk factors, the number of people with complete observations should be enough to train the model; however, the missing covariates are not randomly distributed across all the patients in our cohort. As a result, excluding people who have missing data leads to biased estimation of the parameters in the model and poor calibration, as illustrated in Figure 2.7. For example, the mean age of our cohort is 44.7, while constructing a cohort of people with complete observations leads to a set with mean age of 52.9. As a result, the latter cohort has significantly higher risk than the general population. The missingness, in addition to being influenced by age, is also determined by physicians' decisions regarding the overall well-being of the patient. This implies that a cohort with complete observations is likely to be less healthy than the general population. In fact, after controlling for age and gender, the two most important risk factors, the 5-year CV event rate for the cohort with no missing data is significantly higher than for the general population. Males in the 40-45 year age group in the general population have a 5-year CV event rate of 1.94%, while a similar group in the cohort with complete observations has a 5-year event rate of 2.85%.

Models constructed using the non-missing cohort (Bayes-AC-NM and COX-NM) under-predict the CV risk when applied to people who have part of their observations missing. For these patients, Bayes-AC-NM predicts an average 5-year risk of 3.29% as compared to the observed 5-year risk of 3.80%. However, as expected, for people who have complete observations, both the Bayes-AC-NM and COX-NM evaluate an average 5-year risk very close to the observed risk (Bayes-AC-NM: 5.82%, COX-NM: 6.14%, observed: 6.02%). The miscalibration of the models is to be expected because the Bayes-AC-NM and COX-NM have no support for people with missing data and, as we have shown, these people are likely prognostically different from people with complete measurements.

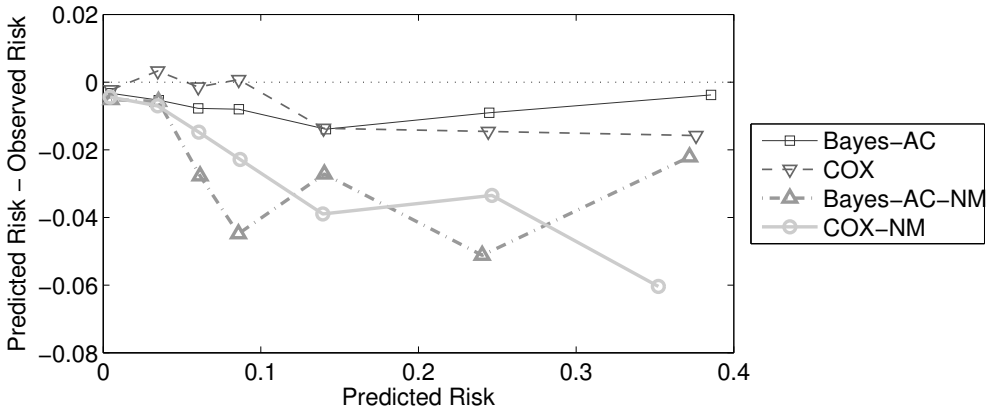


Figure 2.7: Calibration comparison between the Cox proportional hazards model and censoring-aware Bayesian network with model averaging trained using the complete cohort versus the same models trained using the non-missing cohort.

2.7 Discussion

2.7.1 Summary and advantages of proposed method

This paper focuses on the application of a machine learning approach to risk prediction using EHD, when event times may be censored due to unequal individual follow-up. Traditional statistical models for event-time data with censored observations are well-developed, but typically less flexible than established machine learning techniques for classification. On the other hand, most classification techniques do not handle censoring, as they assume that labels in the training data are fully observed (or in the case of semi-supervised classifiers, observed on a random subsample of individuals). Our proposed technique combines features from both of these approaches, using inverse probability weighting to extend the Bayesian network technique for censored event data. Although we apply our approach to a Bayesian network, IPCW can be extended to other machine learning classifiers. Furthermore, we have implemented a novel model averaging approach to control the flexibility in the Bayesian network model. We have shown that the model averaging approach leads to improved prediction of CV events and allows us

to extract more structure from the data without overfitting.

In addition to offering both modeling flexibility and statistical validity, our technique seamlessly handles missing data (which is common in EHD) and offers the opportunity to combine findings from the medical literature with clinical judgment to shape the model. The advantages of our method are illustrated by applying it to a large electronic health database. We show that: (1) ignoring censoring when performing classification results in grossly miscalibrated predictions, (2) the Bayesian network learns non-linear predictor-outcome relationships better than standard proportional hazards regression models of the type typically used to construct cardiovascular risk models from longitudinal studies with censored event data. We also emphasize the importance of using appropriate criteria for assessing the performance of predictive models for censored data. Commonly employed metrics for calibration and discrimination require event indicators that are fully known, and, therefore, can be misleading in the presence of censoring. We present alternate metrics which are tailored to the censored data setting.

2.7.2 Potential limitations of proposed method

As we noted previously, the use of inverse probability of censoring weighting relies on the assumption that the censoring time is independent of the CV event time. Heuristically, we assume that patients more likely to have a CV event are not more or less likely to disenroll from the health system. We could relax this assumption by modeling the censoring time as a function of the risk factors.

Though there are known techniques for searching across multiple DAG structures in the context of Bayesian networks, we chose to focus on techniques for learning parameters for a given, fixed network. This decision was motivated by the fact that our clinical collaborators desire model interpretability and face validity, which may not be achieved with an automated process determining network topology. Further, flexible methods such as Bayesian networks are prone to overfitting. We control overfitting in our method by implementing BIC model averaging and bootstrapping; in our experiments, the method was not highly sensitive to tuning parameter values which determined the maximum number of mixture components in each model and the number of

bootstrap resamples, provided these were set within reasonable ranges (> 3 and > 10, respectively).

2.7.3 Generalizability and future work

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.

Chapter 3

Incorporating legacy effects in risk prediction models using Dynamic Bayesian networks

3.1 Overview

Risk prediction models for a range of health outcomes are an increasingly important component of clinical decision support systems. But despite mounting evidence that the effects of uncontrolled risk factors may persist well after they have been brought under control either by medication or by lifestyle changes, most risk models use only the current or most recent observations to estimate the risk or probability of experiencing a health outcome. Here, we present an approach to modeling risk using historical risk factor trajectories with Dynamic Bayesian networks, accounting for the fact that the time to the event of interest may be censored via inverse probability of censoring weighting. Using electronic health data from a large Midwestern health insurance provider, we construct Dynamic Bayesian network models to predict the risk of cardiovascular events using three years of historical risk factor data. The Dynamic Bayes approach yields more accurate predictions than other approaches, including proportional hazards regression models and non-Dynamic Bayesian networks constructed with data from a

single time period. We also conclude that the highest prediction accuracy is achieved by summarizing data separately in each of the three years prior to baseline. Notably, this approach outperforms both more aggregated (e.g., averaging across three years) and less aggregated (e.g., summarizing data in six-month intervals) alternatives. The gains in predictive accuracy achieved by incorporating history are most pronounced in medically relevant cases where risk factors change substantially over time.

The work in this chapter is being reviewed for publication in Health Services and Outcomes Research Methodology and includes contributions from Julian Wolfson, David Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Paul Johnson, and Dr Patrick O'Connor. Patrick provided us with medical insights that helped us design the model. Paul, Gediminas, Julian and David helped me develop the model by critiquing my different design decisions. Julian, David and Gediminas also helped writing this manuscript. This work was supervised by Paul and Gediminas,

3.2 Background

Models for predicting clinical risk (i.e., the probability of experiencing a particular event over a given future time period) on the basis of observed risk factors play an important role in clinical decision support systems. Risk models can be used to counsel patients on potential lifestyle changes, help physicians decide between alternative treatment strategies, and allow health systems to quantify and track quality of care. Models exist for a wide variety of clinical outcomes, including cardiovascular disease [30, 31, 71], diabetes [26], sepsis [12], breast cancer [11], and hospital readmission [54] among many others.

The vast majority of risk prediction models assume that only the current (i.e., "baseline") risk factor values are known. In some cases, this is because the data used to estimate the risk model does not contain risk factor information collected longitudinally, or data are collected over an insufficiently long period of time to both simultaneously characterize pre-baseline risk factor values and allow for adequate post-baseline follow-up to characterize risk. In other settings, it is assumed that users of the risk model will not

have longitudinal risk factor information available (e.g., models based on measurements easily taken during a single routine check-up).

However, there is evidence in the biomedical literature that the longitudinal trajectories of risk factor values play a role in determining future risk beyond the current value [112, 75, 21, 65]. For example, given two individuals, A and B, with the same current normal blood pressure, the cardiovascular risk profile of person A who historically has always had normal blood pressure is different from the risk profile of person B whose blood pressure has been normal only for a short time after being elevated for an extended period of time [3]. This difference in risk due to different histories has been called the *legacy effect*. Prediction models that are able to capture legacy effects may, therefore, estimate risk more accurately.

Electronic health data (EHD), including electronic medical records (EMRs), insurance claims data, and mortality data obtained from the state government, provide longitudinal observations on a large number of individuals over a significant timespan. Availability of such data allows for constructing dynamic prediction models that can take into account individual history. Additionally, the increased usage of EMRs in clinical practice ensures that longitudinal risk factor information will be routinely stored and can easily be fed into a risk prediction model as part of a clinical decision support system.

EHD databases typically include records on hundreds of thousands to millions of heterogeneous patients who receive care in contemporary clinical settings. With many data-capture systems having been put in place in the early 2000s, up to 15 years of historical data may be available in some cases, but follow-up over 5-10 years is currently more common. Despite their large sample sizes and representativeness, electronic health data are also often "messy", in the sense that risk factors may be measured sporadically and the duration of follow-up may vary widely between individuals.

We seek to develop a flexible risk modeling approach which incorporates information on heterogeneous, longitudinally measured risk factors while accommodating for

the challenging characteristics of EHD. There are several flexible machine learning techniques – among them support vector machines (SVM), decision trees, and neural networks – that can be applied to longitudinal data representing changes of patient state. However, these techniques generally assume that data are "clean", with fully observed (i.e., non-censored) outcomes and no missing predictor information. Standard regression techniques for censored time-to-event data do not natively handle missing predictor information, lack flexibility because the regression relationship between risk factors and outcomes must be pre-specified, and do not exploit the temporal relationships between highly correlated, longitudinally measured predictors.

In this paper, we introduce a risk prediction technique for EHD based on a Dynamic Bayesian network (DBN) model which describes the longitudinal relationships between historical risk factors and the outcome of interest. Dynamic Bayesian networks, unlike other machine learning techniques, are able to natively handle missing data and, unlike standard regression-based techniques, exploit the longitudinal structure covariates measured over time. However, (Dynamic) Bayesian networks assume that the outcome of interest (i.e., whether or not a subject experiences the event of interest within τ years) is observed on all subjects. However, some subjects in the dataset will be observed for fewer than τ years due to disenrollment from the health system. Further, most common Bayesian network implementations assume that the predictors are of the same type, i.e., all continuous or all discrete. But clinical risk prediction models often include both discrete (e.g., sex) and continuous (e.g. age) predictors.

Our method extends the traditional DBN in two key ways: First, by applying inverse probability of censoring weighting (IPCW) to account for the fact that event times are right-censored. And second, by implementing a novel model fitting algorithm to accommodate combinations of discrete and continuous predictors. We illustrate the application of our extended DBN in the context of predicting cardiovascular risk using EHD from a large Midwestern health insurance provider. In this context, we compare the performance of DBN to competing techniques, and study the influence of the temporal resolution of historical observations on prediction accuracy.

3.3 Dynamic Bayesian networks for right-censored data

Let $\boldsymbol{X}$ represent the vector of observed values for a set of possible risk factors (e.g., blood pressure, cholesterol levels, medication use, etc.) which will be used to predict the occurrence of cardiovascular (CV) events. Some of these factors may be recorded several times prior to the "baseline", i.e., the time at which we wish to predict risk. The risk model seeks to estimate $P(E = 1|\boldsymbol{X})$, where E is an indicator variable for whether or not a random subject experienced a health outcome or adverse event within τ years from baseline. There are many techniques which may be applied to estimate this conditional probability. Here, we focus on the Bayesian network technique, which is motivated by the fact that $P(E = 1|\boldsymbol{X})$ can be expressed using Bayes Theorem as:

$$P(E = 1|\mathbf{X}) = \frac{P(\mathbf{X}|E = 1)P(E = 1)}{\sum_{e\in\{0,1\}} P(\mathbf{X}|E = e)P(E = e)}, \tag{3.1}$$

Initially, re-expressing the problem in this way might seem ill-advised, as it replaces estimation of the conditional probability of the scalar E, given $\boldsymbol{X}$, with the challenging task of estimating the joint distribution of $\boldsymbol{X}$, given E. Although estimating the distribution of a high-dimensional $\boldsymbol{X}$ would be challenging in general, the idea of the Bayesian network is to exploit known or assumed relationships between the components of $\boldsymbol{X}$ to make the estimation task more manageable by identifying sets of factors which are conditionally independent. In our case, the conditional dependencies follow naturally from the temporal ordering of the historical risk factor data, yielding what is typically referred to as a Dynamic Bayesian network (DBN).

3.3.1 Dynamic Bayesian networks

Bayesian networks can efficiently represent static systems; however, for dynamic systems, such as speech recognition systems, robotics, etc. that change over time and produce sequential data, Dynamic Bayesian networks provide a more advantageous representation. A Dynamic Bayesian network is a generalization of a Hidden Markov Model (HMM) [74]. As in a HMM, a Dynamic Bayesian network consists of variables that indicate the state of the dynamic system at a given point of time. A DBN imposes

restrictions on the dependency of the states, i.e., any given state depends only on the state preceding it. While states in an HMM are represented as a single node, a "state" in a DBN may encompass multiple nodes that depend on each other. Typically, the structure of network corresponding to each state is identical, and in addition to the intra-state dependencies, any given node in a state depends on the corresponding node in the previous state.

In medicine, Dynamic Bayesian networks are frequently used to model conditions, such as progression of ventilator-associated pneumonia [22], blood glucose variation over time to control insulin administration [6], outcome of a particular treatment for patients with congenital cardiac anomaly [78], or in prediction outbreaks of influenza based on historical records of influenza-related hospitalizations [90]. In most of these systems, the desired prediction is a function of its multiple historical states which generally vary temporally. While in most studies Dynamic Bayesian networks have been used to predict the current state of the system (such as blood glucose based on its previous observations), we use DBNs not to model the progression of risk factors but to model the effect of a sequence of observations (of the risk factors) on the risk of adverse cardiovascular (CV) events, such myocardial infarctions or strokes. As mentioned earlier, an overwhelming majority of CV risk prediction models have focused on a single measurement of risk factors [92]. However, some recent studies have demonstrated that examining the trajectories of risk factors, such as systolic blood pressure (SBP), leads to better prediction of CV events [103], because these trajectories are known to affect the risk of CV events [3, 4, 75]. Dynamic Bayesian networks provides us a technique to capture such temporal trajectories in a model that is open to easy interpretation, can capture the potentially complex relationship between risk factor history and clinical risk, is able handle missing data in a principled fashion, and can be easily adapted for censored observation.

