Journal of Machine Learning Research 24 (2023) 1-26 Submitted 12/19; Revised 11/22; Published 1/23

The Brier Score under Administrative Censoring: Problems and a Solution

Håvard Kvamme HAAVAKVA@MATH.UIO.NO
Ørnulf Borgan BORGAN@MATH.UIO.NO
Department of Mathematics
University of Oslo
P.O. Box 1053 Blindern
0316 Oslo, Norway

Editor: Jon McAuliffe

Abstract

The Brier score is commonly used for evaluating probability predictions. In survival analysis, with right-censored observations of the event times, this score can be weighted by the inverse probability of censoring (IPCW) to retain its original interpretation. It is common practice to estimate the censoring distribution with the Kaplan-Meier estimator, even though it assumes that the censoring distribution is independent of the covariates. This paper investigates problems that may arise for the IPCW weighting scheme when the covariates used in the prediction model contain information about the censoring times. In particular, this may occur for administratively censored data if the distribution of the covariates varies with calendar time. For administratively censored data, we propose an alternative version of the Brier score. This administrative Brier score does not require estimation of the censoring distribution and is valid also when the censoring times can be predicted from the covariates.

Keywords: survival analysis, time-to-event prediction, customer churn, inverse probability weighting, progressive type I censoring, time-dependent case mix

1. Introduction

Recently, there has been an increasing interest in combining machine learning methodology with survival analysis for improved time-to-event prediction. Some methods extend the well known Cox regression with neural networks (Katzman et al., 2018; Ching et al., 2018; Yousefi et al., 2017; Luck et al., 2017; Kvamme et al., 2019), while others consider a more direct approach for optimizing the likelihood for right-censored time-to-event data (Biganzoli et al., 1998; Lee et al., 2018; Fotso, 2018; Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2019). Also worth mentioning is the Random Survival Forest (Ishwaran et al., 2008) which makes decision trees based on the log-rank test and estimates the cumulative hazards with the Nelson-Aalen estimator.

In classical statistical theory, the focus is often on model assumptions and the properties of statistical methods. In machine learning research, however, it is common to put more emphasis on the predictive performance of the methods. Both perspectives do of course have their virtues. But we note that while there is currently much research on applying machine learning methodology for time-to-event prediction, the criteria used for evaluating these prediction methods have received less attention in the machine learning community.

Although survival analysis is applied in a number of areas, the field has very much been driven by medical research. As a consequence of this, censoring is typically treated as a random event that can occur for a number of reasons such as an individual choosing to drop out of the study, the death by a competing cause, or survival past the end of the study. The latter form of censoring is known as administrative censoring, and in many industrial applications, e.g., customer churn (Section 2), this is the only censoring present.

The two most common evaluation criteria for survival predictions are arguably the inverse probability of censoring weighted (IPCW) Brier score (Graf et al., 1999; Gerds and Schumacher, 2006) and different versions of the concordance index (Harrell Jr et al., 1982; Antolini et al., 2005; Uno et al., 2011; Gerds et al., 2013). We will, in this paper, show that the IPCW Brier score can be biased in situations with staggered entry and administrative censoring if there is no loss to follow up and the covariates used in the prediction model contain information about the administrative censoring times. As we will discuss further in Section 3, such situations are much more likely to occur for industrial than medical applications. Furthermore, we will show that due to this bias, approaching the prediction problem by naively applying a binary classifier to the uncensored subset of the data, can result in better scores than more reasonable event-time modeling. In this paper we identify why such issues occur, and we propose a new version of the Brier score, the *administrative Brier score*, that may serve as an alternative to the IPCW score for administratively censored event times.

To give the reader an understanding of the potential problems of the IPCW Brier score, we will start with an illustrative example in Section 2. Then, in Section 3, we will introduce the Brier score in detail and discuss more carefully the issues of the IPCW scheme. We present our proposed alternative, the administrative Brier score, in Section 3.4. Binary classifiers are investigated in Section 4, where we will show their relationship to the potential bias of the IPCW Brier score under administrative censoring. A simulation study is conducted in Section 5 to empirically illustrate our findings, and in Section 6, we investigate a real data set with administrative censoring. A summary and concluding remarks are made in Section 7. Two appendices give some theoretical results for the administrative Brier score and the binary classifier.

The code for the evaluation metrics, the survival methods, the simulations, and the data sets are available at `github.com/havakv/pycox`.

2. A Real-World Example

To illustrate the potential issues with the IPCW Brier score, we consider an example encountered while researching the KKBox Churn data set in Kvamme et al. (2019). The task is to predict whether or not customers continue to subscribe to the KKBox music streaming service t days after their first subscription. If customers leave their subscription they have churned, and these are the events we want to model.

