Using Discrete Hazard Bayesian Networks to Identify which Features are Relevant at each Time in a Survival Prediction Model

Li-Hao Kuan LIHAO@UALBERTA.CA
Dept of Computing Science, University of Alberta, Edmonton Alberta T6G 2E8

Russell Greiner RGREINER@UALBERTA.CA
Dept of Computing Science, University of Alberta, Edmonton Alberta T6G 2E8
Alberta Machine Intelligence Institute, Alberta

Abstract

When predicting the survival time of a patient, different covariates may be important at different times. We introduce a survival prediction model, "discrete hazard Bayesian network", that can provide individual survival curves and also identify which features are relevant for each time interval. This model encodes the discrete hazard function as a sequence of (possibly different) Bayesian networks, one for each time interval. Note each such network includes a "Death" node, which is True iff the person dies in that interval. A set of features relevant for each time interval are the nodes in the Markov blanket around that "Death" node for that interval. We also apply a "discrete hazard computation correction" based on the effective sample size – a correction that avoids biased survival curves.

We first show that our model is effective by demonstrating that it can identify the time-varying relevance of the features, using the synthetic dataset. We then provide two real-world examples by analyzing the relevant features for different times on the North Alberta cancer dataset and the Norway/Stanford breast cancer dataset.

Keywords: survival prediction, feature selection, (discrete) hazard function, Bayesian networks, time-varying effects

1. Introduction

The field of survival prediction provides algorithms that attempt to estimate the time until an event will happen – *e.g.*, the time until the death of a patient. As such, an accurate survival prediction models can help doctors to make treatment decisions. To help people interpret these predictions, it is useful to know which variables are relevant. However, different variables may be relevant, at different times. For example, in the months immediately after an operation, a patient's blood factor are very important. However, if the patient survives more than a year, then BMI and age are typically more important. Indeed, assuming a feature's impact is constant might lead to misleading conclusions (Bellera et al., 2010).

We provide a model that uses Bayesian Networks and Individual Survival Distributions. To define these terms: A *Bayesian network* is a probabilistic graphical model that encodes a probability distribution as a graph structure with nodes corresponding to random variables and arcs that encode conditional dependencies (Koller and Friedman, 2009; Pearl, 1988). The graphical structure allows people to easily "read off" the (conditional) dependencies of

the set of variables – to identify how these covariates affect each other (Heckerman, 2008); in particular, the "Markov Blanket" around a node X – which includes X's parents, X's children, and the other parents of X's children – are a set of features directly relevant to X.

In general, an *ISD (Individual Survival Distribution)* model represents the probability that each *specific patient* will survive until time t, over *all future time points* t (Haider et al., 2020). The fact that it is individualized, and over all future times, mean an ISD can effectively deal with the heterogeneity of patients and the high variation of survival times.

Here, we define and use a novel survival model, Discrete Hazard Bayesian Network (DHBN), which builds one Bayesian network (over the covariates as well as a "Death" random variable Λ_j) for each time interval I_j, to represents the discrete hazard. Here, the "discrete hazard" is the conditional probability that a subject will die in a time interval, given that this subject is alive at the start of this interval. We provide a training process that is fully automatic and driven by the training data. DHBN incorporates right-censored data by applying the probability computation correction inspired by the life-table analysis. More importantly, the Bayesian networks built by DHBN are interpretable, in that they can be used to find the time-varying effects of each covariate. This DHBN model gives an individualized survival curve for each subject, which we demonstrate is calibrated. We also demonstrate that this model can identify important features at different times.

Section 2 presents some previous works that are related to feature selection and time-varying feature effects, and discusses different categories of survival prediction models. Section 3 defines the discrete hazard function, then describes how our DHBN learner can learn a sequence of Bayesian networks to represent a discrete hazard function for each interval, which can be used to produce a survival curve for a patient. Section 4 introduces a computation correction to effective sample size when learning the parameters. Section 5 provides the empirical results on synthetic and real-world data. Finally, Section 6 summarizes the results and discuss some future directions.