3.3.2 Defining the network

In Section 3.4, we will use Dynamic Bayesian networks to build models for predicting the risk of experiencing a CV event. The DBN which we use to model CV risk as a

function of individual risk factors is shown as a directed acyclic graph (DAG) in Figure 3.1. Edges of the DAG are used to encode (conditional) dependencies between different measured variables. In addition to dependencies between measurements of the same covariate across a time interval (such as dependence of systolic blood pressure at time interval 1 on time interval 0), there are also dependencies between different risk factors within time intervals. For example, the blood pressure within each time period depends on the presence or absence of blood pressure medications in the corresponding interval. In the figure, these dependencies are indicated by arrows between nodes in the same box. Other risk indicators, such as gender, which does not depend on time, and age, which is straightforwardly deterministic with respect to time, are represented as nodes external to (and interacting with) all of the time-dependent risk factors.

More generally, we can write $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_T)$ where $\boldsymbol{X}_t$ are measurements at time period t, $0 \leq t \leq T$, and where $\boldsymbol{X}_T$ represents the current (i.e., baseline) status of the patient. The vector $\boldsymbol{X}_t = \{X_t^0 \cdots X_t^i \cdots X_t^N\}$ represents individual risk factors, where X_t^i is the ith risk factor observed at time t. For instance, x_1^0, x_2^0, ..., x_t^0 could correspond to blood pressure measurements at times 1, 2, and t respectively, while x_1^1, x_2^1, ..., x_t^1 could correspond to lipid measurements at times 1, 2, and t respectively.

To simplify exposition, we assume that all the X_i contain information on the same set of risk factors for all i. That is, the same set of risk factors are measured at the same set of times; however, the framework is easily extended to accommodate factors that are measured only once or only at a subset of specified times. The key simplifying assumption, which allows the joint distribution of $\boldsymbol{X}$ to be factored into simpler subcomponents, is the Markov property:

$$P(X_t^j | \boldsymbol{X}_t^{-j}, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots, \boldsymbol{X}_0, E) = P(X_t^j | \boldsymbol{X}_t^{-j}, X_{t-1}^j, E)$$

where $\boldsymbol{X}_t^{-j}$ indicates the omission of risk factor j from $\boldsymbol{X}_t$.

In other words, the value of risk factor j at time t is conditionally independent of its values measured prior to $t-1$ given: the value of risk factor j at $t-1$, current values of other risk factors, and the event indicator E. Although this Markov assumption constrains the joint distribution of the risk factors, it still allows the event indicator to

depend on all the historical risk factors. That is, the entire risk factor trajectory can have an impact on the risk predictions.

Based on the independence assumptions defined in the Dynamic Bayesian network, we can evaluate the probability $P(\boldsymbol{X}, E)$ as follows:

$$P(\boldsymbol{X}|E) = \prod_{i,t,e} P(x_t^i|\mathrm{Pa}^t(x_t^i), \mathrm{Pa}^{t-1}(x_t^i), e) \tag{3.2}$$

where $\mathrm{Pa}^t(x)$ are the parent nodes[1] in time t of the node x. We can further represent all the parents of x_t^i as $\boldsymbol{Y_t^i} = \{\mathrm{Pa}^t(x_t^i), \mathrm{Pa}^{t-1}(x_t^i)\}$. Thus,

$$P(\boldsymbol{X}|E) = \prod_{i,t} P(x_t^i|\boldsymbol{Y_t^i}, E) \tag{3.3}$$

The probability of an event occurring, i.e., $P(E|\boldsymbol{X})$, can be obtained by simply by substituting (3.3) in (3.1) resulting in

$$P(E = 1|\mathbf{X}) = \frac{\prod_{i,t} P(x_t^i|\boldsymbol{Y_t^i}, E = 1)}{\sum_{e\in\{0,1\}} \prod_{i,t} P(x_t^i|\boldsymbol{Y_t^i}, E = e)}, \tag{3.4}$$

3.3.3 Modeling and estimation

There are several possible approaches to evaluating the conditional probabilities in Equation (3.4). Our proposed technique is to represent the joint distribution $P(x_t^i, \boldsymbol{Y_t^i}|E = e)$ using a partitioned mixture of Gaussian distributions, which we describe in Section 3.3.3, for each value of E, then fit the resulting model using an Expectation Maximization (EM) algorithm. Evaluating the conditionals $P(x_t^i|\boldsymbol{Y_t^i}, E = e)$ from the joint distribution is straightforward.

There are three key complicating factors for estimation in our setting. First, the risk factor set $\boldsymbol{X}$ may contain both discrete and continuous predictors, so that assuming a simple mixture of Gaussian distributions to model predictors is not appropriate. Second, risk factor values may be missing. And third, because follow-up times are unequal and

[1] In a directed acyclic graph, the set of *parents* V of a node U are the variables such that, given V, U is independent of the all the remaining variables in the graph (except for the variables of which U itself is a parent). Hence, the joint distribution of all the variables in a graph can be factored into conditionally independent components, where the conditioning is on each node's parents. A variable W is a *child* of U if U is a parent of W.

are subject to censoring, the event indicator E is not observed for all individuals. In the remainder of this section, we describe how our technique deals with these challenges.

Joint modeling of discrete and continuous risk factors

Let $\boldsymbol{Z} = \{x_t^i, \boldsymbol{Y_t^i}, E\}$, where $\boldsymbol{Z}$ can be partitioned into discrete and continuous parts, i.e., $\boldsymbol{Z_d}$ and $\boldsymbol{Z_c}$, respectively. For example, $\boldsymbol{Z_d}$ could consist of information on sex (male/female), smoking history (current/ever/never), and use of SBP medications (0/1/2/3+ classes), and $\boldsymbol{Z_c}$ could consist of information on age, blood pressure, and cholesterol.

The joint distribution of $\boldsymbol{Z}$ can be represented as: $P(\boldsymbol{Z}) = P(\boldsymbol{Z_d}, \boldsymbol{Z_c}) = P(\boldsymbol{Z}_c|\boldsymbol{Z}_d)P(\boldsymbol{Z}_d)$. $P(\boldsymbol{Z}_c|\boldsymbol{Z}_d)$ is modeled as a mixture of M multivariate normal distributions, as follows:

$$P(\boldsymbol{Z}_c = z_c|\boldsymbol{Z}_d = z_d) = \sum_{m=1}^{M} \rho_{m,z_d} \phi\left(\Sigma_{m,z_d}^{-1}(z_c - \mu_{m,z_d})\right), \tag{3.5}$$

where ϕ is the density function of a multivariate standard normal random variable of dimension $|\boldsymbol{Z}_c|$, with mean μ_{m,z_d}, variance matrix Σ_{m,z_d}, and mixing parameters ρ_{m,z_d}. $P(\boldsymbol{Z}_d)$ is estimated as the fraction of samples where $\boldsymbol{Z}_d = \boldsymbol{z_d}$.

Note that distinct values of the mean, variance, and mixing parameters are estimated for each value of $\boldsymbol{z_d}$; thus, this approach yields many parameters to estimate. Specifically, if S is the number of parameters to estimate for a given value of $\boldsymbol{Z_d}$ and E, and Q_k is the number of possible states of the k^{th} dimension of $\boldsymbol{Z}_d$, then there are $2 \times S \times \prod_{k=1}^{|\boldsymbol{Z}_d|} Q_k$ parameters to estimate (the factor of two comes from the fact that the parameters must be estimated separately for the cases where $E = 0$ and $E = 1$).

The number of mixture components, M, is a tuning parameter. For a fixed number of mixing components M, an inverse probability of censoring weighted expectation maximization algorithm (IPCW-EM, described below) is used to solve for the maximum likelihood estimators of the mean, variance, and mixing parameters. In general, to control overfitting, one could consider M to be a tunable parameter and select the number of mixture components using the Bayes Information Criteria (BIC) or some other goodness-of-fit measure (see, e.g., [8]).

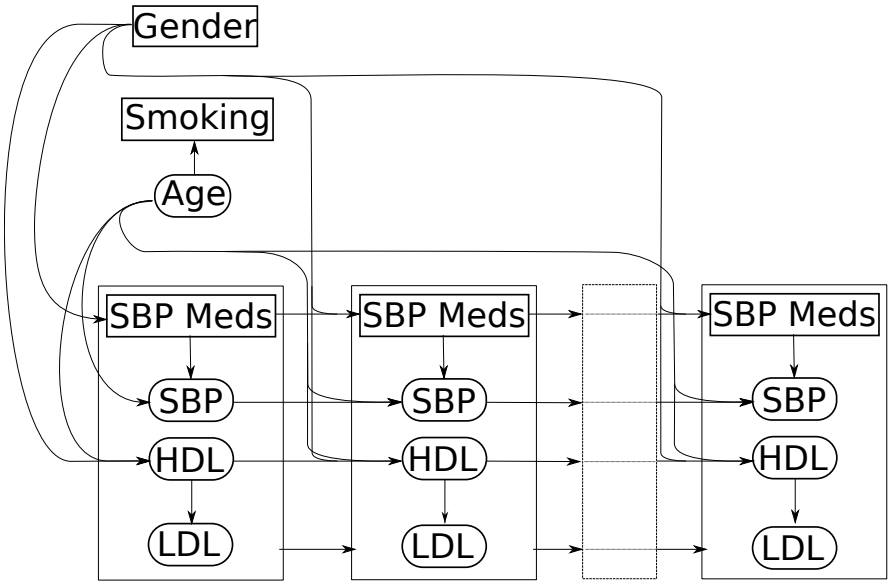


Figure 3.1: Dynamic Bayesian network for CV risk: The large boxes logically group measurements of the "dynamic" variables, i.e., Systolic Blood Pressure (SBP), Systolic Blood Pressure Medications (SBP Meds), High Density Lipoprotein (HDL), and Low Density Lipoprotein (LDL) at time $t = 0, 1 \cdots T$ respectively. The solid arrows indicate dependence between the variable pair connected by the arrow. The variables Gender, Smoking, and Age are either assumed not to change over time (Gender, Smoking) or progress deterministically from their initial value (Age). All the variables in this figure are connected to the event node which is not displayed.

Missing risk factor values

The above description of inference applies when all the risk factors in $\boldsymbol{X}$ are observed. However, this is usually not the case, and the missing covariates must be "marginalized out" of Equation (3.4). Let us partition $\boldsymbol{X}$ into 2 parts, $\boldsymbol{X}_o$ and $\boldsymbol{X}_h$, which indicate the covariates that are observed ($\boldsymbol{X}_o$) and those that are hidden/missing ($\boldsymbol{X}_h$). To evaluate $P(\boldsymbol{X}|E = e)$ in such a scenario, we need to evaluate $\int_{x \in \boldsymbol{X}_h} P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{X}_h = x, E = e) dH(x|e)$ where $H(x|e) = P(\boldsymbol{X}_h \leq x|E = e)$. This integral is relatively straightforward to compute if the $x \in \boldsymbol{X}_h$ all appear in terminal nodes of the network,

i.e, they appear in only a single term in Equation (3.4). If the missing covariates appear in more than one product term, then there is no analytical solution to the resulting integral, in which case we use a Markov-chain Monte-Carlo technique to evaluate the integrals.

To marginalize out missing data (that appear in non-terminal nodes of the network), we construct a group of nodes consisting of the missing node and the nodes that make up its Markov blanket, i.e., the set of nodes consisting of the missing node's parents, children, and children's other parents. If the value of any node in the Markov blanket is also missing, we include additional nodes from its Markov blanket resulting in an enhanced Markov blanket around one or more nodes with missing observations. We continue growing the enhanced blanket until all the nodes making up the blanket are observed. Let the group of nodes with the missing observations and their enhanced Markov blanket be represented by $\boldsymbol{X}^c$, which can further be partitioned into $\boldsymbol{X}_o^c$ and $\boldsymbol{X}_h^c$, representing the observed and hidden/missing nodes in the group respectively. We generate random samples of $\boldsymbol{X}_h^c$ from the joint distribution $P(\boldsymbol{X}^c|E=e) = \prod P(X_k|\text{Pa}(X_k), E=e)$, where $\{X_k, \text{Pa}(X_k)\} = \boldsymbol{X}^c$ by simply holding the values of $\boldsymbol{X}_o^c$ fixed, while the slice sampling algorithm [76] is used to sample from $\boldsymbol{X}_h^c$. In our implementation, we draw 1000 samples for every missing observation, compute $P(E|\boldsymbol{X})$ using every sample, and then average the outcome to arrive at the final risk prediction. This technique for approximate inference is described more completely in [15].

Unequal follow-up times and censoring

EHR data include patients who are under observation for unequal amounts of time. Patients may change their insurance provider, move out of the healthcare system, or die at varying times after "baseline". Patients that leave the healthcare system do not have any more data captured in the electronic health database afterward. Patients who leave the health system or die of causes other than CVD are said to be *right-censored*, and their event status during times when they are no longer under observation is unknown. Typically, CV risk predictions are desired over a fixed time horizon, e.g., 5 years from baseline. A substantial fraction of patients in our motivating dataset have less than 5

years of follow-up, and those who do not experience a CV event are only known to be CV event-free up until the end of their follow-up.

How should such patients contribute to model fitting? A naive approach to evaluating $P(\boldsymbol{X}|E)$ could be to assume that these patients did not have events, but doing so would underestimate the true event rate, since some fraction of patients will experience events during the time that they are not under observation. Conversely, excluding right-censored patients entirely will overestimate the true event rate, since patients who experience events early on during follow-up are over-represented.

To prevent biases due to a naive approach towards censoring, we adjust for censoring using inverse probability of censoring weights (IPCW). . The IPCW approach consists of discarding patients with unknown event status (as with the naive approach described previously) and computing a weight for each patient with known event status that is the inverse of the probability that a patient is observed for a duration greater than the least of: (1) the follow-up period (i.e., τ), (2) the time interval between end of baseline and a CV event (i.e., T), or (3) the duration for which a patient is followed up or enrolled (i.e., C). IPCW will assign larger (smaller) weights to patients with known event status that "represent" a large (small) number of other patients whose event status is unknown. In our setting, patients who experience events shortly after baseline receive small weights, with weights increasing for patients who experience events closer to 5 years after baseline.

More formally, IPCW defines a weight for every observation(i.e., every subject i) as follows:

$$\omega_i = \begin{cases} 0 & \text{if } T_i > C_i \text{ and } C_i < \tau \\ \frac{1}{\hat{G}(\min(C_i, T_i, \tau))} & \text{otherwise} \end{cases} \tag{3.6}$$

where $\hat{G}(t)$ is the Kaplan-Meier estimator [53] of $G(t) \equiv P(C > t)$, defined by:

$$\hat{G}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i^*}{n_i} \right), \tag{3.7}$$

where d_i^* is the number of subjects who were censored at time t_i and n_i are the number of subjects at risk (i.e., number of subjects not previously censored or experiencing a

CV event) at time t_i. We assume that $G(t)$ is independent of all other risk factors.