An illustration of how censored event times arise in the KKBox data set is given in Figure 1. Panel A shows the observations for five (hypothetical) customers, labeled 1–5, as they occur in calendar time. The customers subscribe to the streaming service at different

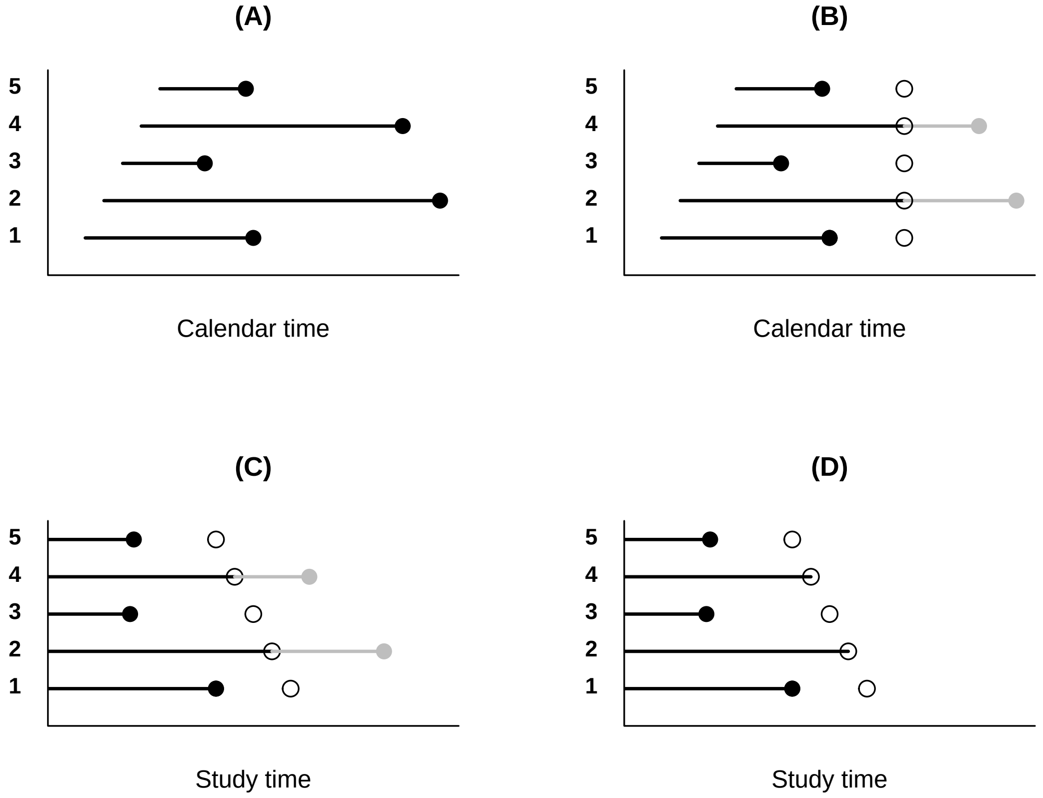

Figure 1: Illustration of how censored observations arise in the KKBox data set. See the text for an explanation.

dates, and are then followed up over time. In panel A it is assumed that all customers are followed until they churn, which is marked with filled circles (●) in the figure. However, in practice, the customers are only followed up until a given date, which for the KKBox data set was January 29, 2017. This is illustrated in panel B, where the open circles (○) indicate the date for end of follow-up. Customers 2 and 4 have not churned by this date, so what happen to them after the end of the study period, marked with grey in the figure, is not known to us. For time-to-event prediction, we do not use calendar time, but rather the time t since subscription to the service. Panel C shows the information of panel B in this study time scale. Finally, in panel D, we omit the grey lines and circles for customers 2 and 4, and show only the information that is actually available to us. In this panel the lines with filled circles correspond to observed events (customers 1, 3, and 5), while the lines with open circles correspond to censored observations (customers 2 and 4). In addition, we know the maximal follow-up times also for the churned customers. So if they had not left their subscription, we would have known when they had been censored. This is marked

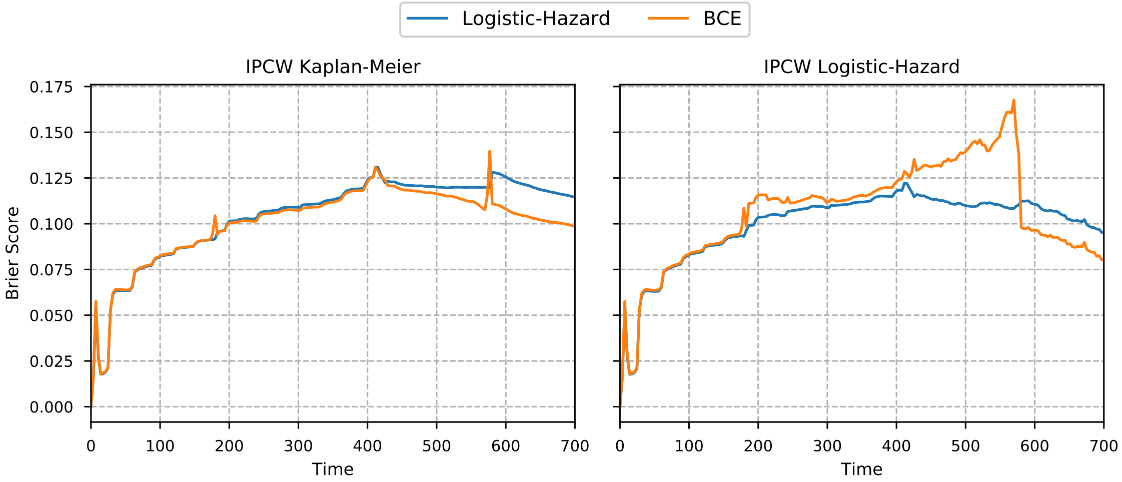

Figure 2: Brier scores of survival estimates from the Logistic-Hazard method and the BCE method on the KKBox data set. The Brier scores are weighted by the inverse probability of censoring estimated with Kaplan-Meier (left) and with a Logistic-Hazard model (right).

in panel D with the open circles for customers 1, 3, and 5. The type of right-censoring described here is called *administrative censoring* or *progressive type I censoring*.