2. Background

For standard classification and regression tasks, standard feature selection methods, such as L1-regularizer methods, mRMR, etc., can identify the features that are important to a target variable (Peng et al., 2005). In the survival prediction setting, one can use the (L1-regularized) Cox Proportional Hazards (CoxPH) model or accelerated failure time model (AFT) to select features that affect the hazard rate or the acceleration of failure time, by removing the features whose associated regression coefficient is essentially 0 (Tibshirani, 1997; Huang and Ma, 2010). However, these approaches can only include, or exclude, a variable, *for all time points*; none can include a variable for only a subset of the time.

These approaches implicitly embody the Proportional Hazard assumption – *i.e.*, that a variable has the same importance at all future times. However, this assumption is often violated (Xue et al., 2013), which means that models like CoxPH would need to exclude those variables. We provide a method that can use variables with time-varying important, appropriately, by providing a way to include different variables at different times.

There are many ways for survival prediction systems to deliver the prediction results. Each of these focuses on different characteristics (Haider et al., 2020). Some systems focus

on the discrimination ability by only providing a risk score – allowing this tool to predict the relevant order in which people will die. The standard such model is CoxPH. Other systems focus on the probability that a subject will survive until a single predefined time point – *e.g.*, 5-year survival (Sailer et al., 2015) or 30-day hospital re-admissions (Liu et al., 2020). There are also survival models that provide a general survival distribution, applicable to entire group of people (*e.g.*, everyone with stage 4 lung cancer) – *e.g.*, the Kaplan-Meier estimator.

Instead, we want a model that provides a more general form of survival, called "Individual Survival Prediction" (Haider et al., 2020), that predicts a unique survival distribution for each individual subject $\boldsymbol{x}_i$ – providing $p(\, t \,|\, \boldsymbol{x}_i \,)$, which is the probably that $\boldsymbol{x}_i$ lives until (at least) time $t > 0$. Note these prediction results can be used for both risk ranking (*e.g.*, by using the mean of the curve, or the median) and estimating survival until a single time point. Some examples of individual survival prediction models includes MTLR (Yu et al., 2011), accelerated failure time (Kalbfleisch and Prentice, 2011) and our model DHBN.

3. DHBN Model

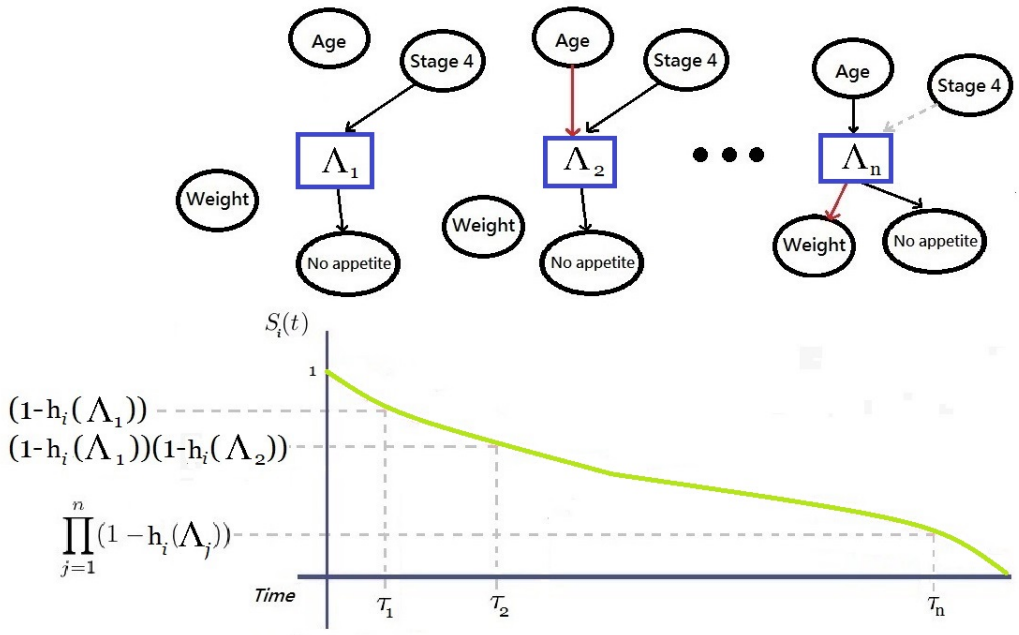

Figure 1: Model architecture for DHBN. An edge is red if it was not in the previous graph; and is dashed if it is deleted from the previous graph.