The IPCW weights are incorporated in the model fitting procedure by applying a weighted version of the EM algorithm to determine $P(\boldsymbol{X}|E)$, where the ω_i is the contribution of the i^{th} subject to likelihood. Further details are provided in Bandyopadhyay et al. [8].

3.4 Application to predicting cardiovascular risk using EHD

Myocardial infarctions (MI) and strokes are among the leading causes of death in United States, with the treatment costs of CV-related conditions approaching $300 billion annually [42]. Risk prediction models are important tools used in management of CV disease, and are particularly relevant for complex patients, i.e., those with multiple risk factors [82, 18]. Accurate CV risk predictions help physicians identify high-risk patients and formulate intervention strategies that are more effective in treating CV disease and ultimately leading to fewer CV events.

There are over 100 CV risk models that are currently used [30, 31, 71] including several popular ones, such as the Framingham Risk Score [34], ACC/AHA Pooled Cohort Equations [43], SCORE [28], ASSIGN-SCORE [113], QRISK1 [47, 46], QRISK2 [48], PROCAM [7], WHO/ISH, and Reynolds Risk Score [83, 84]. These models are constructed using data from longitudinal cohort studies, the populations of which may not be representative of the patients that are regularly seen in clinical practice. The accuracy of these risk models varies substantially across different target populations [52] and, while accuracy can be improved by retraining a model on a specific target population [88, 109], risk prediction performance typically remains relatively poor. Another shared characteristic of these risk models is that they evaluate risk based only on the currently observed state of the patient, and ignore the trajectory of risk factor values prior to the current state. Our goal was to illustrate how historical risk factor information derived from the electronic health record could be used to improve the accuracy of cardiovascular risk prediction models.

3.4.1 Data sources

Our study was conducted using electronic health data from a healthcare network in the Midwestern U.S. The data includes enrollment records, demographics, census data, diagnosis codes, prescription information, claims, vital signs, and laboratory measurements from patients who were enrolled in the network insurance plan or attended a network clinic between 2001 and 2012. While the total number of individuals represented in the database is approximately 450,000, many of these individuals did not seek routine care through the system. The criteria described below are designed in part to exclude these individuals, as their health status is frequently unknown and they are unlikely to be seen regularly within that healthcare system for management of their cardiovascular risk.

3.4.2 Defining a cohort

Our initial cohort consists of patients who have at least one systolic blood pressure (SBP) measurement taken at in-network clinic when they were age 21 or older, and who had been continuously enrolled in the health plan for at least three years prior to this measurement. SBP measurements taken in acute care settings (e.g., ER, urgent care, surgery, etc.) are excluded. The earliest SBP measurement satisfying the above criteria defines the *baseline* time point for each individual. The *observation period* consists of the three years prior to baseline, across which historical risk factor values may be observed and aggregated. The *follow-up period* for a patient begins at baseline and continues until the earliest date on which a patient: (1) experiences a CV event, (2) dies, (3) is no longer enrolled in the health plan (if not re-enrolled within 90 days), (4) reaches the end of the data recording period (i.e., December 31, 2011).

The enrollment criteria for both the observation and follow-up periods ensures that we can accurately determine events, because the events are primarily determined from the claims data which are available only for people enrolled in the insurance plan. Patients with diabetes[2] or other previous cardiovascular comorbidities (including prior

[2] Diabetes was defined based on joint consideration of inpatient and outpatient ICD-9 diagnosis

MI or stroke) were also excluded. We also restricted attention to individuals with complete drug coverage through the same healthcare network; thus, information on prescriptions and refills is likely to be accurate. After applying these inclusion criteria, the final sample size was 168,661.

3.4.3 Baseline and historical risk factors

The variables available to predict CV risk were: age in years, sex (female/male), smoking status (never/current/quit), BMI (in kg/m^2), systolic blood pressure (in mmHg), high-density lipoprotein (HDL, in mg/dL), low-density lipoprotein (LDL, in mg/dL), and number of classes of blood pressure medications (0/1/2/3+). The first three of these variables (age, sex, smoking status) were viewed as "baseline" variables and their values were recorded as the latest available values prior to baseline. The remaining variables were considered "historical" risk factor values whose pre-baseline trajectory during the observation period was of interest.

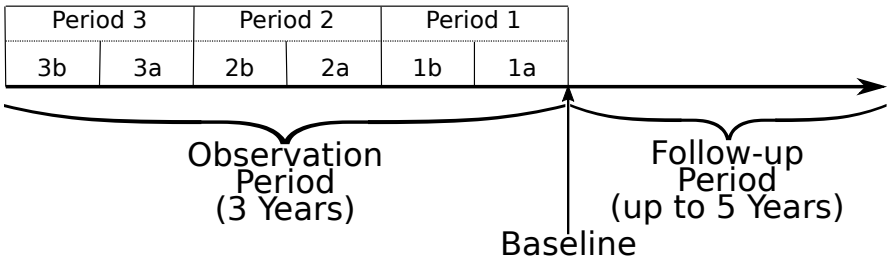


Figure 3.2: Partitioning the observation period.

As shown in Figure 3.2, the observation period was partitioned into three smaller periods (1, 2, and 3, corresponding to number of years prior to baseline) and 2 half-year sub-periods (a and b) within each period. Each historical risk factor may be measured multiple times (or not at all) during each interval, with the exception of blood pressure which by definition of the observation period must have been measured at least once in the six months preceding baseline (i.e., during sub-period 1a). When the same risk

codes, use of glucose-lowering medications, and glucose-related laboratory values using a previously validated algorithm with estimated sensitivity of 0.91 and positive predictive value of 0.94 [36].

factor was measured multiple times during an interval, the reported value was set to the average of measurements. For medications, we considered only the period of 60 days from the end of each period. This helps us minimize the inaccuracies in medication classes when physicians switch medications over a time period. Hence, at most six longitudinal values could be recorded for each risk factor. We also considered some models where data were over one-year and three-year rather than six-month periods, in which case available measurement values were averaged over the longer time period. While shorter periods allow closer tracking of changes in the risk factors, in many cases risk factor measurements were not made at a high enough frequency to provide risk factor values in all periods. Table 3.1 summarizes the baseline and historical risk factors over three one-year periods prior to baseline, including the proportion of each that are missing. One of the advantages of the Dynamic Bayesian model is that it naturally handles such missing data without the need to explicitly impute missing values prior to analysis.

Variable	Mean (SD) or N (%)			Missing, %		
	Period 3	Period 2	Period 1	Period 3	Period 2	Period 1
Historical risk factors						
SBP (in mmHg)	119.65(14.72)	118.90(14.28)	118.73 (14.32)	48.89	38.13	0
# of BP meds						
0	155,781 (92.4)	151,742 (89.9)	147,312 (87.3)	0	0	0
1	7,981 (4.7)	10,186 (6.0)	12,330 (7.3)	0	0	0
2	3,789 (2.2)	5,084 (3.0)	65,37 (3.8)	0	0	0
3+	1,110 (0.7)	1,649 (1.0)	2,482 (1.4)	0	0	0
LDL (in mg/dL)	154.56(37.04)	149.68 (36.49)	148.72 (37.16)	85.48	82.64	75.10
HDL (in mg/dL)	50.66 (14.32)	51.26 (14.58)	51.63 (14.44)	79.75	77.25	67.45
Baseline risk factors						
Sex (female)	98,508 (58.4)			0		
Age (in years)	41.62 (15.26)			0		
Smoking						
Never	120,967 (71.7)			0		
Current	28,394 (16.8)			0		
Quit	19,300 (11.4)			0		

Table 3.1: Summary of historical and baseline risk factors used to predict CV risk for the entire population.

3.4.4 Cardiovascular events

We used the definition of a composite "CV event" similar to that used in the development of the Framingham Risk Score [34] except that our definition did not include events corresponding to congestive heart failure. Event times were recorded as the time to first cardiovascular event or death from a cardiovascular cause; subjects were censored if they disenrolled from the health plan, died of a non-cardiovascular cause, or did not experience an event during the follow-up period. Major CV events were ascertained as the first occurrence based on the date of primary hospital discharge ICD-9 diagnosis codes as follows: 1) myocardial infarction (MI)/acute coronary syndrome (ACS) (ICD-9 codes 410.0-410.91, 411.1,and 411.8); 2) ischemic and hemorrhagic stroke (433-434.91

and 430-432.9); 3) heart failure (HF) (428-428.9); or 4) peripheral artery disease (PAD) (intermittent claudication, 440.21 and 443.9). Time to event was calculated as days elapsed from baseline to the hospital discharge date associated with the given event. Mortality data, including cause of death (CV-related or not), were extracted from administrative and state death registries and were available with a 1-year lag.

3.4.5 Model Comparisons

Table 3.2 describes the prediction models we compared. The Dynamic Bayesian networks used the DAG structure in Figure 3.1 and were fitted according to the methods described above. Cox proportional hazards models were fitted using the time to event and censoring indicators. The regression models included a main effect term for each of the variables in Figure 3.1, including multiple historical risk factor values where applicable.

The models in Table 3.2 vary according to the fitting technique used (Dynamic Bayesian network vs. Cox proportional hazards regression) and the way in which historical information is used and aggregated. The time periods are labeled as in Figure 3.2, with ✓and ✗ indicating whether or not data from that time period is used in the model. For example, the model "Bayes 123(a)" is constructed using 3 six month periods (1a, 2a, 3a) and has a higher temporal resolution than the model "Bayes 123" constructed using 3 one year periods (1, 2, 3), even though both models are constructed using 3 years of observations. The model "Bayes merged", constructed using a single data period which consists of measurements averaged across the entire three-year observation period, has the lowest temporal resolution. Note that models "Bayes 1(a)", "Bayes 1", and "Bayes merged" use only one set of baseline predictor values (measured in period 1(a), averaged over period 1, and averaged over all three periods, respectively) and, hence, represent standard (i.e., non-dynamic) Bayesian networks, but with different temporal resolutions.

Prediction model	Period						Description
	3		2		1		
	3b	3a	2b	2a	1b	1a	
Bayes 1(a)	✗	✗	✗	✗	✗	✓	1 six-month period, regular Bayes
Bayes 1(ab)	✗	✗	✗	✗	✓	✓	2 six-month periods, 1-year history, Dynamic Bayes
Bayes 123(a)	✗	✓	✗	✓	✗	✓	3 six-month periods, 3-year history, Dynamic Bayes
Bayes 123(ab)	✓	✓	✓	✓	✓	✓	6 six-month periods, 3-year history, Dynamic Bayes
Bayes 1	✗	✗	✗	✗	✓		1 one-year period, regular Bayes
Bayes 123	✓		✓		✓		3 one-year periods, 3-year history, Dynamic Bayes
Bayes merged	✓						1 three-year period, 3-year history, regular Bayes
Cox 1(a)	✗	✗	✗	✗	✗	✓	1 six-month period, Cox model
Cox 123(a)	✗	✓	✗	✓	✗	✓	3 six-month periods, 3-year history, Cox model
Cox 1	✗	✗	✗	✗	✓		1 one-year period, Cox model
Cox 123	✓		✓		✓		3 one-year periods, 3-year history, Cox model
Cox merged	✓						1 three-year period, 3-year history, Cox model

Table 3.2: Model Descriptions.

To ensure unbiased assessments of prediction accuracy, we fit our models using 75% of the available data ("training data") and estimate prediction accuracy of the fitted models on the remaining 25% of the data ("test data"). We summarize the prediction accuracy of risk prediction models by assessing their calibration and discrimination using established metrics.

Calibration

The calibration of a model indicates whether the model predicts events at a rate that is comparable to event rates that are observed in the cohort under study. We compute a Hosmer-Lemeshow type calibration statistic [49]:

$$K = \sum_{j=1}^{B} \frac{(\bar{p}_j - p_j^{KM})^2}{var(p_j^{KM})} \tag{3.8}$$

where the sum is taken over B equispaced partitions of the test set determined by quantiles of predicted risk. p_j is the average predicted risk in partition j ("Expected")

p^{KM} is the Kaplan-Meier estimate of experiencing an event before ("Observed"). The variance $var(p^{KM})$ is calculated using Greenwood's formula [44].

Discrimination

The concordance index (C-index) is a measure of discrimination of the model. One of the most popular measures of discrimination is the area under the ROC curve (AUC). In the context of risk prediction model, this measure is equivalent to a measure of concordance which can be thought of as the probability that a person having an event is assigned a higher risk than a person without an event by the model. While it is difficult to evaluate AUC because the outcomes are censored, the traditional concordance measure (which is exactly equivalent to AUC for non-censored data) can be extended to application settings where outcomes are censored.

As described in [45], the C-index adapted for censoring considers the concordance of survival outcomes versus predicted survival probability among pairs of subjects whose survival outcomes can be ordered; namely, among pairs where both subjects are observed to experience a CV event, or one subject is observed to experience a CV event before the other subject is censored. Pairs in which both subjects are censored or in which the censoring time of one precedes the failure of the other do not contribute to this metric. Let $P(E_i|\boldsymbol{X}_i)$ be the estimated probability that the i^{th} subject experiences an event within τ years. Formally, this adapted C-index is given by:

$$C_{cens}(\tau) = \frac{\sum_{i \neq j} \delta_i \mathbb{I}[V_i < V_j] \mathbb{I}[\hat{P}(E_i|X_i) < \hat{P}(E_j|X_j)]}{\sum_{i \neq j} \delta_i \mathbb{I}[V_i < V_j]} \tag{3.9}$$

3.5 Results

The results of applying the models described in Table 3.2 to the test data are presented in Table 3.3. Calibration is generally better in the Bayes models (k-statistic ranging from 3.4-16.7) compared to the Cox models (k-statistic ranging from 16.1-22.1). Discrimination performance is similar across all the models, with Bayes models exhibiting a slight advantage. While the differences are relatively modest, Bayes models constructed

using data spanning a longer duration (e.g., Bayes 123(a), 123(ab), and 123) appear to have better calibration and discrimination than those constructed with data spanning a shorter interval (e.g., Bayes 1(a), 1(ab), and 1). In contrast, the calibration and discrimination of Cox models was insensitive to the way in which historical information was incorporated.

Prediction model	Period						k-statistic (SE)	C-index (SE = 0.001)
	3		2		1			
	3b	3a	2b	2a	1b	1a		
Bayes 1(a)	✗	✗	✗	✗	✗	✓	11.9 (0.5)	0.874
Bayes 1(ab)	✗	✗	✗	✗	✓	✓	14.3 (0.6)	0.884
Bayes 123(a)	✗	✓	✗	✓	✗	✓	8.8 (0.4)	0.891
Bayes 123(ab)	✓	✓	✓	✓	✓	✓	9.8 (0.4)	0.885
Bayes 1	✗				✓		4.8 (0.3)	0.880
Bayes 123	✓		✓		✓		3.4 (0.3)	0.893
Bayes merged	✓						16.6 (0.5)	0.879
Cox 1(a)	✗	✗	✗	✗	✗	✓	19.6 (0.6)	0.870
Cox 123(a)	✗	✓	✗	✓	✗	✓	16.1 (0.6)	0.871
Cox 1	✗				✓		18.1 (0.5)	0.870
Cox 123	✓		✓		✓		18.5 (0.6)	0.871
Cox merged	✓						22.0 (1.5)	0.869

Table 3.3: Calibration and discrimination statistics of models evaluated on test data.

