CSBN: A Hybrid Approach For Survival Time Prediction With Missing Data

Simon Rabinowicz\(^1\), Raphaela Butz\(^2,3\), Arjen Hommersom\(^3,4\), and Matt Williams\(^5,6\)

\(^1\) Faculty Of Medicine, Imperial College London, UK
\(^2\) Institute for Computer Science, TH Köln, Germany
\(^3\) Department of Computer Science, Open University of the Netherlands
\(^4\) Department of Software Science, Radboud University, The Netherlands
\(^5\) Department of Radiotherapy, Charing Cross Hospital, London, UK
\(^6\) Computational Oncology Laboratory, Imperial College London, UK

Abstract. Survival prediction models most commonly use Cox Proportional Hazards (CPH) models, and are frequently used in medical statistics and clinical practice. However, such models underperform when the predictor variables are missing. By building Bayesian networks we automatically construct a model with the most important risk factors and relationships between risk factors and Bayesian networks are able to infer the likely values of missing data. We therefore propose a hybrid solution, consisting of a CPH model and a BN, where the predictive variables in the CPH model are the child nodes of a BN, which we call CSBN. We learn the CPH and BN models separately, using standard techniques, with the only constraint being that the variables that are predictors in the CPH model are child nodes in the BN. This allows us to fuse the two models, using the predictors of the CPH models as the join points. We test our approach by examining the performance of the CPH model, against the hybrid CSBN model, using both complete data cases and in cases with missing data. We calculate the performance of the survival prediction for both CPH and CSBN using the C-index and a normalised error function as metrics. For the CPH model, predictive error was significantly larger for missing data (±3120.8 days) compared to complete data (±1171.5 days; \(p = 3.6e-07\)). This was also true for the CSBN ±1387.3 days for missing data compared with ±1171.5 days with complete data (\(p = 0.01568\)). However, with missing data, the predictive error was significantly larger for the CPH model (±3120.8 days) than the CSBN (±1171.5 days; \(p = 0.03274\)). In conclusion the CSBN methodology provides a more effective method of predicting survival when using incomplete data.

Keywords: Bayesian Network · Survival Analysis · Prognosis.

1 Introduction

Survival analysis (time to event analysis) and modelling provides valuable information for clinicians. Determining survival accurately is important for patient
wellbeing and allows patients to make better decisions about what treatment is appropriate for them and identifies important prognostic indicators [6]. This can help identify new targets for pharmacological, lifestyle or surgical intervention, and may improve clinical accuracy, as clinicians tend to substantially overestimate survival [1]. Survival modelling uses multivariate data to identify predictors influencing survival by looking at how they affect the time to an event, i.e. death [3]. This works by looking at the time from an intervention such as initiation of a therapy and assessing both the time to event and whether the event occurs [12]. Time to event analysis is a powerful technique as it provides details on both event occurrence and time of event [12]. However, real-world clinical data is often missing data on key predictive variables. Here, we present a hybrid method that integrates a Bayesian Network and Cox Proportional Hazards (CPH) survival model that we call CSBN and show that, when presented with incomplete data, it achieves better predictive accuracy that the CPH model alone.

This paper is structured as follows. In Section 2 we discuss the necessary methods with respect to survival analysis, BN and predictive error. In Section 3, we discuss related work. The experimental setup for our work is presented in Section 4 and discussed in Section 5. Finally, in Section 6, we consider future research and conclude.

2 Methods

2.1 Survival Analysis

In traditional survival analysis, the main aim is to model the survival function $S$, i.e., $S(t) = P(T > t)$, with $T$ being a random variable representing the time of death defined by a density function $p$. Equivalently, we can consider a hazard function, which is defined as:

$$h(t) = \frac{p(t)}{S(t)} = -\frac{d}{dt} \log S(t)$$

which can be thought of as the density of death at $t$ given that the patient survived until $t$. In practice, one is usually interested in how the hazard varies in response to some explanatory variables $X = X_1, \ldots, X_n$, sometimes also called predictor variables or risk variables. A popular parameterised model is the CPH model, which estimates the relative risk or hazard ratio $hr$ as a log-linear model:

$$hr(X) = \frac{h(t \mid X)}{h_0(t)} = \exp \left( \sum_{j=1}^{n} w_jX_j \right)$$

where $h_0$ is a baseline hazard function and $w = \{w_1, \ldots, w_n\}$ a set of weights. Note that this ratio is assumed to be constant for all $t$, and parameters $w$ are assumed to be independent of censoring time. Various baseline hazard can be used,
e.g. $h_0$ can be estimated non-parametrically using a Kaplan-Meier estimate [7]. Once derived, the conditional survival function can be obtained:

$$S(t \mid X) = \exp \left[ - \int_0^t h_r(X)h_0(s)ds \right]$$

(3)

### 2.2 Conditional Survival Bayesian Networks

**Definition** In order to combine survival analysis with Bayesian networks, we defined in Rabinowicz et al. [11] a **conditional survival Bayesian network** (CSBN). These models are defined as a Bayesian network $B = (G, P)$, such that $G$ is an acyclic directed graph with with vertices $V$, which represent random variables. The set of vertices $V$ consists of a set of discrete nodes $D$ and a (continuous) survival nodes $T$, such that $T \cap D = \emptyset$ and $T \cup D = V$.