Consider the dataset $DS = \{(\boldsymbol{x_i}, t_i, c_i)\}_{i=1}^{K}$, where $\boldsymbol{x_i}$ is a d-dimension vector of covariates for subject i, t_i is a future time, which is the time to death if this instance is uncensored (c_i=1), or the time of censoring if the death event has not happened yet (c_i=0). K is the total number of subjects in the dataset.

First, we set up N time points by the quantile of the event happened or censored time $\{t_i\}$, producing the time points $\{\tau_j\}_{j=1}^{N}$ that we use to divide the time axis into N discrete intervals, from $I_1 = [\tau_0 = 0, \tau_1)$ to $I_N = [\tau_{N-1}, \tau_N)$. Here, we use $N = 10$. For each of the intervals I_j, we collect a dataset D_j, which includes the relevant subset of the instances who are alive at the start τ_{j-1}, each described by values $\boldsymbol{x}_i$ of the covariate variables X, and the appropriate value λ_{ij} for its "Death" variable Λ_j. Because we want to model the discrete hazard, which is the conditional probability of the subject's death given that the subject is alive at the start of this interval, D_j only includes the subjects known to be alive in the previous time interval I_{j-1} – *i.e.*, it does not include any subjects who died or were censored before interval I_j. The death value λ_{ij} is 0 if the subject i is still alive in interval I_j, is 1 means the subject dies in interval I_j (note that D_j excludes subject that died before τ_{j-1}), and is NA if the subject becomes censored in interval I_j.

$$
\begin{aligned}
D_j &= \{ (\boldsymbol{x}_i, \lambda_{ij}) \mid t_i > \tau_{j-1} \} \\
\lambda_{ij} &= \begin{cases} 0, & t_i \geq \tau_j \\ 1, & \tau_{j-1} \leq t_i < \tau_j \wedge c_i = 1 \\ NA, & \tau_{j-1} \leq t_i < \tau_j \wedge c_i = 0 \end{cases}
\end{aligned}
$$

The probability of $\Lambda_j = 1$ given $\Lambda_{j-1} = 0$ and $X = \boldsymbol{x}_i$ is the discrete hazard for subject i. Discrete hazard, denoted by

$$
h_i(\Lambda_j) = \begin{cases} p(\Lambda_j = 1 \,|\, X = \boldsymbol{x}_i), & j = 1 \\ p(\Lambda_j = 1 \,|\, \Lambda_{j-1} = 0, X = \boldsymbol{x}_i), & j > 1 \end{cases}
$$

is the conditional probability of the subject's death given that the subject is alive in the previous time interval and all the other covariates $\boldsymbol{x}_i$.

We learn a graphical structure G_j and estimate the parameters using data D_j for each interval I_j. The variables in each graphical structure G_j includes all the covariates X and the death variable Λ_j. To incorporate the censored data, we apply a calculation correction when estimating the parameters for the discrete hazard $h_i(\Lambda_j)$, which will be described in the Section 4.

Finally, we build a smooth survival curve $S_i(t)$ for the i^{th} patient (described by $\boldsymbol{x}_i$) by smoothly interpolating through the points $\{ [\tau_n, S_i(\tau_n)] \}_{n=1}^{N}$ where

$$
S_i(\tau_n) = \prod_{j=1}^{n} (1 - h_i(\Lambda_j))
$$

see bottom of Figure 1.

4. Estimating Discrete Hazard

When estimating the discrete hazard with censored data, we need to apply a correction to the probability calculation because the instances are grouped into discrete time intervals. This correction has been used in the life-table analysis (Igwenagu, 1993).