Our models also considered differing degrees of aggregation across data periods covering the same duration. Overall, periods that were short (e.g., six months in Bayes 123(a)) appeared to yield poorer calibration and discrimination than longer periods (e.g., one year in Bayes 123), but aggregating data into periods that were too long (e.g., three years in Bayes merged) also decreased predictive performance. We hypothesize that overly short periods perform poorly because of lower data quality; specifically, the corresponding models have a higher fraction of missing observations within each period

which must be internally imputed. For overly long (i.e., 3-year) periods, the loss of accuracy can be attributed to the fact that averaging observations removes information about variation of risk indicators, such as SBP and SBP medications, which are known to significantly affect the risk of CV events. The best calibrated model – which also achieves the highest C-index – is Bayes 123, which uses yearly data from all three periods.

3.5.1 Results stratified by historical changes in blood pressure

The risk of CV events depends not only on the current state of risk indicators, but also on the past state of these indicators. As noted earlier, blood pressure is particularly prone to longitudinal fluctuations which may affect future CV risk. To demonstrate the legacy effect of blood pressure and to further evaluate the performance of our model among those individuals who are most likely to experience such legacy effects, we identify three distinct cohorts from our population:

1. **Stable BP:** Individuals whose systolic blood pressure in period 3 is **within 5mm Hg** of their blood pressure in period 1.

2. **Rising BP:** Individuals whose systolic blood pressure in period 3 is **more than 10mm Hg lower** than their blood pressure in period 1.

3. **Falling BP:** Individuals whose systolic blood pressure in period 3 is **more than 10mm Hg higher** than their blood pressure in period 1.

Some characteristics of the three cohorts are summarized in Table 3.4. The Stable BP and Rising BP cohorts are similar on most dimensions except for the observed change in SBP and the 5-year CV event rate. The Falling BP cohort is quite different from the other two: it is slightly older on average, and more likely to be taking 1 or more blood pressure medications. Also, the 5-year CV event rate is higher, at 5.5%. It might seem counterintuitive that a group whose blood pressure is decreasing has a higher event rate, but this finding makes sense when one considers that approximately

one-third of these individuals are taking blood pressure medications, which indicates that they were judged at some point to be at higher risk.

	BP Cohorts		
	Stable (N = 59,959) Mean (SD) or N (%)	*Rising* (N = 24,933) Mean (SD) or N (%)	*Falling* (N = 33,541) Mean (SD) or N (%)
Age	44.6 (17.5)	46.4 (18.1)	50.0 (18.9)
Sex			
Male	22,864 (38)	11,010 (44)	14,188 (42)
Female	37,095 (62)	13,923 (56)	19,353 (58)
# of BP meds			
0	48,147 (80)	19,274 (77)	22,038 (66)
1	5,782 (10)	2,631 (11)	4,378 (13)
2	3,732 (6)	1,831 (7)	4,125 (12)
3+	2,298 (4)	1,197 (5)	3,000 (9)
3-year change in SBP	-0.1 (2.7)	17.2 (8.7)	-18.2 (9.2)
5-year CV event rate	3.2%	4.3%	5.5%

Table 3.4: Characteristics of three cohorts defined by change in SBP over three years prior to baseline. *Stable* = Change in SBP of less than 5 mmHg, *Rising* = Increase of more than 10 mmHg, *Falling* = Decrease of more than 10 mmHg.

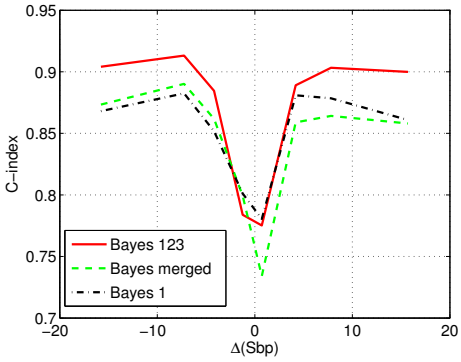
Prediction model	Period						k-statistic			C-index		
	3		2		1		BP cohort			BP cohort		
	3b	3a	2b	2a	1b	1a	Stable	Rising	Falling	Stable	Rising	Falling
Bayes 1(a)	✗	✗	✗	✗	✗	✓	4.64	3.50	3.96	0.834	0.851	0.861
Bayes 1(ab)	✗	✗	✗	✗	✓	✓	3.87	1.28	3.04	0.841	0.883	0.882
Bayes 123(a)	✗	✓	✗	✓	✗	✓	4.22	2.63	6.02	0.866	0.880	0.896
Bayes 123(ab)	✓	✓	✓	✓	✓	✓	5.41	2.56	2.96	0.862	0.878	0.878
Bayes 1	✗	✗	✗	✗	✓		3.36	4.37	5.17	0.847	0.850	0.860
Bayes 123	✓		✓		✓		2.07	1.88	1.70	0.860	0.898	0.902
Bayes merged	✓						3.44	3.21	3.98	0.863	0.869	0.893
Cox 1(a)	✗	✗	✗	✗	✗	✓	2.82	6.16	1.61	0.836	0.842	0.855
Cox 123(a)	✗	✓	✗	✓	✗	✓	2.49	5.50	1.56	0.838	0.835	0.854
Cox 1	✗	✗	✗	✗	✓		2.49	4.38	1.37	0.837	0.839	0.854
Cox 123	✓		✓		✓		3.79	3.56	2.43	0.835	0.835	0.853
Cox merged	✓						7.95	2.46	3.31	0.828	0.828	0.855

Table 3.5: Calibration and discrimination of Bayesian network, Dynamic Bayesian network, and Cox models in Stable, Rising, and Falling BP cohorts.

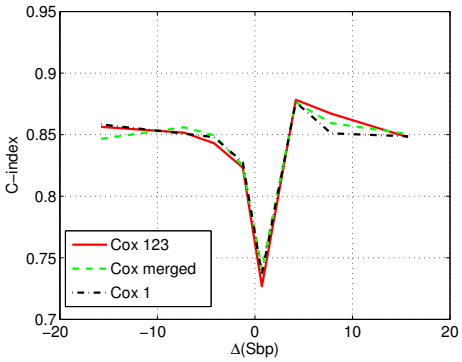
Table 3.4 summarizes model performance for the Stable, Rising, and Falling BP groups. In the Stable BP group, calibration is virtually identical across all the models. There are some differences in discrimination, however, with the best performing models being Bayes 123(a), 123(ab), 123, and merged, all of which incorporate the full three years of history (albeit in different ways). Note that Bayes merged, which averages values over the three year observation history of the patient, performs essentially equivalently to Bayes 123 and variants. This finding is unsurprising in the Stable BP cohort – when BP is stable, the Dynamic Bayes models do not gain any new information about blood pressure compared to Bayes merged. In the Rising BP cohort, the Dynamic Bayes models that capture trends in the values of BP outperform static models which use a single most recent or averaged value. The best performing dynamic model, Bayes 123, has better calibration and discrimination (k-statistic = 1.169, c-index = 0.895) than the best static model, Bayes merged (k-statistic = 4.5, c-index = 0.887). Bayes 1,

1(a), and 1(ab), which use only recent risk factor values, have the worst calibration and discrimination among the Bayes models. Interestingly, this gap in performance based on how historical risk factor values are included in the model does not appear for the Cox models. This may be because the Cox models do not contain interaction terms between the risk factors measured at different times. Results for the Falling BP cohort are similar to the Rising BP cohort: Calibration is good for all models, but the C-index is substantially higher for Bayes models which use more historical data. As in other cohorts, the use of historical data does not alter the performance of Cox models, and these models have lower C-indices than the Bayesian networks.

Figures 3.3a and 3.3b show the variation in the C-index as the threshold for three-year change in SBP varies continuously from -15 mmHg to +15 mmHg. Consistent with the pattern in Table 3.5, performance is better among groups with larger changes in SBP, and worse when SBP is stable over time. Note also the vertical separation between Bayes 123, Bayes merged, and Bayes 1 in Figure 3.3a which is absent in Figure 3.3b for the Cox models. As noted above, this supports the argument that the flexibility of the Dynamic Bayesian network allows it to use the information on longitudinal risk factor trajectories to improve accuracy, whereas the less flexible Cox model does not gain accuracy when historical information is included.



(a) Predictions using (Dynamic) Bayesian networks



(b) Predictions with Cox models

Figure 3.3: C-index by three-year pre-baseline change in SBP (Period 1 - Period 3) for three Bayesian network models and three Cox models.

3.6 Discussion

With detailed electronic health data now readily available at the point of care, there is a great opportunity to use this information to build risk models to predict a variety of clinical outcomes. However, these EHD remain "messy" and are not easily analyzed using traditional statistical methods. This paper presents a novel, general-purpose machine learning method for building clinical risk prediction models that can incorporate multiple historical risk factor values for each individual. Our method extends the standard Dynamic Bayesian network using inverse probability of censoring weighting to account for the fact that the clinical outcome of interest may be right-censored, which occurs frequently in practice. Missing values are automatically imputed as part of model fitting, so there is no need to use a separate imputation procedure to handle missing risk factor values. Using electronic health data from a large Midwestern health insurance provider, we applied our technique to predict cardiovascular risk over a 5-year period on the basis of up to three years of historical risk factor measurements. The Dynamic Bayesian networks which used multiple historical risk factor values outperformed both less flexible

Cox regression-based approaches and non-dynamic Bayesian network models which used only a single baseline value for each risk factor. The gains in prediction accuracy were largest for the subsets of individuals whose blood pressures trended up or down during the 3 years prior to baseline (as opposed to staying relatively constant). We therefore conclude that the sequential structure of the Dynamic Bayesian network is likely an important contributor to the observed improvements in model performance. For patients whose risk indicators do not vary substantially over time, a Dynamic Bayes model does not provide significantly better predictions than standard (i.e., non-dynamic) models. In fact, for these individuals, averaging historical measurements over time might be more advantageous. Our work focuses on prediction of a binary outcome since the vast majority of prediction models aim to estimate the probability that a particular event occurs (or not). However, the technique could be extended to categorical or continuous outcomes provided the requisite adjustments are made to distributional assumptions. The approach we present could also be applied to model risk for a wide variety of clinical outcomes.

Our work has several limitations. In our setting, we pre-specified the directed acyclic graph structure connecting the risk factors based on input from our clinical collaborators. This was feasible since the interplay between factors which affect cardiovascular risk is relatively well understood. For other clinical outcomes, and when possibly using novel biomarkers, the appropriate graph structure may be unclear. In ongoing work, we are developing techniques for constructing Dynamic Bayesian network DAGs in the presence of censored outcome data. We did not compare our proposed technique to other popular machine learning methods, e.g., support vector machines, regression trees, etc. Most of these techniques are not designed for use with right-censored outcomes, and hence would need to be modified to ensure a fair comparison; a recent paper [111] describes how these modifications might be carried out. The Cox regression models we did make comparisons to were relatively inflexible, including only main effects terms for the historical risk factors. Better accuracy might have been achieved if, e.g., we had used higher-order terms such as interactions and/or regression splines. However, we note that many popular cardiovascular risk prediction models (including the Framingham

Risk Score) are based on very simple Cox models including only main effects, so our comparisons have practical relevance.

The dimension of model performance which seemed to be impacted the most by the use of Dynamic Bayesian networks was model discrimination. While the C-indices of the Dynamic Bayesian network models were frequently higher than competing approaches, they were high (> 0.8) for all models. So, while Dynamic Bayesian networks provides more accurate risk predictions in a statistical sense, we did not address the question of whether these improvements are clinically meaningful. More broadly, there are multiple complex issues to consider (most of which are beyond the scope of this paper) when contemplating whether to use DBNs to predict risk as part of an EHD-backed clinical decision support system.

Though Bayesian networks are often associated with analyses that seek to establish the causal relationships between variables, we use the Dynamic Bayesian network DAG structure only to represent conditional independence relationships between variables, and make no causal claims about the resulting model. In particular, it would be inappropriate to use our proposed technique (or any risk prediction method which estimates cross-sectional associations between risk factors and outcomes) to, e.g., estimate the reduction in risk that would be achieved by starting a blood pressure medication or decreasing LDL by 20 points. Decision support tools which aim to help guide treatment decisions by predicting their potential impacts should use statistical techniques which move beyond simple prediction of future risk to explicitly estimate the causal effects of interventions are behavior changes. Few of these techniques have been implemented in primary care settings thus far, but we anticipate that historical risk factors will play an important role in improving their accuracy as well.

Chapter 4

A data driven approach to optimize Bayesian networks for EHR data

4.1 Overview

One of the primary reasons of using Bayesian networks to construct predictive risk models in medicine has been their ability to use data with missing features, handle nonlinear and complex interactions, and yet remain relatively easy to interpret. While it is possible to hand-craft (i.e., manually define a network structure) Bayesian network models consisting of relatively few risk factors for a homogeneous population comprising of largely healthy people without previous comorbidities such as heart disease or diabetes, determining relationships between risk factors is less practical when a large number of risk factors are used to construct a predictive model for a heterogeneous population cohort that includes individuals diagnosed with heart disease or diabetes or who have previously experienced a CV event. Further, hand crafting the structure of a Bayesian network that may not include all relevant relationships in a diverse cohort, because relationships that experts tend to model are typically a product of studies of carefully controlled cohorts whose characteristics might differ significantly from the cohort under

study; additionally, experts typically tend to think in terms of causality which may not indicate conditional dependencies which are represented by the edges in a Bayesian networks. These factors result in hand crafted Bayesian networks not being optimal.

The space of all possible networks increases as a factorial of the number of nodes (risk factors) in the model; as a result, it is not possible to exhaustively search the network space for an optimal network structure in settings where the number of potential risk factors is large. Therefore, we have to employ heuristic techniques to determine optimal Bayesian network structures that produce predictive models with high calibration and discrimination performance. In this chapter, we introduce a novel scoring function based on a the models discrimination and calibration for a hill-climbing heuristic, combined with simulated annealing to determine an optimal Bayesian network structure.

The results in this chapter is currently being prepared for submission to a journal. This work includes contributions from Julian Wolfson, David Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Paul Johnson, and Patrick O'Connor. Dr Patrick O'Connor provided us with medical insights that helped us design the model. This work was supervised by Paul Johnson, Gediminas Adomavicius Julian Wolfson and David Vock.

4.2 Background

Accurate prediction models that determine risk of cardiovascular events, such as heart attack and stroke, are important tools that can guide medical practitioners in providing treatment that helps reduce fatalities, especially for complex patients who have multiple uncontrolled risk factors. The most widely used prediction models are regression based and have several drawbacks, such as inability to handle missing data natively and model non monotonic relationships between risk factors. Trying to construct a regression model that can predict CV risk for a large heterogeneous population and that can resolve non linearities will include interaction terms making the model difficult to interpret, thus negating one of its features that make them appealing to medical practitioners.