We approach the modeling of the event-time distribution for the KKBox data set in two ways. The first is with the Logistic-Hazard method (Brown, 1975; Biganzoli et al., 1998; Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2019) which accounts for censored observations by considering the likelihood for right-censored event times. We use the version of the method described by Kvamme and Borgan (2019), meaning that we parameterize the discrete hazards with a neural network in the form of a multilayer perceptron (MLP).

The second approach is to fit a binary classifier for each time t and remove all customers censored before this time. The responses (labels) given to the classifiers are indicators of whether or not each customer has churned. We denote this as the BCE method, as it minimizes the binary cross-entropy of the survival estimates, where survival means that a customer has not yet churned. The BCE method is an MLP with equivalent network structure to that of the Logistic-Hazard, with each output node corresponding to a binary classifier at time t. The method will be described in detail in Section 4. As censored individuals are removed from the data set at their time of censoring, the BCE method has a bias towards higher churn probabilities. We would, therefore, expect the Logistic-Hazard to perform better for the KKBox data set.

The IPCW Brier scores with Kaplan-Meier censoring estimates (Graf et al., 1999) are calculated for the survival estimates of the two methods, and a plot of the results is shown in the left panel of Figure 2. For higher times, we see that the BCE method has better scores than the Logistic-Hazard method. This is not expected, as the censoring proportion increases with time, meaning the bias of the BCE estimates increases with time. However, the Kaplan-Meier censoring estimates are not covariate dependent, so one might argue that the results could be explained by poor estimation of the censoring distribution, a problem that has been addressed by Gerds and Schumacher (2006).

In the right hand panel of Figure 2, we have plotted the IPCW Brier scores where the censoring distribution is estimated with a Logistic-Hazard model with the same hyperparameters and covariates as those used to obtain the survival estimates. We now see that the BCE generally performs worse than the Logistic-Hazard, but we still find that it gets better scores than the Logistic-Hazard for the largest follow-up times. In this paper, we will argue that the Logistic-Hazard likely gives better survival estimates than the BCE method, and the results above might be explained by a weakness in how the IPCW Brier score handles predictions that are biased towards zero after the administrative censoring time. We will, also, propose the *administrative Brier score*, which does not suffer from the same vulnerabilities as the IPCW Brier score, and does not even require estimation of the censoring distribution. In Section 6, we will revisit the KKBox data set for a more in-depth analysis.

3. Brier Scores

In the following, we present Brier scores for evaluating time-to-event predictions in the form of survival estimates. We also introduce the topic of right-censoring in survival analysis, and show how the IPCW Brier score accounts for censored observations. This is followed by a discussion of administratively censored observations and how they affect the IPCW Brier score. We end the section by introducing a new Brier score for handling administratively right-censored observations, and discuss problems concerning estimation of the censoring distribution.

3.1 The Brier Score for Uncensored Data

We start out with the situation without censoring, and assume that we have event times $T_1^*, T_2^*, \dots, T_n^*$ for n independent individuals. The distribution of the event time for individual i depends on a vector of covariates $\mathbf{x}_i$. So if we denote the density function of T_i^* by $f(t \,|\, \mathbf{x}_i)$, the survival function of individual i is given by

$$S(t \,|\, \mathbf{x}_i) = \mathrm{P}(T_i^* > t \,|\, \mathbf{x}_i) = \int_t^\infty f(u \,|\, \mathbf{x}_i)\, du.$$

Note that in observational studies, like the KKBox study, the vector of covariates $\mathbf{x}_i$ is the observed value of a random vector $\mathbf{X}_i$. Then the pairs $(T_i^*, \mathbf{X}_i)$; $i = 1, \dots, n$; are assumed independent. But, as further addressed in Sections 3.3 and 3.5, we do not assume that the $(T_i^*, \mathbf{X}_i)$'s are identically distributed. For such situations, $f(t \,|\, \mathbf{x}_i)$ and $S(t \,|\, \mathbf{x}_i)$ are the *conditional* density and *conditional* survival function of T_i^* given $\mathbf{X}_i = \mathbf{x}_i$. However, as is common in regression modeling, we will in the following consider the observed $\mathbf{x}_i$'s as fixed, non-random quantities, and omit the term "conditional" when we talk about density functions, survival functions, and similar quantities.

In time-to-event prediction, we want to estimate the survival functions for all individuals, and we let $\pi(t \,|\, \mathbf{x}_i)$ denote the estimate of the survival function for individual i. To simplify the presentation, we will for now disregard the estimation uncertainty, and consider the $\pi(t \,|\, \mathbf{x}_i)$'s as known (non-random) functions.