Censoring can happen at any time within the interval. The censored subject should not be counted as a whole person-interval but should also not be ignored because they live a

fraction of the interval. By assuming the censorship distributed uniformly across the time interval, the mean survival period of all subjects whose censor time falls in the interval I_j should be the half of the whole interval. Thus, the effective sample size (*i.e.*, people at risk) should be:

$$n'_j \quad = \quad n_j - \frac{c_j}{2}$$

where n_j is the number of people alive at the start of time interval I_j, which excludes subjects who are already dead or censored before the current interval; and c_j is the number of instances whose censor time falls in the interval I_j. With this effective sample size n'_j, we can calculate the discrete hazard for the whole population in interval I_j, defined as $h(\Lambda_j)$:

$$\begin{aligned} h(\Lambda_j) &= \begin{cases} p(\,\Lambda_j = 1\,), & j = 1 \\ p(\,\Lambda_j = 1 \,|\, \Lambda_{j-1} = 0\,), & j > 1 \end{cases} \\ &= \frac{d_j}{n'_j} \end{aligned}$$

where d_j is the number of (uncensored) people who die during the interval I_j. This calculation correction is applied to the parameter estimation for the conditional probability table of the Λ node. Similar to the calculation of $h(\Lambda_j)$, for $h_i(\Lambda_j)$ the subjects whose censoring time falls in I_j are count as half alive. The empirical results on a real-world dataset show that using n'_j rather than n_j or $n_j - c_j$ is necessary to avoid over optimistic or pessimistic estimation; see Table 1. n'_j has the least bias and the overestimate percentage closest to 50%. (The "Overestimate (%)" is the proportion of the test set whose estimated time of death is greater than the true time of death, using the "death time" computation used for computing L1 loss. See Section 5.1 for details.)

Table 1: The comparison of using n_j, n'_j, and $n_j - c_j$ for effective sample size tested on NACD colorectal cancer dataset.

	Bias	Overestimate(%)
n_j	2.674	55.894
n'_j	**0.518**	**53.157**
$n_j - c_j$	-5.032	47.368

5. Empirical results

5.1. Evaluation

We will measure each survival model's performance using concordance, D-calibration, and L1 loss. The concordance (aka "c-index") is a measure for discrimination; here, the "risk" for each survival curve is the median of survival time from the survival curve $S_i(t)$ – *i.e.*, where the curve crosses 0.5. (If necessary, we extending the curve beyond our maximum survey time window using the average slope of the whole curve (Haider et al., 2020, Appendix A).) D-calibration, a statistical test for the overall calibration of the survival curve (Haider et al.,

2020), holds if the probability of the times of the actual deaths t_i, over the distribution of patients, is distributed uniformly within the testing data – *i.e.*, if $S_i(t_i) \sim U[0,1]$. The D-calibration test produces a p-value. Stated informally, the closer the p-value to 1, the more calibrated the death time probability distribution. Here, we use D-calibration to refer to the statistical quantity computed, rather than the actual "reject/no-reject" test result.

We use the median of the survival curve to calculate the L1 loss, extending the survival curve if necessary, as described above. For censored data, we estimate a specific survival time, by adding to the censored time, the expected value of the Kaplan-Meier distribution (estimated over the training sample), conditioned on living until at least that censored time; see (Haider et al., 2020). All the results shown are five-fold cross-validation.

Note we also explored a few other models such as Cox-KP, which applies the Kalbfleisch-Prentice extension to the Cox risk scorer (Kalbfleisch and Prentice, 2011), and MTLR (Yu et al., 2011).

5.2. Synthetic data

To show that the DHBN can capture the important factors for different times, we generate synthetic data, based on a set of Bayesian networks, where the hazard for different time intervals depend on different covariates. We use six random variables, { A, B, C, D, E, F }, and consider 3 time intervals. The effects of the variables are constant hazard within an interval but different in other intervals. Before t1 = 13, the hazard depends on variables { A, B, F}; between t1 and t2 = 32, the hazard depends on { D, E, F}, and after t2, the hazard only depends on {F}; see Figure 2. We generated 2400 samples and censored the data, by randomly selecting a subject, then randomly picking a censoring time, uniformly wrt the whole survey time window [0,60]. If that censoring t_c happens before death t_d, the subject is censored. Here, 34% of the subjects were censored.