Let $\pi(V_i)$ indicate the parents of node $V_i$ in graph $G$. It is assumed in CSBNs that for each $V_i \in V$, $\pi(V_i) \subseteq D$, i.e., survival nodes cannot be used as parents of other nodes. Furthermore, the joint density of a CSBN factorises as follows:

$$p(V) = \prod_{D_i \in D} P(D_i \mid \pi(D_i)) \prod_{T_i \in T} p(T_i \mid \pi(T_i))$$

(4)

such that $p(T_i \mid \pi(T_i))$ is a density function for survival, which in this paper is assumed to be derived from a CPH model, i.e.,

$$p(t_i \mid \pi(T_i)) = h(t \mid \pi(T_i)) \exp \left[ - \int_0^t h(s \mid \pi(T_i))ds \right].$$

(5)

**Learning** A popular approach for structure learning is **score-based learning**, where a model selection criterion such the AIC or MDL is maximised. Assuming we have complete data, such scores decompose to individual random variables, as the scores are a linear combination of the log-likelihood and a penalty term:

$$\text{score}(B) = \sum_{D_i} \text{score}(D_i \mid \pi(D_i)) \sum_{T_i} \text{score}(T_i \mid \pi(T_i))$$

(6)

Since the continuous random variables $T_i$ are restricted to being children of the discrete variables in our approach, the structure of the discrete variables can be estimated separately and the optimal parent set for each $T_i$ can be searched, e.g. by a exhaustive search on the set of potential parents.

However, real data is commonly incomplete. In that case the score does not decompose, which means that structure learning is more difficult. Effective methods include structural **expectation maximisation** (EM) algorithms [4]. In these algorithms, the expected score (e.g. MDL or AIC) is iteratively improved by computing expected sufficient statistics (in the E-step) and then learning the structure and parameters that maximises the score given these expectations (in the M-step), until convergence has been reached. In addition to our approach defined in [11] we are using structural EM to learn the structure of the discrete Bayesian network.
**Inference** Similar to the learning problem, inference can be performed in a two-step process. Suppose for simplicity that $|T| = 1$. Then:

$$p(t \mid D_i) = \sum_{\pi(T)} p(t \mid \pi(T)) P(\pi(T) \mid D_i)$$

where $P(\pi(T) \mid D_i)$ can be computed on a Bayesian network with discrete nodes only and $p(t \mid \pi(T))$ is the event rate. Arbitrary other quantities can be computed using Bayes’ rule, e.g.

$$P(D_i \mid T < t) \propto P(T < t \mid D_i) P(D_i)$$

where $P(T < t \mid D_i)$ can be computed from the density as given in Equation 7.

### 2.3 Predictive Error

**Concordance Index** The concordance index (c-index) is a measure of predictive discrimination and estimates the probability of concordance between predicted and observed outcomes [5]. However, it is defined as the order of a pair of outcomes, and so is insensitive to the magnitude of error. We therefore developed a **Normalised error function** to assess the prediction error of our models:

$$e = \frac{(Sa - Sp)^2}{Sa}$$

This error function has three advantages: firstly, it is symmetrical (i.e. under-prediction is as important as over-prediction), secondly it is normalised to actual survival, and finally it reports errors in days.

### 3 Related work

Bayesian networks, especially dynamic Bayesian networks, have been proposed for modelling prognosis, e.g. prognostic Bayesian networks [13]. These networks are very strong at modelling the care process over time, and eventually the clinical outcome. However, these models are not suitable for modelling simple time-to-event data, where most of the data is of a non-temporal nature.

Modified Bayesian networks for learning from censored time-to-event data have been introduced more recently [2, 14]. These use weighting schemes to handle censored data, but do not appear to outperform CPH models [14] for predicting survival. In this paper, we show that if the problem domain can be modelled using a CSBN, then standard and well-understood Bayesian network methods can be applied, and no specialized learning schemes are required.