A reasonable metric for evaluating the predictive performance of the $\pi(t \,|\, \mathbf{x}_i)$'s, is to calculate the mean squared error of the estimates to the true survival functions:

$$\mathrm{MSE}(t, \pi) = \frac{1}{n} \sum_{i=1}^{n} \left[S(t \,|\, \mathbf{x}_i) - \pi(t \,|\, \mathbf{x}_i)\right]^2. \tag{1}$$

However, the $S(t \,|\, \mathbf{x}_i)$'s are not known outside of simulations; what we observe are the event times T_i^* drawn from the event time distributions. The Brier score for uncensored data approximates the true survival functions with step-functions with jumps at the event times, giving

$$\begin{aligned}\mathrm{BS}(t, \pi) &= \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)\right]^2 \qquad (2)\\ &= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i^* \leq t\} + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i^* > t\}\Big].\end{aligned}$$

The expectation of the Brier score for uncensored data is

$$\begin{aligned}\mathbb{E}\left[\mathrm{BS}(t, \pi)\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i^* \leq t\} + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i^* > t\}\Big]\\ &= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2\, \mathrm{P}(T_i^* \leq t \,|\, \mathbf{x}_i) + [1 - \pi(t \,|\, \mathbf{x}_i)]^2\, \mathrm{P}(T_i^* > t \,|\, \mathbf{x}_i)\Big]\\ &= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2\, [1 - S(t \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2\, S(t \,|\, \mathbf{x}_i)\Big] \qquad (3)\\ &= \frac{1}{n} \sum_{i=1}^{n} \Big\{[S(t \,|\, \mathbf{x}_i) - \pi(t \,|\, \mathbf{x}_i)]^2 + S(t \,|\, \mathbf{x}_i)\, [1 - S(t \,|\, \mathbf{x}_i)]\Big\}\\ &= \mathrm{MSE}(t, \pi) + \frac{1}{n} \sum_{i=1}^{n} S(t \,|\, \mathbf{x}_i)\, [1 - S(t \,|\, \mathbf{x}_i)].\end{aligned}$$

So the expected Brier score is the sum of the MSE in (1) and a constant that does not depend on our survival estimates $\pi(t \,|\, \mathbf{x}_i)$. This constant is the irreducible error of approximating the true survival functions $S(t \,|\, \mathbf{x}_i)$ with the step-functions $\mathbb{1}\{T_i^* > t\}$. So minimizing the expected Brier score is equivalent to minimizing the MSE, and the minimum is obtained for the true survival functions, i.e., $\pi(t \,|\, \mathbf{x}_i) = S(t \,|\, \mathbf{x}_i)$.

3.2 The IPCW Brier Score for Right-Censored Data

In most applications, only a subset of the event times T_i^* is observed. For some individuals, we will only know that the event time occurs after some censoring time C_i^*. As is common in survival analysis, we assume that T_i^* and C_i^* are conditionally independent given the vector of covariates, and we let $G(t \,|\, \mathbf{x}_i) = \mathrm{P}(C_i^* > t \,|\, \mathbf{x}_i)$ denote the survival distribution of the censoring time. For ease of notation we use $\mathbf{x}_i$ to denote the vector of covariates for the survival distributions of both T_i^* and C_i^*. But we note that different components of $\mathbf{x}_i$ may be of importance for the two distributions.

For data sets with right-censoring, we consider the right-censored event times $T_i = \min\{T_i^*, C_i^*\}$ and the event indicators $D_i = \mathbb{1}\{T_i^* \leq C_i^*\}$; $i = 1, 2, \ldots, n$. We here follow the common convention in survival analysis that when an event and censoring time coincide, we observe the occurrence of the event.

As we now only have partial information, the Brier score (2) cannot be calculated. We can, however, approximate it by weighting the scores of the observed event times by the inverse probability of censoring (Graf et al., 1999; Gerds and Schumacher, 2006). This is called *inverse probability of censoring weighting* (IPCW), and for $G(t \,|\, \mathbf{x}_i) > 0$ the IPCW Brier score is given by

$$\mathrm{BS}_{\mathrm{IPCW}}(t, \pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right], \tag{4}$$

where $G(t - \,|\, \mathbf{x}_i) = \mathrm{P}(C_i^* \geq t \,|\, \mathbf{x}_i)$. By the weighting in (4), each individual in the sum can be considered to represent more than one individual. If, e.g., $G(t \,|\, \mathbf{x}_i) = 0.50$, an individual i with $T_i > t$ represents $1/G(t \,|\, \mathbf{x}_i) = 1/0.50 = 2$ individuals (itself plus one individual that has been censored before time t).