In the Chapters 2 and 3, we have demonstrated that Bayesian networks can be

used to construct CV risk models using Electronic Health Record (EHR) data that represents a heterogeneous cohort. These models can outperform traditional regression models, they can represent non-monotonic relationships between risk and risk factors and are yet relatively easy to interpret. In fact, the Bayesian networks in used Chapters 2 and 3 were developed based on consultations with physicians, where each directed edge in the graphical model indicated a causal relationship known or hypothesized in medical literature between two or more risk factors. These relationships where either uncovered in separate clinical studies or were part of the physician's experiential knowledge.

Bayesian networks constructed based on expert knowledge have been widely used to construct prediction models for monitoring patients in intensive care [13], diagnosing oesophageal cancer [106], mammography [20], and diagnosing liver disorder[77]. Such approaches to construct a network require access to an expert which may be time consuming and further assume that the expert is aware of all possible relationships between the risk factors. Alternatively, one can learn network structure from the underlying data; an approach that has been used to construct models of emergency medical service [1], tuberculosis epidemiology [41], and diagnostics for endocrinology and lymphoma [55].

4.3 Structure learning for Bayesian networks

Structure learning techniques find a Directed Acyclic Graph (DAG) representing the Bayesian network that best fits the data that the Bayesian network represents. There are three types of structure learning techniques [59]: search and score techniques, where the algorithm tries to find the structure that maximizes the score function; constraint based methods, where constraints on the graph are defined based on conditional independence relationships that are discovered in the data; and hybrid algorithms that use both constraint based and search and score techniques. We briefly overview these techniques in the remainder of this section.

4.3.1 Search and score

Search and score algorithms operate by evaluating a score for different structures and selecting the structure with the highest score. There are several possible scoring functions, some of which assign a score based on how well the individual joint distribution terms $P(x_i|\text{Pa}(x_i))$ from Equation (2.2) fit the underlying data $x_i, \text{Pa}(x_i)$, where x_i is the i^{th} variable (in our case the i^{th} risk factor) and $\text{Pa}(x_i)$ are the parent node(s) of x_i. These scores are usually the product of log likelihoods of all the joint distributions in the network with a penalty for the number of edges in the network [32]. Scoring functions corresponding to a network structure can also be determined based on the predictive performance the model using the structure to evaluate predictions made on a data validation set. While finding the structure that maximizes the score is an optimal solution, it is not usually achievable because the number of distinct structures increases as a factorial ($O(n!)$) with the number of nodes. Therefore to find a structure that is close to optimal, we employ various heuristic search techniques, such as: Sparse candidate algorithm, Optimal reinsertion and Greedy Search.

Sparse candidate algorithm

The sparse candidate algorithm [39] identifies a relatively small number of candidate parents for each variable based on simple local statistics, such as mutual information or correlation. This construction creates a small subset of all possible graph structures from which the highest scoring structure is assumed to be optimal. This algorithm may not find the global optimum because the local independence assumption may not always be valid when we take into account other variables. However, its advantage lies in the fact that selecting a "configurable" number of parents of a node allows one to greatly reduce the search space, thus allowing to quickly generate structures for systems with a large number of variables. This technique of structure search has been extensively used to model gene expression networks [40].

Optimal reinsertion

The optimal reinsertion algorithm [72] is an iterative search-and-score algorithm that at each step chooses a target node and deletes all edges associated with it. The algorithm then inserts edges between the target node and its parents and its children such that maximizes a score associated with the target node and nodes in its Markov blanket i.e., the set of nodes consisting of the target node's parents, children, and children's other parents. The insertion strategy can be combined with the sparse candidate algorithm which restricts the search space for insertion.

Greedy search

The Greedy Search algorithm [23] is an iterative algorithm that starts with a graph with no edges. Every iteration of this algorithm consists of a "forward phase" and a "backward phase". During the forward phase, this algorithm tries adding exactly edge from every node to other remaining nodes as long as the addition of the edge does not create a cycle. This phase results in the selection of the graph which results from the addition of the edge which led to the highest increase of the scoring metric. If none of the additions lead to an increase in score, no edges are added at the end of this phase .i.e. the graph remains unchanged. The "backward phase" scores all the graphs that results from the removal of exactly one edge the graph resulting from the "forward phase". This phase terminates when a resulting graph that leads to maximum increase in score is found. These two steps are repeated until no changes can be made that increases the score.

4.3.2 Constraint based methods

Constraint based techniques for structure learning proceed in three phases: (1) learning the Markov blanket of each node, (2) learning the neighbors of the nodes, and finally (3) learning the edge directions between the nodes. To learn Markov blanket $B(X_i)$ of random variable (node) X_i, all pairs of nodes X_j, X_k (where $j, k \neq i$ and $X_j \in B(X_i)$) are examined to check if X_i is independent of X_k given X_j. The Markov blanket is

then pruned by verifying that the membership of the Markov blanket is symmetric .i.e. $X_i \in B(X_j) \implies X_j \in B(X_i)$. The second phase learns the neighbors of each node X_i. This is achieved by testing if X_i and X_j are independent given all other nodes and when $X_j \in B(X_i)$. If the two nodes are not independent, then an "undirected" edge is placed between the nodes. The third phase of this algorithm consists of learning the direction of the "undirected" edge. Constraint based methods, unlike search and score methods are not iterative and therefore terminate after a definite number of steps. Though this may result in structures that are not as representative of the data as structures obtained using search and scoring techniques especially for small data samples, they are well suited to determine structures in situations where there are very large number of variables.

4.3.3 Hybrid methods

Hybrid algorithms combine constraints with search and score algorithm. These algorithms first construct a "skeleton" network which may be derived from expert input followed by greedy hill climbing heuristic to add additional edges to find an optimal network [104].

4.4 Structure learning using EHR data

Let $\boldsymbol{X} = X_i \cdots X_N$ represent the vector of observed values for risk factors such as blood pressure, smoking, age, etc. which predicts the risk of occurrence of cardiovascular (CV) event, and E be an indicator variable that represents the occurrence of a CV events τ years after the baseline (defined in section 2.4). The risk model tries to estimate $P(E = 1|\boldsymbol{X})$, which can be expressed using Bayes theorem as follows:

$$P(E = 1|\mathbf{X}) = \frac{P(\mathbf{X}|E = 1)P(E = 1)}{\sum_{e\in\{0,1\}} P(\mathbf{X}|E = e)P(E = e)}, \tag{4.1}$$

The joint distributions of $\mathbf{X}|E = e$ can be represented using a directed acyclic graph (DAG) which indicates conditional independence relationships between variables $X_i \in$

$\boldsymbol{X}$ as a result joint distribution to be decomposed into a product of individual terms conditioned on their parent variables [100]:

$$P(\boldsymbol{X}|E) = \prod_{i=1}^{N} P(X_i|\text{Pa}(X_i), E) \tag{4.2}$$

where $\text{Pa}(X_i)$ are the parents of X_i. The goal of learning the structure of the Bayesian network is to determine the parent nodes of all X_i's that maximizes a score based on the model accuracy. Computing the joint distributions $P(X_i|\text{Pa}(X_i), E)$ for EHR data is not straight forward because the outcome variable E is not always observed as this data is often right censored. To estimate joint distributions of right censored data without introducing significant bias, we weigh the observations by the inverse of the probability that the observation is right censored. This approach of weighing observation is called inverse probability of censoring weight . The application of IPCW to learn distributions from right censored data is discussed in detail in section 2.3.5. The joint distributions $P(X_i|\text{Pa}(X_i), E)$ is represented as a mixture of normals as discussed previously in section 3.3.3.

4.4.1 A greedy search algorithm for learning Bayesian network structure

Optimizing the structure of a Bayesian network using a greedy search strategy requires the maximization of a function that is representative of the predictive performance of the Bayesian network. One way of quantifying the predictive performance is by using a combination of quantities that are derived from the calibration and discrimination performance of the model. The scoring function that we use (Equation 4.3) to learn the structure of Bayesian network consists of a linear combination of a calibration and a discrimination term.

Let us assume that the structure of a Bayesian network which we are trying to optimize consists of N random variables $X_i \in \boldsymbol{X}$. An edge in this network from the random variable (node) X_i to X_j is indicated by E_{ij}. The optimal structure consists of

edges $E_{ij} \in E$ which maximizes the scoring function $F(\boldsymbol{X}, E)$ defined as:

$$F(\boldsymbol{X}, E) = p(2/(1 + \exp(K/p_s)) + (1 - p)(C - p_c)/(1 - p_c) \tag{4.3}$$

The calibration term in the scoring function is: $2/(1 + \exp(K/p_s)$. This term maps the calibration statistic K (described in section 2.3.9) to $[0, 1]$. The parameter p_s in this term is used to define a range in which the score F is sensitive to variations in calibration. The discrimination term in the scoring function; $(C - p_c)/(1 - p_c)$ maps the concordance C (described in section 2.3.9) to $[0, 1]$ using a scale factor p_c. The parameter p in Equation 4.3 is used the tune the contribution of the calibration and discrimination terms to the over all score. To evaluate this score the calibration and concordance metrics are determined using the training data which is also used to calculate the joint distributions of the random variables $\{X_i, \mathrm{Pa}(X_i)\}$.

We adapt the greedy search algorithm [23] to find an optimal network structure by adding a move to reverse edges and incorporating a simulated annealing process to avoid converging to a locally optimal solution. This algorithm starts with the set $E = \emptyset$. For every variable we try adding an edge e_{ij} between X_i and X_j. If adding the edge doesn't introduce a cycle in the graph and increases $F(\boldsymbol{X}, E)$ then the edge is retained in the network. Following the $N(N - 1)$ addition we delete edges from the graph, if the deletion results in increasing the score. In addition to adding and deleting edges, we also reverse edges if it does not introduce a cycle and leads to an increase in the score. This set of $3N(N - 1)$ moves corresponds to a single iteration. We continue iterating till we arrive at a point where two successive iterations does not lead to a score change. To prevent graph structure from converging to a local minimum we randomly permute the order in which the edge changes are attempted in each iteration, further, we employ a simulated annealing scheme where we accept a sub optimal edge change (addition, deletion or reversal) with a probability of $\exp(-n/T)$, where n is the number of edge changes attempted, and T is a decay constant. This algorithm is described in details in Algorithm 2.

Algorithm 2 Greedy structure search algorithm

1: $E \leftarrow \emptyset$

2: $n \leftarrow 0$
3: **while** $F(X, E)$ changes **do**
4: $\quad \boldsymbol{I} \leftarrow$ random permutation$(\mathbf{1} \cdots \boldsymbol{N})$
5: $\quad \boldsymbol{J} \leftarrow$ random permutation$(\mathbf{1} \cdots \boldsymbol{N})$
6: $\quad$ **for all** (i, j) such that $i \in I$ **and** $j \in J$ **do** **Add edges**
7: $\quad\quad$ **if** $E \cup e_{ij}$ is a valid graph **then**
8: $\quad\quad\quad n \leftarrow n + 1$
9: $\quad\quad\quad$ **if** $F(E \cup e_{ij}, \boldsymbol{X}) > F(E, \boldsymbol{X})$ **OR** uniform random$(0, 1) < \exp(-n/T)$ **then**
10: $\quad\quad\quad\quad E \leftarrow E \cup e_{i,j}$
11: $\quad\quad\quad$ **end if**
12: $\quad\quad$ **end if**
13: $\quad$ **end for**
14: $\quad$ **for all** (i, j) such that $i \in I$ **and** $j \in J$ **do** **Delete edges**
15: $\quad\quad$ **if** e_{ij} exists **then**
16: $\quad\quad\quad n \leftarrow n + 1$
17: $\quad\quad\quad$ **if** $F(E \setminus e_{ij}, \boldsymbol{X}) > F(E, \boldsymbol{X})$ **OR** uniform random$(0, 1) < \exp(-n/T)$ **then**
18: $\quad\quad\quad\quad E \leftarrow E \setminus e_{i,j}$
19: $\quad\quad\quad$ **end if**
20: $\quad\quad$ **end if**
21: $\quad$ **end for**
22: $\quad$ **for all** (i, j) such that $i \in I$ **and** $j \in J$ **do** **Reverse edges**
23: $\quad\quad$ **if** e_{ij} exists **then**
24: $\quad\quad\quad n \leftarrow n + 1$
25: $\quad\quad\quad$ **if** $F((E \setminus e_{ij}) \cup e_{ji}, \boldsymbol{X}) > F(E, \boldsymbol{X})$ **OR** uniform random$(0, 1) < \exp(-n/T)$ **then**
26: $\quad\quad\quad\quad E \leftarrow (E \setminus e_{i,j}) \cup e_{ji}$
27: $\quad\quad\quad$ **end if**
28: $\quad\quad$ **end if**
29: $\quad$ **end for**
30: **end while**

In the above implementation of the algorithm, the T parameter of the simulated annealing process is set to $\approx N^2$, which ensures that most of the random perturbations occur in the first few iterations of the algorithm. For predictive models that are used in the context of health care, it is important that the model be well calibrated therefore we set $p = 0.5$ in equation 4.3 .i.e. equal weight to both calibration and discrimination. Further we set $p_s = 400$ so that the score is relatively sensitive for a large calibration range, and we set $p_c = 0.5$ in equation 4.3 because the lowest possible discrimination is 0.5.

4.4.2 Data description

Our study was conducted utilizing the HMO Research Network Virtual Data Warehouse (HMORN VDW) from a healthcare system from the Midwestern United States. The VDW stores data in standardized data structures including insurance enrollment, demographics, pharmaceutical dispensing, utilization, vital signs, laboratory, census and death records. These data are obtained from both the EMR and insurance claims. 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. This initial selection identified 448,306 subjects. A description of selection criteria that were used to select individuals for inclusion into the cohort for training the model, the risk factors and a description of events can be found in Section 2.4

4.4.3 Study design for learning network structure

The network structure of a Bayesian network that is used to predict CV risk of a relatively homogeneous population can be defined based on expert knowledge of the dependencies between risk factors primarily because the variability of the population is characterized by relatively few risk factors. However for more heterogeneous (complex) populations with larger number of risk factors, using expert information to construct a network becomes difficult. To validate this hypothesis we have constructed cohorts

with different degrees of heterogeneity (complexity) characterized by different risk variables. These cohorts are constructed by putting together population subgroups that by themselves are relatively homogeneous as shown in table 4.1. Cohorts are composed of three subgroups:

1. The "healthy" subgroup that is primarily composed of individuals whose risk indicators are in the normal range. A small fraction of individuals in this subgroup may have elevated risk factors but they are usually medicated and not diagnosed as having heart failure or peripheral arterial disease.

2. The "Comorbidities" subgroup may have had CV events such as heart attack or stroke in the past or, they have been diagnosed with heart failure. The individuals in this sub group are almost always medicated and have a significantly higher CV risk compared to the "healthy subgroup".

3. The "Diabetic" subgroup is composed of individuals who have diabetes. The CV risk of this subgroup depends on A1c in addition to the risk factors (table 4.2).

The composition of each of the cohorts are shown in table 4.1, where, a ✓ indicates that a particular subgroup is included in the cohort while a ✗ indicates an absence of the subgroup from the cohort.