A translation of CPH model into a discrete Bayesian networks was proposed by Kraisangka and Druzdzel [10]. Their approach converts every explanatory and survival variable into discrete random variables in a Bayesian network and introduces a time index variable. The time indexing node and every explanatory node is then modelled as parent of the survival node. The conditional probability
table for the survival node is populated subsequently with the probabilities of survival conditional on every combination of values of the explanatory variables. The resulting Bayesian network produces survival probabilities for each index of the time variable that are identical to those of the CPH model. The main limitation of this approach is that it requires time to be discretised, which makes reasoning about statistics such as median survival times inaccurate. Another issue with discrete time is that the number of parameters is exponential in the number of discrete time points that is chosen, which makes inference intractable if one would consider a fine-grained discrete-time solution. Therefore, we would argue that it is better to keep the continuous-time variable as part of the Bayesian network.

4 Experimental Work

4.1 Experimental Setup

We used a publicly-available dataset from a cancer genetics study [9]. We removed observations that were missing data on both follow-up or death and split the data (3:1) into training and test datasets. We constructed a CPH model following the approach outlined by Harrell [6] using the rms package in R. We checked for evidence of violation of CPH assumptions, and retained variables based on Aikake’s Information Criteria. To learn the Bayesian network model, we used the decomposition outlined in Section 2.2. The discrete-network part was learned using the R gRain package using expectation-maximization. The survival node in the network was the same as the CPH model, and was attached to the BN. The resulting network is shown in Figure 1.

We validated our approach by comparing survival prediction from the CPH and CBSN using patients with complete data, using the C-index and normalised error function as metrics. To assess the performance in patients with missing data, we removed data on one of the predictor variables (tumour stage) from patients in the test dataset, and attempted to predict survival. We compared the estimated survival from each approach with actual survival. Since the CPH package requires complete data, we randomly resampled the missing variable, based on the distribution in the original dataset, and took the median survival prediction from 1000 iterations. The paired Wilcoxon signed rank test was used to assess significance between the normalised error in the prediction from the CPH and CBSN for each case. We used the Holm-Bonferroni correction to adjust for multiple comparisons.

4.2 Results

The cleaned dataset consisted of 1444 patients. The median survival was 1971 days and the median age 63 years. 329 patients were alive at the end of the follow-up. There were 1083 patients in the training dataset and 361 in the test dataset. Our CPH model met the proportional hazards assumptions (global score 3.52 $p = 0.642$) and appeared well calibrated with a mean error of 0.03282158
Fig. 1. Bayesian network structure for cancer genetics. Predictor variables shown attached to the survival node by red arrows indicates that inference from the CPH model. The black arrows indicate inference from the Bayesian network.

The median error for survival time predictions with complete data was 1172 days. The median error for CPH predictions based on incomplete data was 3120.8 days and was 1373 days with the CSBN which were significantly greater than with complete data ($p = 0.01568$ and $p = 3.6e-07$ respectively). The C-index for the predictions using complete data was 0.6128 and with missing data was 0.5254 (CPH) and 0.5694 (CSBN).

5 Discussion

The CPH model was shown to fit the proportional hazards assumption and was well calibrated and validated. It is therefore a robust comparator to our novel approach. When asked for survival predictions in patients with missing data, the CSBN gave predictions with substantially less error than compared to CPH with random sampling. As we expected, because the CPH model is part of the CSBN, and because of the effect of d-separation, the error was the same for the CPH model and CSBN when predicting survival on complete data. Our findings are in line with related but distinct work [8]. Our previous work [11] presented the theoretical outline for CSBN, but used a simpler CPH model, a much smaller dataset and a much simpler metric for evaluation. This work therefore represents an important confirmation that the CSBN approach outperforms CPH when asked for survival predictions in patients with missing data.

There is no ideal performance metric for survival models, and we have used two different measures. Other methods exist, such as the Brier score, and our approach is limited by not considering the distribution of censoring in the data. Although we also used the C-index to assess performance, the differences between...
the C-index for the two models would normally lead one to consider them as being broadly equivalent, and modest changes in C-index have limited clinical significance, whereas the difference in error in the predictions (3120 vs. 1373 days) is more understandable.

The novel aspect of this approach is that the CSBN uses inference to better estimate survival in patients with missing data. The coupling of a validated CPH model to the CSBN as a way of identifying predictors allows the Bayesian Network to impute missing data amongst the predictor variables, and allows us to make use of the extensive body of work on both BN and CPH. A further advantage of our approach is that the CSBN could be used to explore the impact of variables that have been eliminated from the CPH model, by removing one or more of the direct predictors and then seeing the effect that the values of the indirect predictors have on survival. We also avoid some of the weaknesses of previous work, such as [10] which requires time to be discretised which makes reasoning about statistics such as median survival times inaccurate and results in an exponential expansion in data.

6 Further work

The CSBN approach is fundamentally a hybrid model, consisting of two parts. Although this work uses a BN and a CPH, in theory one could apply any inferential technique to infer predictive variables and use any existing survival model. Further work will develop a more generalisable approach for integrating the inferential and survival parts of the network.

References

4. Friedman, N.: Learning belief networks in the presence of missing values and hidden variables 97 (06 2000)