In practice, one has to estimate the $G(t \,|\, \mathbf{x}_i)$'s; cf. Section 3.5, but for simplicity we will for now assume that the $G(t \,|\, \mathbf{x}_i)$'s are known functions. Then the expected value of the IPCW Brier score may be given as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{BS}_{\mathrm{IPCW}}(t, \pi)\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \pi(t \,|\, \mathbf{x}_i)^2 \, \mathbb{E}\left[\frac{\mathbb{1}\{T_i^* \leq t, T_i^* \leq C_i^*\}}{G(T_i^* - \,|\, \mathbf{x}_i)} \right] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, \frac{\mathrm{P}(T_i^* > t, C_i^* > t \,|\, \mathbf{x}_i)}{G(t \,|\, \mathbf{x}_i)} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \pi(t \,|\, \mathbf{x}_i)^2 \int_0^t \frac{G(u - \,|\, \mathbf{x}_i)\, f(u \,|\, \mathbf{x}_i)}{G(u - \,|\, \mathbf{x}_i)} \, du + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, \frac{G(t \,|\, \mathbf{x}_i)\, S(t \,|\, \mathbf{x}_i)}{G(t \,|\, \mathbf{x}_i)} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \pi(t \,|\, \mathbf{x}_i)^2 \left[1 - S(t \,|\, \mathbf{x}_i)\right] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, S(t \,|\, \mathbf{x}_i) \right\}.
\end{aligned} \tag{5}$$

We see that (5) is identical to the expected value (3) of the Brier score for uncensored data, so the IPCW Brier score is a reasonable approximation of the uncensored Brier Score.

3.3 The IPCW Brier Score with Administrative Censoring

In Section 2, we discuss how right-censored time-to-event data occur when individuals are recruited to a study population at different calendar times (i.e., we have staggered entry), and then followed up to the occurrence of an event of interest or to the closure of the study at a given date. As is commonly done, we will use time since entry as the time-scale for statistical modeling and time-to-event prediction. Then, when closure of the study is the only reason for censoring, the censoring time C_i^* for individual i is the difference between the closure time and the entry time of the individual. For such administrative censoring, *all* the censoring times C_i^* are observed, regardless of whether an individual experiences the event of interest or is censored, cf. Figure 1.D.

One may envisage two strategies for obtaining an IPCW Brier score for administratively censored time-to-event data. One option is to consider C_i^* as a random variable. Its survival distribution $G(t \,|\, \mathbf{x}_i) = \mathrm{P}(C_i^* > t \,|\, \mathbf{x}_i)$ may in general depend on the vector of covariates $\mathbf{x}_i$. One may then estimate the $G(t \,|\, \mathbf{x}_i)$'s as will be described in Section 3.5 and use the IPCW Brier score (4). The covariates will typically describe characteristics of the individual, like age and gender. But, in particular in industrial applications, they may also describe properties of the recruitment process, as will be the case for customer churn data when a campaign takes place in a limited period of time. In the latter case, the distribution of the covariates (when considered as observed values of random variables; cf. Section 3.1) will depend on the calendar time of inclusion to the study, and we will have a time-dependent case mix.

Another option is to condition on the observed values of the censoring times. We then use the fact that we know that the censoring time C_i^* for individual i takes the value c_i^*. This corresponds to using a one-point censoring distribution for C_i^* with survival distribution $\mathbb{1}\{c_i^* > t\}$. If we insert $\mathbb{1}\{c_i^* > t\}$ for $G(t \,|\, \mathbf{x}_i)$ in (4), the IPCW Brier score becomes

$$\mathrm{BS}_C(t, \pi) = \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \le t, D_i = 1\} + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\} \Big] . \qquad (6)$$

This is proportional to the unweighted Brier score (2) for the subset of individuals that are not censored, meaning that the score (6) just disregards the set of censored individuals $\{i : T_i \le t, D_i = 0\}$. As this is not a sensible thing to do, one would in practice not use the score (6). Nevertheless, it is of interest to study its properties. One reason for this is that (6) can be seen as an approximation of the IPCW Brier score (4) when (estimates of) the $G(t \,|\, \mathbf{x}_i)$'s are close to the step functions $\mathbb{1}\{c_i^* > t\}$. As further discussed in Section 3.5, this may be the case when there is enough information in the covariates $\mathbf{x}_i$ to predict the censoring times c_i^*.

If we denote the conditional probability and conditional expectation given the censoring times by P_C and $\mathbb{E}_C$, and use that survival and censoring times are independent (given covariates), we may obtain the conditional expectation of (6) given that $C_i^* = c_i^*$ for $i = 1, 2, \ldots, n$ as follows:

$$\begin{aligned}
\mathbb{E}_C \left[\mathrm{BS}_C(t, \pi)\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_C \Big[\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \le t, D_i = 1\} + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\} \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \, \mathrm{P}_C(T_i \le t, D_i = 1 \,|\, \mathbf{x}_i) + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, \mathrm{P}_C(T_i > t \,|\, \mathbf{x}_i) \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \, \mathrm{P}(T_i^* \le \min\{t, c_i^*\} \,|\, \mathbf{x}_i) + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, \mathrm{P}(T_i^* > t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\} \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \; [1 - S_i(\min\{t, c_i^*\} \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\} \Big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^{n} \Big[\pi(t \,|\, \mathbf{x}_i)^2 \,[1 - S(t \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, S(t \,|\, \mathbf{x}_i)\Big] \, \mathbb{1}\{c_i^* > t\} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \pi(t \,|\, \mathbf{x}_i)^2 \,[1 - S_i(c_i^* \,|\, \mathbf{x}_i)] \, \mathbb{1}\{c_i^* \leq t\}.
\end{aligned}
\tag{7}
$$