Table 2 and Figure 3 show the Markov blanket (neighboring variables and the children's parents) around the death variable in the Bayesian network structures learned by DHBN. The graphical structures recover the important factors successfully by connecting them to the death node Λ_j. We see that no nodes connect to the death node in interval I_8; we think this is because I_8 includes the transition of variable effect from {D,E,F} to only {F}. The transition makes the data more noisy in this interval. In other intervals, DHBN does not find some relevant variables because those variables' impact on the discrete hazard are small or the samples are unbalanced.

For comparison, Table 3 shows each feature's contribution in the learned Cox proportional hazard model. where coefficients with larger absolute values are more important. The variables selected by Cox-KP match the DHBN with the most important variables A, D, and F appear the most in the network structure. However, the Cox-KP can not show the time dependence of those variables. We also compare our DHBN with MTLR models (Yu et al., 2011). Table 4 shows that DHBN has the best concordance and L1-loss, and passes D-calibration.

5.3. North Alberta Cancer Dataset

We applied our model to the North Alberta Cancer Dataset for colorectal cancer, which contains 34 covariates for 950 individuals, with 51.8% censoring. Table 5 lists the important

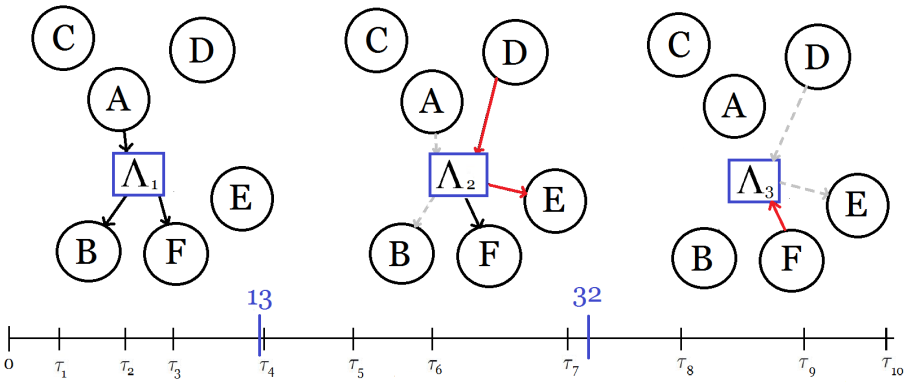

Figure 2: Graphical structures for synthetic dataset. Dashed lines are the non-edges, deleted from previous graph. Red lines are the new edges. The tick-marks on the time axis separate the different time intervals.

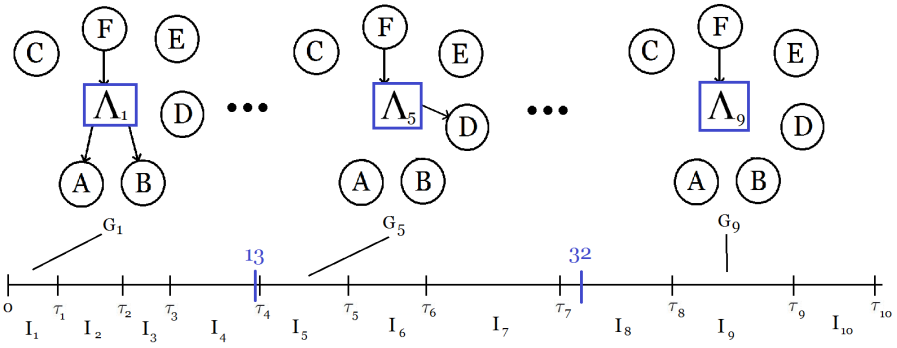

Figure 3: Graphical structures learned by DHBN, for the synthetic dataset.
Here we show the graphical structures for interval I_1, $I_5$and I_9 – based on one fold.

Table 2: The Markov blanket around the "Death" variable Λ_j in the Bayesian network structures for synthetic data. The changing time points are at t1 = 13 and t2 = 32.