"Simple" cohorts

Cohort names	Sub groups			Descriptions
	"healthy"	comorbidities	diabetics	
Cohort-H	✓	✗	✗	Relatively healthy population without diabetes & without previous CV events or diagnosis
Cohort-C	✗	✓	✗	Individuals with comorbidities
Cohort-D	✗	✗	✓	Individuals with diabetes
Composite "Complex" cohorts				
Cohort-HC	✓	✓	✗	Cohort-H ∪ Cohort-C
Cohort-HCD	✓	✓	✓	Cohort-H ∪ Cohort-C ∪ Cohort-D

Table 4.1: Description and composition of cohorts with different variability.

To compare the difference between expert defined structure and data driven structure definitions of Bayesian networks we have constructed several models using different risk factors from Cohorts of different complexities as seen in table 4.2. In addition to listing the cohort used to train the model, this table also lists the risk factors that the model includes. For example the models BNx-H-v1 and BNx-HCD-v2 are constructed from cohorts "Cohort-H" and "Cohort-HCD" , which is significantly more heterogeneous and is described by more risk factors. In addition to studying the effect of cohort complexity on model building, we also would like to study how expert defined models perform when there are more risk factors. A comparison of the models in terms of their training cohort complexity and variable complexity is displayed in Figure 4.1. In this figure (4.1), we see, BNx-H-v1 and BNx-H-v2 are both constructed using Cohort-H, however, BNx-H-v2 makes use of three risk factors LDL, HDL and Triglycerides to model lipids while, BNx-H-v1 uses only cholesterol (which is a linear combination of the 3 risk factors). The use of two additional risk factors in BNx-H-v2 can complicate the construction of

the network because the relationship between the lipid components and the other risk factors such as blood pressure is not clearly established [17]. As a result network created by experts may not reflect the true relationships that exist within the cohort resulting in poor model performance. Similarly models such as BNx-HCD-v2 which are trained on cohorts more heterogeneous than that used to train BNx-H-v2 or BNx-HC-v2. In addition to the cohort (Cohort-HCD) being more complex, it also includes additional risk factors such as A1c levels, and indicators of previous conditions. It may be difficult for an expert to specify the network for a model constructed using this cohort because the relationships between all the risk factors under different conditions (such as for people having diabetes) may not be well understood.

Models	Training cohort	Risk factors											
		age	gender	Comorbidity	has Diabetes	SBP	SBP Med Count	Smoking	Cholesterol	LDL	Trig	HDL	A1C
BNx-H-v1	Cohort-H	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
BNx-H-v2	Cohort-H	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓	✓	✗
BNx-HC-v2	Cohort-HC	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
BNx-HCD-v2	Cohort-HCD	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓

Table 4.2: Risk factors and cohorts that are used to construct the different Bayesian networks. The ‘x’ in the model names stand for either ‘e’ or ‘d’, where ‘e’ indicates a model with an expert defined network and ‘d’ indicates one where the network is learned from the data.

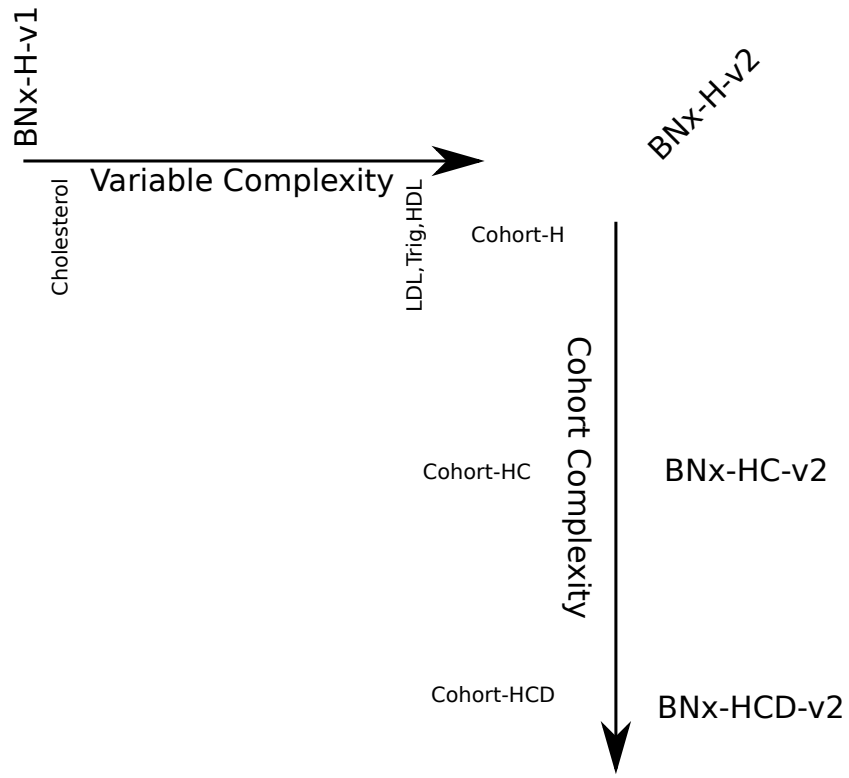


Figure 4.1: Comparison of variable and cohort complexity of the models being studied.

4.5 Results

There is a clear difference between the performance of the models whose structures are determined by experts and those whose structures are discovered using the greedy search algorithm. As shown in table 4.3, the data driven approach to defining the structure is clearly superior to the expert defined model structure for all subgroups and cohorts. However the degree by which the models vary offers insight into the situations where structure learning makes a difference.

The calibrations and discrimination of BNe-H-v1 and BNd-H-v1 are almost identical because the two networks are trained using a relatively homogeneous "easy to learn" cohort which consists of primarily healthy individuals. Moreover, BNx-H-V1 uses a single risk factor (cholesterol) to model the effect of lipids on risk thus making it relatively

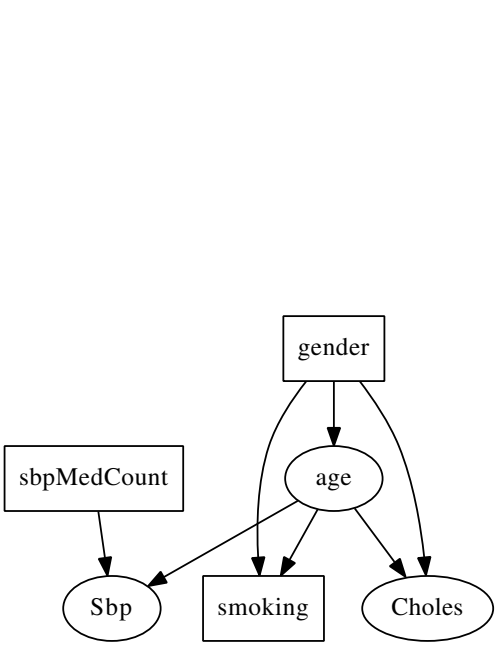
simple to manually specify relationships between risk factors.

Models	Cohorts and subgroups		
	Cohort-H	Cohort-HC	Cohort-HCD
	k-statistic		
BNe-H-v1	9.76	-	-
BNd-H-v1	9.21	-	-
BNe-H-v2	18.02	-	-
BNd-H-v2	5.57	-	-
BNe-HC-v2	47.43	127.13	-
BNd-HC-v2	13.69	2.96	-
BNe-HCD-v2	42.71	101.22	303.54
BNd-HCD-v2	10.73	3.21	2.43
	c-Index		
BNe-H-v1	0.869	-	-
BNd-H-v1	0.873	-	-
BNe-H-v2	0.867	-	-
BNd-H-v2	0.872	-	-
BNe-HC-v2	0.864	0.892	-
BNd-HC-v2	0.871	0.898	-
BNe-HCD-v2	0.859	0.886	0.880
BNd-HCD-v2	0.871	0.895	0.891

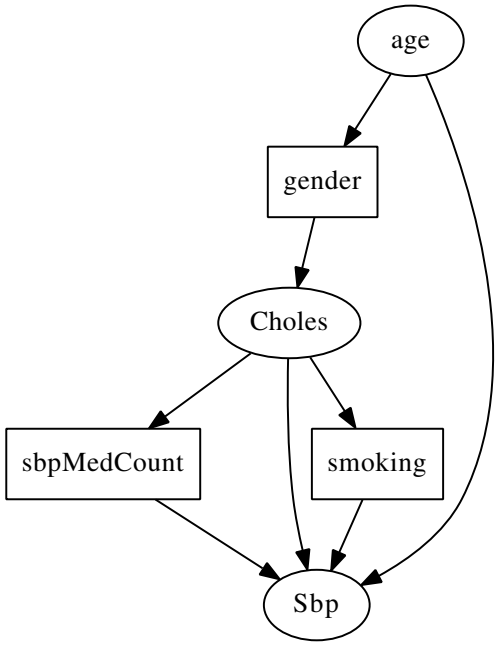
Table 4.3: Model performance for different population subgroups and cohorts

The calibration error of BNe-H-v2 is 18.02 while that of BNd-H-v2 is significantly lower at 5.57 in-spite of both being trained on Cohort-H. We think this difference is due to fact that BNx-H-v2 is trained using LDL, HDL and Triglycerides instead of just cholesterol. The presence of these additional risk factors lead to inaccuracies in the expert defined network because a relationship between these risk factors is not clearly

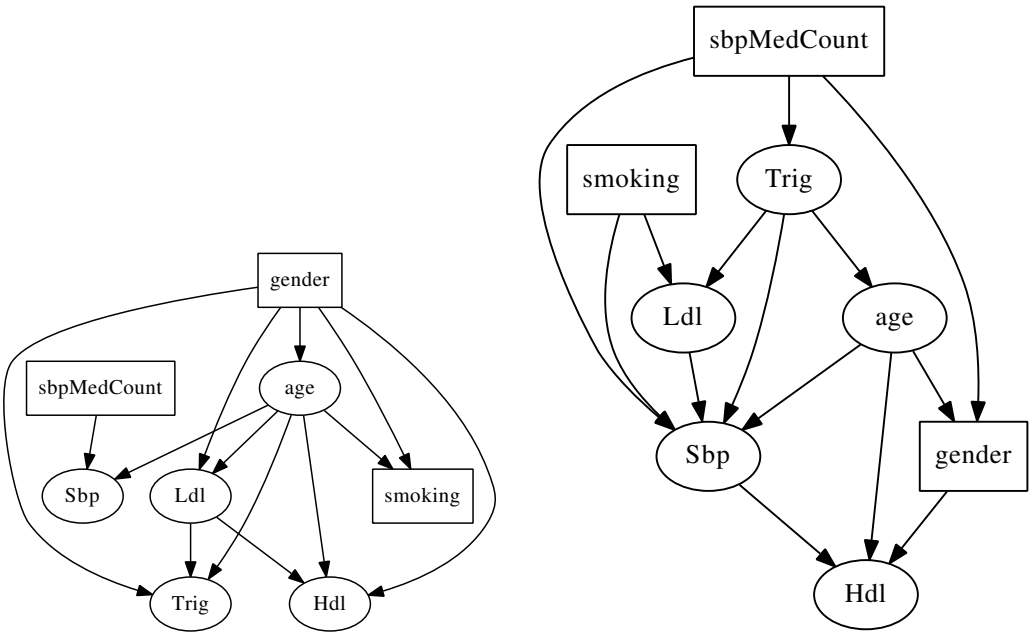
known. The difference between the relationship of the lipids between the expert defined and the data-driven network is clearly visible in figures 4.2d and 4.2c. On the other hand the differences between the networks corresponding to BNx-H-v1 are less pronounced (figures 4.2b and 4.2a).



(a) Expert defined network for BNe-H-v1



(b) data defined network for BNd-H-v1

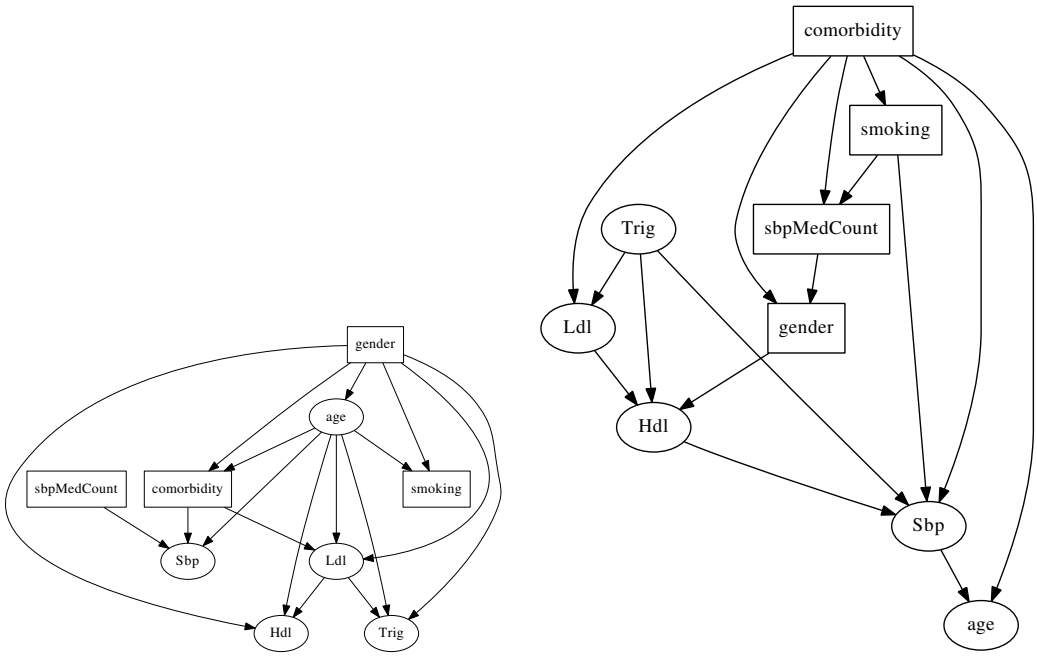


(c) Expert defined network for BNe-H-v2 (d) data defined network for BNd-H-v2

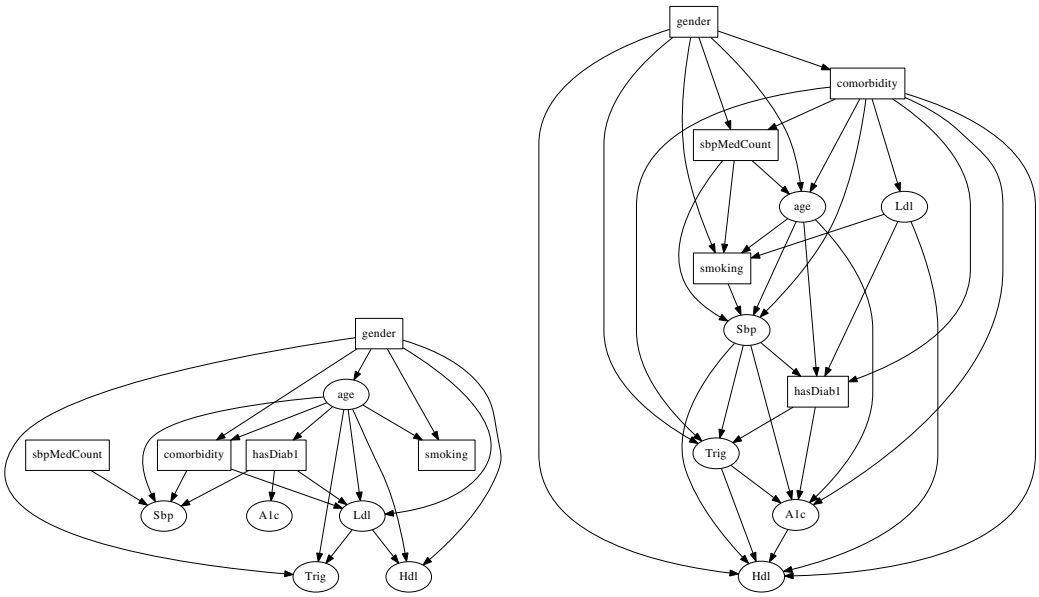
Figure 4.2: A comparison of expert defined network structure with the corresponding data defined network structure using a greedy search algorithm for Bayesian networks trained on cohort-H, a relatively homogeneous cohort consisting primarily of health individuals