Note that the conditional expectation (7) is only equal to the expected value (3) of the uncensored Brier score if all individuals have censoring times $c_i^* > t$. For this group, with $c_i^* > t$, we still have that the minimizer of the expected score is $\pi(t \,|\, \mathbf{x}_i) = S(t \,|\, \mathbf{x}_i)$. However, for the other group, with $c_i^* \leq t$, we see that the minimizer is $\pi(t \,|\, \mathbf{x}_i) = 0$. Hence, according to the score (6), the optimal survival estimates would be $\pi(t \,|\, \mathbf{x}_i) = S(t \,|\, \mathbf{x}_i) \, \mathbb{1}\{c_i^* > t\}$. As we would like to have an evaluation metric that decreases when our survival predictions approach the true survival functions $S(t \,|\, \mathbf{x}_i)$, this is a problematic property of the score (6).

More generally, we obtain from (7) that for any predictors $\pi(t \,|\, \mathbf{x}_i)$ of the survival functions, the conditional expectation of (6) becomes smaller if we replace the predictors by

$$
\pi_{\mathrm{c}^*}(t \,|\, \mathbf{x}_i) = \pi(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}. \tag{8}
$$

A similar result holds for the IPCW Brier score (4). To see that this is the case, we in (4) replace $\pi(t \,|\, \mathbf{x}_i)$ by (8). Then the IPCW Brier score becomes

$$
\begin{aligned}
&\mathrm{BS}_{\mathrm{IPCW}}(t, \pi_{c^*}) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\pi_{\mathrm{c}^*}(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi_{\mathrm{c}^*}(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{[\pi(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}]^2 \mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}]^2 \mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\pi(t \,|\, \mathbf{x}_i)^2 \, \mathbb{1}\{T_i^* \leq t < c_i^*\}}{G(T_i^* - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi(t \,|\, \mathbf{x}_i)]^2 \, \mathbb{1}\{T_i^* > t, c_i^* > t\}}{G(t \,|\, \mathbf{x}_i)} \right] \\
&\quad = \mathrm{BS}_{\mathrm{IPCW}}(t, \pi) - \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i^* \leq c_i^* \leq t\}}{G(T_i - \,|\, \mathbf{x}_i)} \\
&\quad \leq \mathrm{BS}_{\mathrm{IPCW}}(t, \pi).
\end{aligned}
$$

The inequality is strict if there are individuals with $T_i^* \leq c_i^* \leq t$ and $\pi(t \,|\, \mathbf{x}_i) > 0$. In particular, it follows that the predictors $S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}$ give a smaller IPCW Brier score than the true survival functions $S(t \,|\, \mathbf{x}_i)$.

3.4 The Administrative Brier Score

We have shown that the IPCW Brier score may have undesirable behavior under administrative censoring. We here propose an alternative that may be useful in such situations. Our approach is to use the uncensored Brier score (2) and simply remove individuals from evaluation after their administrative censoring times. This is possible when there is no loss to follow up before the administrative censoring, as we then observe the censoring times for all individuals. Consequently, the score is not applicable for studies with loss to follow up, or for administratively censored data sets that do not provide the administrative censoring time for every individual. This excludes most medical data sets, making the score more targeted towards industrial applications like churn prediction.

To define the *administrative Brier score*, first note that when the observed censoring time c_i^* for individual i is at least t, then we know whether $T_i^* \leq t$ or $T_i^* > t$ (remember that if $T_i^* = c_i^*$ we assume that we observe the occurrence of the event). The administrative Brier score is then

$$\mathrm{BS}_{\mathrm{A}}(t, \pi) = \frac{1}{\tilde{n}_{\mathrm{A}}(t)} \sum_{i=1}^{n} \left[\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)\right]^2 \mathbb{1}\{c_i^* \geq t\}, \tag{9}$$

where we scale by the number individuals contributing to the score

$$\tilde{n}_{\mathrm{A}}(t) = \sum_{i=1}^{n} \mathbb{1}\{c_i^* \geq t\}$$

instead of n. Note that the score (9) only use information from individuals with an observed event ($D_i = 1$) when $t \leq c_i^*$. This is different from the naive (and biased) score one obtains by removing the set of censored individuals $\{i : T_i < t,\, D_i = 0\}$, as given by (6), which include information for individuals with an observed event ($D_i = 1$) also after the administrative censoring time c_i^*.

The conditional expectation of the administrative Brier score, given that $C_i^* = c_i^*$ for $i = 1, 2, \ldots, n$, is given by

$$\begin{aligned}\mathbb{E}_C\left[\mathrm{BS}_{\mathrm{A}}(t, \pi)\right] &= \frac{1}{\tilde{n}_{\mathrm{A}}(t)} \sum_{i=1}^{n} \mathbb{E}_C\left[\left[\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)\right]^2\right] \mathbb{1}\{c_i^* \geq t\} \\ &= \frac{1}{\tilde{n}_{\mathrm{A}}(t)} \sum_{i=1}^{n} \left[\pi(t \,|\, \mathbf{x}_i)^2 \left[1 - S(t \,|\, \mathbf{x}_i)\right] + \left[1 - \pi(t \,|\, \mathbf{x}_i)\right]^2 S(t \,|\, \mathbf{x}_i)\right] \mathbb{1}\{c_i^* \geq t\},\end{aligned}$$

which, for the subset of individuals with $c_i^* \geq t$, is the same as for the uncensored Brier score. Individuals with $c_i^* < t$ give no contribution to the score. Note that this means that the score is not able to penalize survival predictions that deviate from $S(t \,|\, \mathbf{x}_i)$ for $t > c_i^*$. But contrary to the IPCW scores, it does not reward biased predictions that are smaller than $S(t \,|\, \mathbf{x}_i)$ for $t > c_i^*$.