Interval id	Time intervals	Connected nodes
1.	(0, 2.17)	A, B, F
2.	(2.17, 4.49)	A, F
3.	(4.49, 8.2)	A, B, F
4.	(8.2, 13.1)	A, B
5.	(13.1, 17.9)	D, F
6.	(17.9, 23.7)	D, E
7.	(23.7, 30.4)	D, F
8.	(30.4, 37.1)	{}
9.	(37.1, 46.6)	F
10.	(46.6, 59.4)	{}

Table 3: The parameters in Cox proportional hazard models for synthetic data.

Variables	Coefficients
A	0.19480
B	0.05259
C	-0.02280
D	0.21360
E	-0.07883
F	0.11807

Table 4: Results on the synthetic dataset

	Concordance	D-calibration	L1 loss
CoxKP	0.60121614	0.0000000035	34.569668
MTLR	0.60568394	**0.444995**	34.57228
DHBN	**0.63226048**	**0.07371909**	**34.221828**

factors found by DHBN, showing for example, that the "STAGE-4" variable is important (only) at the beginning, and after time 42 months, "AGE" is critical. Chi et al. (2017) similarly found that the effects of stage and age are time-dependent. We also see that "LDH_SERUM" is important in 13 to 19 month, etc. Table 6 shows the concordance, L1-loss, and D-calibration. Seeing concordance > 0.5 shows that the DHBN can provide some basic discrimination, but still worse than MTLR and Cox-KP.

5.4. Norway/Stanford breast cancer dataset

The Norway/Stanford breast cancer dataset (NSBCD) (Sørlie et al., 2003) contains 115 instances, 274 features, and 66% censoring. Table 7 shows that all three models are calibrated; DHBN has the lowest L1 loss and MTLR has the highest concordance. Table 8 shows the features that DHBN finds relevant at different times.

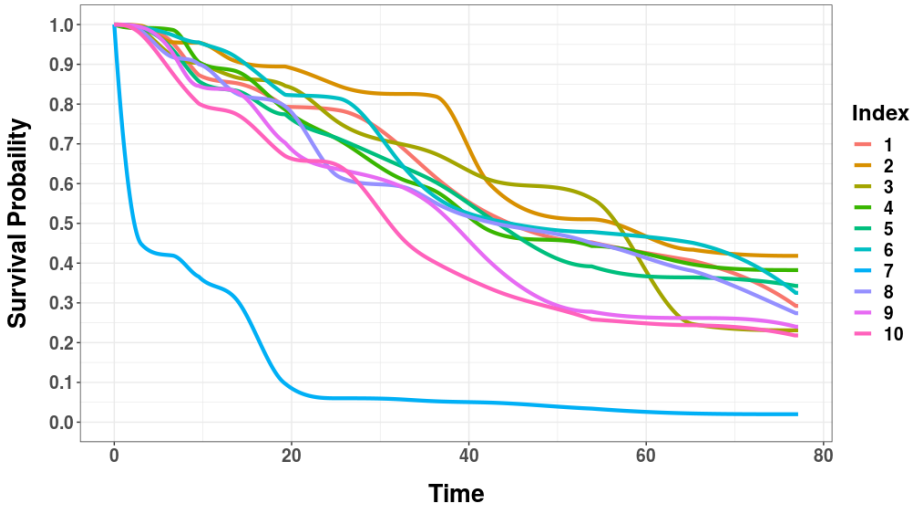

Figure 4: Survival curves for some colorectal patients, produced by DHBN

6. Conclusion

We list below three future directions that might be worth investigating. First, because the survival curve $S_i(t)$ is the product of all the discrete hazards before time t, the errors at earlier times might accumulate, leading to larger errors at later times. We are considering extensions that address this, perhaps by combining both hazard prediction with survival prediction. Second, as the number of subjects that are alive decrease over time, there may be only a small number of instances available for learning a Bayesian network at the later time intervals. To include more data at later time, we plan to explore ways to "fractionally include" instances. Finally, here we have fixed the number of time intervals, $N = 10$. However, the number of time intervals could be based on the amount of training data and the number of features, motivated us to develop an automatic way to divide the time axis into discrete intervals.