The performance difference between the expert defined and the data derived models are far more pronounced when they are trained on cohorts which are constructed using multiple dissimilar subgroups. For example BNd-HCD-v2 which is trained on Cohort-HCD, a combination of three cohorts, has a calibration error of 2.43 for Cohort-HCD while the the expert defined BNe-HCD-v2 has a much higher calibration error of 303.54. BNd-HCD-v2 has good calibration across all risk ranges as we see in figure 4.4d. The discrimination of the model with data defined network (BNd-HCD-v2) is 0.891 which is significantly better than that of the expert defined model. Further, the calibration and the discrimination of BNd-HCD-v2 is better than that of BNe-HCD-v2 across all cohorts and subgroups even when the subgroup is relatively homogeneous. For example

the c-Index of BNd-HCD-v2 computed for Cohort-H is 0.871, which is significantly higher than that of BNe-HCD-v2 at 0.859. This is probably because the risk factors in the BNe-HCD-v2 network have not been conditioned correctly, as we see in figures 4.3d where the indicator variables "comorbidity" and "has Diabetes" are used to condition many more risk factors such as blood pressure, Triglycerides, etc compared to the BNe-HCD-v2 network (figure: 4.3c). In summary, more variables in the model provides a greater opportunity for an expert to make incorrect assumptions about relationships between risk factors.

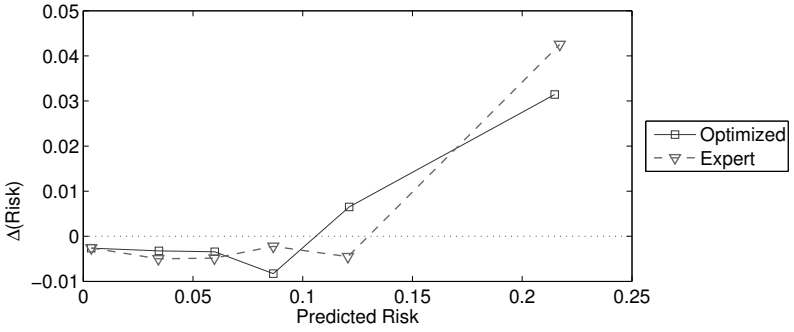


(a) Expert defined network for BNe-HC-v2 (b) data defined network for BNd-HC-v2

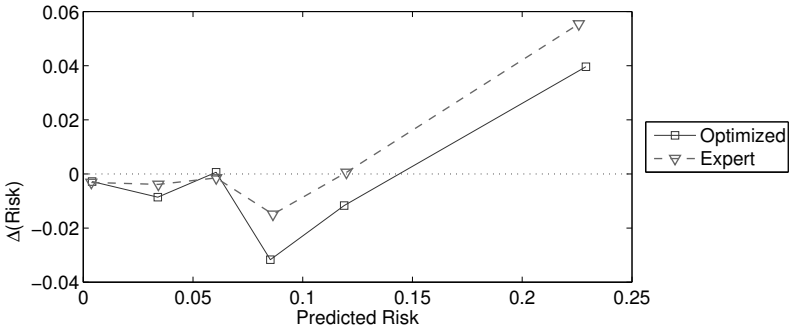


(c) Expert defined network for BNe-HCD-v2 (d) data defined network for BNd-HCD-v2

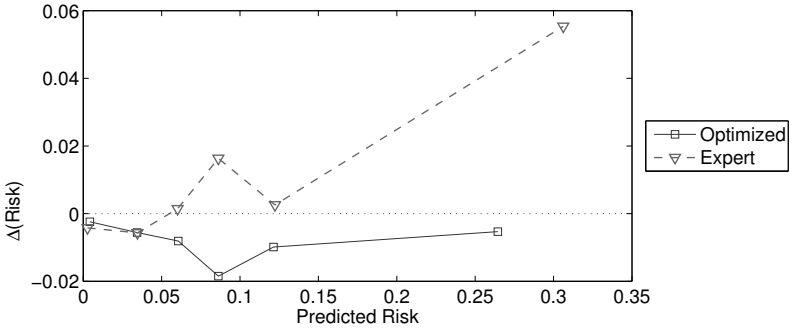
Figure 4.3: A comparison of expert defined network structure with the corresponding data defined network structure using a greedy search algorithm for Bayesian networks trained on heterogeneous cohorts, cohorts-HC and cohort-HCD



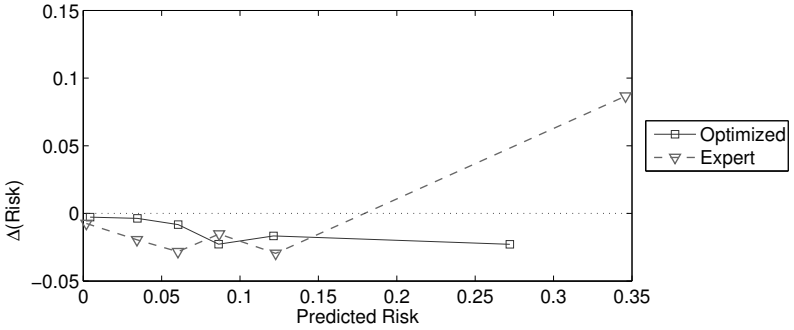
(a) Calibration: BNx-H-v1 on Cohort-H



(b) Calibration: BNx-H-v2 on Cohort-H



(c) Calibration: BNx-HC-v2 on Cohort-HC



(d) Calibration: BNx-HCD-v2 on Cohort-HCD

Figure 4.4: Comparison of calibration of the Bayesian networks with expert defined network structure against that of the the respective data defined network structure

4.6 Conclusion

We have seen that the ability of Bayesian networks to predict CV risk with sufficient accuracy largely depends on how well the independence assumptions i.e. the network structure reflects the relationships between the risk factors of the population for which the CV risk is being predicted. Experts are often able to define the relationships if there are relatively few risk factors; such models can be used only for relatively homogeneous populations. With more complex populations or in situations where a population is represented by large number of risk factors it becomes difficult for an expert to define the network because the relationships between risk factors might not have been studied or it maybe outside the experts domain of knowledge. Additionally even if relationship between certain risk factors have been studied, it might not be representative of the population for which the predictive model is being constructed.

A data driven approach to discover the network structure is unaffected by the gaps in knowledge and can determine relationships between risk factors specific to the population under study. We have used a greedy algorithm to determine the network structure that minimizes the prediction error of a Bayesian network. The Bayesian network derived from data is as good as one defined by experts for risk prediction involving relatively few risk factors in a homogeneous cohort. For more complex populations with more risk factors where there are a larger number of possible interactions among the risk factors, the data driven approach to defining the network structure of a Bayesian network is significantly more accurate than an expert defined Bayesian network structure.

Chapter 5

Conclusion and future work

Electronic health records are maintained by almost all organizations that provide health care, leveraging this data to provide additional information regarding patients risk to-health care providers can improve the quality of care a patient receives. This information is particularly useful for physicians to prioritize treatment strategies for complex patients because this system lets physicians evaluate different potential treatment options based on the potential reduction in risk and select a strategy that is optimal for a given individual

Bayesian networks are a powerful machine learning technique that can be used to construct predictive models for diverse applications. In this thesis we have shown how Bayesian networks can be applied to construct predictive models to estimate risk of cardiovascular events by using electronic health record data. However there are several challenges to applying traditional machine learning techniques, including Bayesian networks to this data because the outcome of the patient is not always known, and there is a large fraction of data that is not recorded in the EHR. We have applied inverse probability of censoring weights to train Bayesian networks on this censored data and have used a novel model averaging technique to prevent over fitting for Bayesian networks. We have also demonstrated that Bayesian networks can be used to construct risk predictive models from EHR data that outperform traditional regression based survival models and are flexible enough to recover non monotonic relationships between the risk

of CV events and its risk factors, while still retaining interpretability, which makes it appealing to medical practitioners.

The risk of CV events, in addition to current state of an individual, also depends on the person history. To make use of the historical data which is well represented in EHR data we extended the standard Dynamic Bayesian network model using inverse probability of censoring weight to account for the fat that observations of events might be censored. The Dynamic Bayesian model network which used multiple historical risk factors outperformed both less flexible Cox regression based approaches and non-dynamic Bayesian network models which used only a single baseline value for each risk factor. The increase in prediction accuracy was most prominent in patient subgroups whose blood pressure trended up or down during the period prior to the baseline. For patients whose blood pressure did not change appreciably during this time, the Dynamic Bayesian model does not provide significantly better predictions compared to non-dynamic Bayesian network models. For these patients, we discovered that it might be more advantageous to average of all their historical measurements and use a non-dynamic model because averaging over a large time period reduces the chance that a particular variable may be missing at baseline.

One of the reasons why prediction models using Bayesian networks are popular in medicine is due to their interpretabily and their ability to easily incorporate expert knowledge into the learning process. However, in situations where the underlying population is diverse or where there are large number of variables characterizing the population, it may not be possible for a modeler to rely solely on an expert to define the relationships between risk factors. To overcome this limitation in applying Bayesian network models to complex populations, we made use of a greedy search heuristic that starts with an empty network and gradually populates it based on a "goodness of fit" score. This approach to building a network solely based on data characteristics significantly outperforms expert defined networks for complex populations. For homogeneous populations having few risk factors, the difference between the data derived Bayesian network model and the expert defined Bayesian network model is not significant. In addition, the networks generated by the heuristic contains partial structures which are

similar to structures that would be defined by experts.

Throughout this thesis, we have made use of inverse probability of censoring weights to adapt Bayesian networks and their variants to learning from EHR data. These weights assume that the probability of an individual being censored is independent of the characteristics of the patient, which we know is not entirely accurate. Individuals who are more likely to need health care support are less likely to be censored. This assumption can be relaxed by using some of the risk factors to predict the probability of censoring, which might lead to more accurate predictions. IPCW can be extended to other machine learning classifiers. It would be interesting to compare the performance of Bayesian networks to other classifiers such as regression trees, random forests. etc. Finally, we have used only a small subset of the data that is present in EHR. We have condensed variables like medications to a count, physician diagnoses to an indicator variable and have ignored risk factors which were not commonly reported. In future, the data driven approach to defining the network could also include a data driven approach to variable selection for the network. It would be interesting to see if we can develop new data features based on prescriptions and diagnosis codes recorded in the EHR to further improve the accuracy of risk predictions. Using these features might let us predict risk of different types of events. Developing advanced and accurate predictive models will boost the importance of using machine learning techniques in the medical community which will ultimately lead to better and more personalized treatment for individual patients.

References

[1] Silvia Acid, Luis M de Campos, Juan M Fernández-Luna, Susana Rodrıguez, José Marıa Rodrıguez, and José Luis Salcedo. A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service. *Artificial intelligence in medicine*, 30(3):215–232, 2004.

[2] Israel T. Agaku, Brian A. King, and Shanta R. Dube. Vital signs: current cigarette smoking among adults aged $\geq$ 18 years–united states, 2005-2010. *MMWR. Morbidity and mortality weekly report*, 60(35):1207, 2011.

[3] Norrina Allen, Jarett D Berry, Hongyan Ning, Linda Van Horn, Alan Dyer, and Donald M Lloyd-Jones. Impact of blood pressure and blood pressure change during middle age on the remaining lifetime risk for cardiovascular disease: The cardiovascular lifetime risk pooling project. *Circulation*, 125(1):37–44, 2012.

[4] Norrina B Allen, Juned Siddique, John T Wilkins, Christina Shay, Cora E Lewis, David C Goff, David R Jacobs, Kiang Liu, and Donald Lloyd-Jones. Blood pressure trajectories in early adulthood and subclinical atherosclerosis in middle age. *JAMA*, 311(5):490–497, 2014.

[5] S. Andreassen, C. Riekehr, B. Kristensen, HC Schonheyder, and L. Leibovici. Using probabilistic and decision-theoretic methods in treatment and prognosis modeling. *Artif Intell Med*, 15(2):121–134, FEB 1999.

[6] Steen Andreassen, Roman Hovorka, Jonathan Benn, Kristian G Olesen, and Ewart R Carson. *A model-based approach to insulin adjustment*. Springer, 1991.

[7] G. Assmann, P. Cullen, and H. Schulte. 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):900–900, 2002.

[8] Sunayan Bandyopadhyay, Julian Wolfson, David M Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Mohamed Elidrisi, Paul E Johnson, and Patrick J OConnor. 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*, 29(4):1033–1069, 2015.

[9] H. Bang and A. A. Tsiatis. Estimating medical costs with censored data. *Biometrika*, 87(2):329–343, 2000.

[10] H. Bang and A. A. Tsiatis. Median regression with censored cost data. *Biometrics*, 58(3):643–649, 2002.

[11] William E Barlow, Emily White, Rachel Ballard-Barbash, Pamela M Vacek, Linda Titus-Ernstoff, Patricia A Carney, Jeffrey A Tice, Diana SM Buist, Berta M Geller, Robert Rosenberg, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, 2006.

[12] Steven L Barriere and Stephen F Lowry. An overview of mortality risk prediction in sepsis. *Critical care medicine*, 23(2):376–393, 1995.

[13] Ingo A Beinlich, Henri J Suermondt, R Martin Chavez, and Gregory F Cooper. *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks*. Springer, 1989.

[14] Jeff A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, University of California, Berkeley, 1998.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information*

Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[16] R. Blanco, M. Inza, M. Merino, J. Quiroga, and P. Larranaga. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *J Biomed Inform*, 38(5):376–388, 2005.

[17] KH Bonaa and DS Thelle. Association between blood pressure and serum lipids in a population: the tromsø study. *Circulation*, 83(4):1305–1314, 1991.

[18] British Cardiac Society, British Hypertension Society, Diabetes UK, and HEART UK. Jbs 2: Joint british societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart*, 91:v1–v52, 2005.

[19] J. Buckley and I. James. Linear-regression with censored data. *Biometrika*, 66 (3):429–436, 1979.