In Appendix A, we derive the approximation (A.2) for the unconditional expected value of the administrative Brier score, i.e., the expected value when we also average over the distribution of the censoring times. We further show that the expected value is the same for the predictors $\pi(t \,|\, \mathbf{x}_i)$ and the predictors $\pi_{c^*}(t \,|\, \mathbf{x}_i)$ given by (8). For the special case where the censoring distributions do not depend on covariates, i.e., $G(t \,|\, \mathbf{x}_i) = G(t)$ for all $i = 1, 2, \ldots, n$, the expected value of the administrative Brier score is approximately equal to the expected value of the Brier score when there is no censoring and the expected value of the IPCW Brier score; cf. (3) and (5). This means that when censoring does not depend on covariates, the administrative Brier score will yield similar scores to both the uncensored Brier score and the IPCW Brier score.

3.5 Estimation of the Censoring Distribution

In practice, we need to estimate the censoring distributions $G(t \,|\, \mathbf{x}_i)$ to use the IPCW Brier score (4). This estimation is, in itself, a time-to-event prediction problem, and can be addressed in the same manner as the original time-to-event problem. Graf et al. (1999) proposed to use the Kaplan-Meier estimates of the censoring distribution, and this is still the most common approach. However, the Kaplan-Meier estimator disregards the covariates, meaning that all individuals are assumed to have the same censoring distribution. This can lead to biased censoring estimates, as addressed by Gerds and Schumacher (2006).

In predictive modeling, we typically split our data set into a training set used to fit the models, a validation set used for hyperparameter tuning, and a test set used for evaluating the models' predictions. We, therefore, only consider Brier scores calculated on a test set in this paper. Both Graf et al. (1999) and Gerds and Schumacher (2006) use the test set to estimate the censoring distribution, which is reasonable when simple methods such as the Kaplan-Meier estimator and Cox regression are used. If we, however, want to use more flexible methods, such as the Logistic-Hazard with neural networks, fitting to the test set is likely to results in overfitted censoring estimates. To the best of our knowledge, this topic has not been addressed in the literature, so there are no "best practices" for how to approach such estimation problems. In this paper, we will treat the censoring distribution in the same manner as the event-time distribution, meaning we fit the censoring model to the training set, and use a validation set to verify that the estimates are reasonable.

When the censoring distribution is obtained from the Kaplan-Meier estimator on the test set, the weights of the IPCW Brier score (4) sum to n. But for more flexible methods, and methods fitted to the training set, this is not necessarily the case. We therefore use a slightly modified version of the IPCW Brier score where n is replaced by the sum of the weights

$$\tilde{n}(t) = \sum_{i=1}^{n} \left[\frac{\mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{\mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right]. \tag{10}$$

This ensures that the modified IPCW Brier score

$$\mathrm{BS}_{\mathrm{IPCW}}^{(\tilde{n})}(t, \pi) = \frac{1}{\tilde{n}(t)} \sum_{i=1}^{n} \left[\frac{\pi(t \,|\, \mathbf{x}_i)^2 \mathbb{1}\{T_i \leq t, D_i = 1\}}{G(T_i - \,|\, \mathbf{x}_i)} + \frac{[1 - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{T_i > t\}}{G(t \,|\, \mathbf{x}_i)} \right]. \tag{11}$$

is between 0 and 1. In a similar manner we replace n in the score (6) by

$$\tilde{n}_C(t) = \sum_{i=1}^{n} \left[\mathbb{1}\{T_i \leq t, D_i = 1\} + \mathbb{1}\{T_i > t\} \right] \tag{12}$$

These are the versions of the scores (4) and (6) we use in all our experiments. We will, however, continue to refer to $\mathrm{BS}_{\mathrm{IPCW}}^{(\tilde{n})}(t, \pi)$ as the IPCW Brier score.

By an argument similar to the one giving (5), we may show that the expected value of (10) equals n. If we use this result and the argument in Appendix A, we may show that the expected value of the version (11) of the IPCW Brier score has (approximately) the same expected value as the classical version (4) of the score.

If there is enough information in the covariates $\mathbf{x}_i$ to identify the censoring times c_i^*, the estimates of the $G(t \,|\, \mathbf{x}_i)$'s will approach the step-functions $\mathbb{1}\{c_i^* > t\}$, meaning the scores approach the biased Brier score (6). As an example of this, consider a study where all censoring is due to administrative censoring at the closure of the study, and one includes the start date for each individual as a covariate, then the c_i^*'s can be identified. A more realistic example would be that a combination of certain covariates can predict the start date of some individuals and, consequently, a subset of the censoring times can be predicted. This may happen if the case mix of the individuals varies with time, as will be the case for customer churn data when a campaign takes place in a limited period of time.