Table 5: NACD colorectal cancer features that DHBN considers relevant, for different times.

j	Time intervals	Features
1	(0, 3.03)	PERFORMANCE_STATUS_3, STAGE_4, GRANULOCYTES
2	(3.03, 6.56)	DRY_MOUTH, STAGE_4
3	(6.56, 9.50)	STAGE_4
4	(9.50, 13.5)	PERFORMANCE_STATUS_3, STAGE_4
5	(13.5, 19.1)	STAGE_4, LDH_SERUM
6	(19.1, 25.4)	PERFORMANCE_STATUS_2, STAGE_4
7	(25.4, 32.6)	{}
8	(32.6, 36.2)	{}
9	(36.2, 42.3)	HGB
10	(42.3, 53.9)	AGE65

Table 6: Results for North Alberta Cancer Dataset colorectal cancer

	Concordance	D-Calibration	L1 loss
CoxKP	**0.72188**	0.52085	**27.3413**
MTLR	0.72101	0.86689	28.0722
DHBN	0.7127766	0.394343	29.7309

Table 7: Results for NSBCD Dataset

	Concordance	D-Calibration	L1 loss
CoxKP	0.6166667	5.09E-60	264.02
MTLR	**0.7833333**	**0.9975128**	227.24
DHBN	0.6916667	**0.9906823**	**156.35**

Table 8: NSBCD features that DHBN considers relevant, for different times.

j	Time intervals	Features
1	(0, 7.33)	f088
2	(7.33, 9.66)	{}
3	(9.66, 12.0)	{}
4	(12.0, 14.3)	f011, f271
5	(14.3, 17.6)	{}
6	(17.6, 22.0)	{}
7	(22.0, 29.3)	f120, f458
8	(29.3, 38.0)	f400
9	(38.0, 55.3)	f420, f448, f459, f488, f511
10	(55.3, 72.6)	{}

To conclude: Finding relevant features for different times can provide much more information to help doctors make better decisions. We introduce the DHBN survival prediction model that produces individual survival curves for each subject, and identifies the factors important to the hazard at different times. We demonstrate that this model works effectively, on both synthetic and real-world datasets.

References

Carine A Bellera, Gaëtan MacGrogan, Marc Debled, Christine Tunon de Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. Variables with time-varying effects and the cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10(1):20, 2010.

Sheng-Qiang Chi, Yu Tian, Jun Li, Dan-yang Tong, Xiang-Xing Kong, Graeme Poston, Ke-Feng Ding, and Jing-Song Li. Time-dependent and nonlinear effects of prognostic factors in nonmetastatic colorectal cancer. *Cancer medicine*, 6(8):1882–1892, 2017.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63, 2020.

David Heckerman. A tutorial on learning with bayesian networks. In *Innovations in Bayesian networks*, pages 33–82. Springer, 2008.

Jian Huang and Shuangge Ma. Variable selection in the accelerated failure time model via the bridge method. *Lifetime data analysis*, 16(2):176–195, 2010.

Chinelo Mercy Igwenagu. The application of life table functions: A demographic study. 1993.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Wenshuo Liu, Cooper Stansbury, Karandeep Singh, Andrew M Ryan, Devraj Sukul, Elham Mahmoudi, Akbar Waljee, Ji Zhu, and Brahmajee K Nallamothu. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PloS one*, 15(4):e0221606, 2020.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

Fabian Sailer, Monika Pobiruchin, Sylvia Bochum, Uwe M Martens, and Wendelin Schramm. Prediction of 5-year survival with data mining algorithms. In *ICIMTH*, pages 75–78, 2015.

Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the national academy of sciences*, 100(14):8418–8423, 2003.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Xiaonan Xue, Xianhong Xie, Marc Gunter, Thomas E Rohan, Sylvia Wassertheil-Smoller, Gloria YF Ho, Dominic Cirillo, Herbert Yu, and Howard D Strickler. Testing the proportional hazards assumption in case-cohort analysis. *BMC medical research methodology*, 13(1):88, 2013.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853, 2011.