[20] Elizabeth S Burnside, Daniel L Rubin, Jason P Fine, Ross D Shachter, Gale A Sisney, and Winifred K Leung. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience 1. *Radiology*, 240(3):666–673, 2006.

[21] John Chalmers and Mark E Cooper. Ukpds and the legacy effect. *New Engl J Med*, 359(15):1618–1620, 2008.

[22] Theodore Charitos, Linda C Van Der Gaag, Stefan Visscher, Karin AM Schurink, and Peter JF Lucas. A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Systems with Applications*, 36(2): 1249–1258, 2009.

[23] David Maxwell Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2003.

[24] P M Clarke, A M Gray, A Briggs, A J Farmer, P Fenn, R J Stevens, D R Matthews, I M Stratton, and R R Holman. A model to estimate the lifetime health outcomes

of patients with type 2 diabetes: The United Kingdom Prospective Diabetes Study (UKPDS) outcomes model. *Diabetologia*, 47(10):1747–1759, 2004.

[25] Gary S. Collins and Douglas G. Altman. An independent external validation and evaluation of QRISK cardiovascular risk prediction: A prospective open cohort study. *Brit Med J*, 339:b2584, JUL 7 2009.

[26] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):1, 2011.

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

[28] 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. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Euro Heart J*, 24(11):987–1003, 2003.

[29] N R Cook and P M Ridker. The use and magnitude of reclassification measures for individual predictors of global cardiovascular risk. *Ann Interm Med*, 150(11): 795–802, 2009.

[30] Marie Therese Cooney, Alexandra L. Dudina, and Ian M. Graham. Value and limitations of existing scores for the assessment of cardiovascular risk a review for clinicians. *J Am Coll Cardiol*, 54(14):1209–1227, 2009.

[31] Marie Therese Cooney, Alexandra Dudina, Ralph D'Agostino, and Ian M. Graham. Cardiovascular risk-estimation systems in primary prevention: Do they differ? Do they make a difference? Can we see the future? *Circulation*, 122(3): 300–310, 2010.

[32] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[33] D. R. Cox. Regression models and life-tables. *J Roy Stat Soc A Met*, 34(2): 187–220, 1972.

[34] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*, 118(4):E86–E86, 2008.

[35] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met*, 39(1):1–38, 1977.

[36] Jay R Desai, Pingsheng Wu, Greg A Nichols, Tracy A Lieu, and Patrick J OConnor. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Medical care*, 50:S30, 2012.

[37] Ralph B DAgostino Sr, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care. *Circulation*, 117(6):743–753, 2008.

[38] Chris Fraley, Adrian E Raftery, T Brendan Murphy, and Luca Scrucca. MCLUST version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, University of Washington, Department of Statistics, 2012.

[39] Nir Friedman, Iftach Nachman, and Dana Peér. Learning bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc., 1999.

[40] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian

networks to analyze expression data. *Journal of computational biology*, 7(3-4): 601–620, 2000.

[41] Lise Getoor, Jeanne T Rhee, Daphne Koller, and Peter Small. Understanding tuberculosis epidemiology using structured statistical models. *Artificial Intelligence in Medicine*, 30(3):233–256, 2004.

[42] Alan S. Go, Dariush Mozaffarian, Veronique L. Roger, Emelia J. Benjamin, Jarett D. Berry, Michael J. Blaha, Shifan Dai, Earl S. Ford, Caroline S. Fox, Sheila Franco, Heather J. Fullerton, Cathleen Gillespie, Susan M. Hailpern, John A. Heit, Virginia J. Howard, Mark D. Huffman, Suzanne E. Judd, Brett M. Kissela, Steven J. Kittner, Daniel T. Lackland, Judith H. Lichtman, Lynda D. Lisabeth, Rachel H. Mackey, David J. Magid, Gregory M. Marcus, Ariane Marelli, David B. Matchar, Darren K. McGuire, Mohler Emile R. III, Claudia S. Moy, Michael E. Mussolino, Robert W. Neumar, Graham Nichol, Dilip K. Pandey, Nina P. Paynter, Matthew J. Reeves, Paul D. Sorlie, Joel Stein, Amytis Towfighi, Tanya N. Turan, Salim S. Virani, Nathan D. Wong, Daniel Woo, and Melanie B. Turner. Heart disease and stroke statistics 2014 update: A report from the American Heart Association. *Circulation*, 129(3):E28–E292, JAN 21 2014.

[43] David C Goff, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B DAgostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J ODonnell, 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. *Journal of the American College of Cardiology*, 63(25_PA), 2014.

[44] Major Greenwood et al. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects. Ministry of Health*, (33), 1926.

[45] F E Harrell. *Regression Modeling Strategies*. Springer-Verlag, New York, 2001.

[46] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, and P. Brindle. Performance of the QRISK cardiovascular risk prediction algorithm in an independent

UK sample of patients from general practice: A validation study. *Heart*, 94(1): 34–39, 2008.

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

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

[49] D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression-model. *Commun Stat A-Theor*, 9(10):1043–1069, 1980.

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

[51] Christopher H. Jackson, Simon G. Thompson, and Linda D. Sharples. Accounting for uncertainty in health economic decision models by using model averaging. *J Roy Stat Soc A Sta*, 172:383–404, 2009.

[52] Amy C Justice, Kenneth E Covinsky, and Jesse A Berlin. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130(6):515–524, 1999.

[53] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, Hoboken, NJ, 2002.

[54] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

[55] HJ Kappen, WAJJ Wiegerinck, E Akay, MJ Nijman, JP Neijt, AP van Beek, and E de Koning. promedas: a probabilistic decision support system for medical diagnosis. 2002.

[56] M W Kattan. Comparison of Cox regression with other methods for determining prediction models and nomograms. *J Urology*, 170(6):S6–S9, 2003.

[57] M W Kattan, K R Hess, and J R Beck. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. *Comput Biomed Res*, 31(5): 363–373, 1998.

[58] Joanna Kazmierska and Julian Malicki. Application of the naive Bayesian classifier to optimize treatment decisions. *Radiother Oncol*, 86(2):211–216, FEB 2008.

[59] Timo JT Koski and John M Noble. A review of bayesian networks and structure learning. *Annales Societatis Mathematicae Polonae. Series 3: Mathematica Applicanda*, 40(1):53–103, 2012.

[60] Stefania Lamon-Fava, Peter WF Wilson, and Ernst J Schaefer. Impact of body mass index on coronary heart disease risk factors in men and women the framingham offspring study. *Arteriosclerosis, thrombosis, and vascular biology*, 16(12): 1509–1515, 1996.

[61] Martijn Lappenschaar, Arjen Hommersom, Peter J. F. Lucas, Joep Lagro, and Stefan Visscher. Multilevel Bayesian networks for the analysis of hierarchical health care data. *Artif Intell Med*, 57(3):171–183, MAR 2013.

[62] P. Larranaga, B. Sierra, M J Gallego, M J Michelena, and J M Picaza. Learning Bayesian networks by genetic algorithms: A case study in the prediction of survival in malignant skin melanoma. *Artif Intell Med*, 1211:261–272, 1997.

[63] Malcolm R Law, Nicholas J Wald, and AR Rudnicka. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *Bmj*, 326(7404):1423, 2003.

[64] S. Lemeshow and D. W. Hosmer. A review of goodness of fit statistics for use in the development of logistic-regression models. *Am J Epidemiol*, 115(1):92–106, 1982.

[65] Shengxu Li, Wei Chen, Dianjianyi Sun, Camilo Fernandez, Jian Li, Tanika Kelly, Jiang He, Marie Krousel-Wood, and Paul K Whelton. Variability and rapid increase in body mass index during childhood are associated with adult obesity. *International Journal of Epidemiology*, 44(6):1943–1950, 2015.

[66] A. M. Lipsky and R. J. Lewis. Placing the Bayesian network approach to patient diagnosis in perspective. *Ann Emerg Med*, 45(3):291–294, MAR 2005.

[67] Donald M. Lloyd-Jones. Cardiovascular risk prediction basic concepts, current status, and future directions. *Circulation*, 121(15):1768–1777, 2010.

[68] P. J. F. Lucas, H. Boot, and B. G. Taal. Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Method Inform Med*, 37 (3):206–219, SEP 1998.

[69] P. J. F. Lucas, N. C. deBruijn, K. Schurink, and A. Hoepelman. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artif Intell Med*, 19(3):251–279, JUL 2000.

[70] P. J. F. Lucas, L.C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artif Intell Med*, 30(3):201–214, MAR 2004.

[71] 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. Systematic review of cardiovascular disease risk assessment tools. Technical report, Agency for Healthcare Research and Quality (US), 2011.

[72] Andrew Moore and Weng-Keen Wong. Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In *ICML*, volume 3, pages 552–559, 2003.

[73] Kevin P Murphy. Inference and learning in hybrid Bayesian networks. Technical report, University of California, Berkeley, Computer Science Division, 1998.

[74] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[75] Patrick Murray, Gary W Chune, and Vasudevan A Raghavan. Legacy effects from DCCT and UKPDS: what they mean and implications for future diabetes trials. *Current atherosclerosis reports*, 12(6):432–439, 2010.

[76] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[77] Agnieszka Onisko, Marek J Druzdzel, and Hanna Wasyluk. A probabilistic causal model for diagnosis of liver disorders. In *Proceedings of the Workshop held in Malbork, Poland, Malbork, Poland*, 1998.

[78] Niels B Peek. Explicit temporal models for decision–theoretic planning of clinical management. *Artificial Intelligence in Medicine*, 15(2):135–154, 1999.

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

[80] Michael J. Pencina, Ralph B. D'Agostino Sr., and Ewout W. Steyerberg. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*, 30(1):11–21, 2011.

[81] Margaret S. Pepe. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol*, 173(11):1327–35, 2011.

[82] Joep Perk, Guy De Backer, Helmut Gohlke, Ian Graham, Željko Reiner, Monique Verschuren, Christian Albus, Pascale Benlian, Gudrun Boysen, Renata Cifkova, et al. European guidelines on cardiovascular disease prevention in clinical practice (version 2012). *European Heart Journal*, 33(13):1635–1701, 2012.

[83] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds risk score. *J Am Med Assoc*, 297(13):1433–1433, 2007.

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

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

[86] Michael RH Rockwood and Susan E Howlett. Blood pressure in relation to age and frailty. *Can Geriatr J*, 14(1):2–7, 2011.

[87] A. G. Rotnitzky and J. M. Robines. Inverse probability weighted estimation in survival analysis. In Peter Armitage and Theodore Colton, editors, *The Encyclopedia of Biostatistics*. John Wiley & Sons, second edition, 2004.

[88] Patrick Royston and Douglas G Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1):33, 2013.

[89] Shantanu Sarkar and Jodi Koehler. A dynamic risk score to identify increased risk for heart failure decompensation. *IEEE T Biomed-Eng*, 60(1):147–150, JAN 2013.

[90] Paola Sebastiani, Kenneth D Mandl, Peter Szolovits, Isaac S Kohane, and Marco F Ramoni. A bayesian dynamic model for influenza surveillance. *Statistics in Medicine*, 25(11):1803–1816, 2006.

[91] M. Berkan Sesen, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. Bayesian networks for clinical decision support in lung cancer care. *PLOS One*, 8(12):e82349, DEC 6 2013.

[92] Sudha Seshadri, Alexa Beiser, Margaret Kelly-Hayes, Carlos S Kase, Rhoda Au, William B Kannel, and Philip A Wolf. The lifetime risk of stroke estimates from the Framingham Study. *Stroke*, 37(2):345–350, 2006.

[93] B. Sierra and P. Larranaga. Predicting survival in malignant skill melanoma using Bayesian networks automatically induced by genetic algorithms: An empirical comparison between different approaches. *Artif Intell Med*, 14(1-2):215–230, 1998.

[94] Wade P. Smith, Jason Doctor, Juergen Meyer, Ira J. Kalet, and Mark H. Phillips. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model. *Artif Intell Med*, 46(2):119–130, JUN 2009.

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

[96] Ivan Stajduhar and Bojana Dalbelo-Basic. Learning Bayesian networks from survival data using weighting censored instances. *J Biomed Inform*, 43(4):613–622, AUG 2010.

[97] Ivan Stajduhar and Bojana Dalbelo-Basic. Uncensoring censored data for machine learning: A likelihood-based approach. *Expert Syst Appl*, 39(8):7226–7234, JUN 15 2012.

[98] Ivan Stajduhar, Bojana Dalbelo-Basic, and Nikola Bogunovic. Impact of censoring on learning Bayesian networks in survival modelling. *Artif Intell Med*, 47(3):199–217, 2009.

[99] Jeremiah Stamler, Rose Stamler, and James D Neaton. Blood pressure, systolic and diastolic, and cardiovascular risks: Us population data. *Archives of internal medicine*, 153(5):598, 1993.

[100] Russell Stuart and Norvig Peter. *Artificial Intelligence: A Modern Approach*, volume 2. Prentice Hall, Upper Saddle River, New Jersey, 2003.

[101] T M Therneau, P M Grambsch, and T R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[102] Jin Tian, Ru He, and Lavanya Ram. Bayesian model averaging using the k-best Bayesian network structures. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 589–597, Corvallis, Oregon, 2010. AUAI Press.

[103] Susanne MAJ Tielemans, Johanna M Geleijnse, Alessandro Menotti, Hendriek C Boshuizen, Sabita S Soedamah-Muthu, David R Jacobs, Henry Blackburn, and Daan Kromhout. Ten-year blood pressure trajectories, cardiovascular mortality, and life years lost in 2 extinction cohorts: the Minnesota Business and Professional Men Study and the Zutphen Study. *Journal of the American Heart Association*, 4(3):e001378, 2015.

[104] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65 (1):31–78, 2006.

[105] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

[106] Linda C van der Gaag, Silja Renooij, CLM Witteman, Berthe MP Aleman, and Babs G Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in medicine*, 25(2):123–148, 2002.

[107] Marina Velikova, Josien Terwisscha van Scheltinga, Peter J. F. Lucas, and Marc Spaanderman. Exploiting causal functional relationships in bayesian network modelling for personalised healthcare. *Int J Approx Reason*, 55(1):59–73, JAN 2014.

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

[109] Yvonne Vergouwe, Karel GM Moons, and Ewout W Steyerberg. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology*, 172(8):971–980, 2010.

[110] Joan Vila-Frances, Juan Sanchis, Emilio Soria-Olivas, Antonio Jose Serrano, Marcelino Martinez-Sober, Clara Bonanad, and Silvia Ventura. Expert system for predicting unstable angina based on Bayesian networks. *Expert Syst Appl*, 40 (12):5004–5010, SEP 15 2013.

[111] David M Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E Johnson, Gabriela Vazquez-Benitez, and Patrick J OConnor. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 2016.

[112] Massimo Volpe, Francesco Cosentino, Giuliano Tocci, Francesca Palano, and Francesco Paneni. Antihypertensive therapy in diabetes: the legacy effect and RAAS blockade. *Current hypertension reports*, 13(4):318–324, 2011.

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

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

[115] B. Zupan, J. Demsar, M W Kattan, J R Beck, and I. Bratko. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artif Intell Med*, 1620:346–355, 1999.