If we estimate the censoring distribution with a flexible method, we might experience that some of the estimates of $G(t \,|\, \mathbf{x}_i)$ become very small. This corresponds to very large weights, meaning that a single individual can potentially dominate the score. To prevent this from occurring, we set a maximum weight allowed. As an example, if we have a maximum weight of 100, we do not allow estimates of $G(t \,|\, \mathbf{x}_i) < 0.01$, giving the interpretation that a single individual can maximally represent 100 individuals in the IPCW Brier score. In practice, this is ensured by setting weights larger than 100 to be 100. By decreasing the maximum allowed weight, we reduce the variance of the IPCW Brier score at the expense of introducing some bias.

4. Binary Classifiers for Time-to-Event Prediction

In machine learning, binary classifiers are sometimes used for time-to-event prediction instead of methods based on survival analysis methodology. A binary classifier for time-to-event prediction can be constructed for a given time t by disregarding individuals who were censored before time t and minimizing the binary cross-entropy (negative log-likelihood for Bernoulli data). This gives the loss function,

$$\begin{aligned}\text{loss}_{\text{BCE}}(t, \pi) &= -\sum_{i=1}^{n} \Big(\mathbb{1}\{T_i > t\} \log[\pi(t \,|\, \mathbf{x}_i)] + \mathbb{1}\{T_i \leq t,\, D_i = 1\} \log[1 - \pi(t \,|\, \mathbf{x}_i)] \Big) \qquad (13)\\ &= -\sum_{i=1}^{n} \Big(y_i \log[\pi(t \,|\, \mathbf{x}_i)] + (1 - y_i) \log[1 - \pi(t \,|\, \mathbf{x}_i)] \Big) \Big(1 - \mathbb{1}\{T_i \leq t,\, D_i = 0\} \Big).\end{aligned}$$

were the labels $y_i = \mathbb{1}\{T_i > t\}$ denote if the events happen after time t. If we want survival estimates for a range of times $\tau_1 < \tau_2 < \cdots < \tau_m$, we can fit a model for each τ_j. Alternatively, if we use a model for $\pi(\tau_j \,|\, \mathbf{x}_i)$ that can be estimated for multiple τ_j's simultaneously, such as a neural network with m output nodes, we can fit the model to the sum of the m losses

$$\text{loss}_{\text{BCE}}(\pi) = -\sum_{j=1}^{m} \sum_{i=1}^{n} \Big(\mathbb{1}\{T_i > \tau_j\} \log[\pi(\tau_j \,|\, \mathbf{x}_i)] + \mathbb{1}\{T_i \leq \tau_j,\, D_i = 1\} \log[1 - \pi(\tau_j \,|\, \mathbf{x}_i)] \Big). \tag{14}$$

We refer to this approach as the BCE method.

The binary classifiers and BCE method are clearly biased, as the removal of censored individuals decreases the survival estimates. In fact, if there is sufficient information in the covariates to identify the censoring times c_i^*, the survival estimates of the binary classifiers will approach

$$\pi(t \,|\, \mathbf{x}_i) = S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}. \tag{15}$$

The derivations leading to (15) are given in Appendix B. We recognize these estimates as the minimizers of the score (6), cf. Section 3.3. If the censoring times c_i^* can be identified from the covariates, and we have a sufficiently large data set, the estimates of the $G(t \,|\, \mathbf{x}_i)$'s will be close to the step functions $\mathbb{1}(c_i^* > t)$, and then the binary classifiers are essentially optimal for minimizing the IPCW Brier score. If the covariates only identify a subset of the c_i^*'s, it is not clear whether or not a binary classifier will give smaller IPCW Brier scores than a corresponding survival method. We believe this might explain why the BCE method for some values of t got a lower Brier score than the Logistic-Hazard method for the KKBox data set in Section 2.

4.1 Simulation with BCE

To illustrate the concerns addressed in the previous sections, we conduct a simple simulation study. We draw event times T_i^* with a constant hazard rate $h(t) = 0.0084$, the same for all individuals. The censoring times C_i^* are uniformly distributed over the interval from 0 to 100. We consider the censoring to be administrative, meaning that we observe all the censoring times c_i^*, regardless of whether they are larger or smaller than the event times. To mimic the effect of having covariates that can identify the censoring times, we consider a covariate obtained by standardizing the c_i^*'s.

We fit two neural networks with the BCE loss (14) to 10,000 simulated samples, the first without the covariate identifying the c_i^*'s and the second with this covariate. In Figure 3 we have plotted the estimated survival functions for six individuals with distinct censoring times c_i^*, marked by the vertical red dotted lines. The plots also contain the true survival function (in blue) for comparison. From the orange lines, we see a clear bias of the BCE method (binary classifier) applied to right-censored data, as it always underestimates the survival. On the other hand, the survival estimates obtained by the BCE method with censoring information, represented by the green lines, follow the true survival function reasonably well up till the censoring times and then fall to zero. This agrees well with the optimal estimates in (15).

Next, we investigate the resulting Brier scores of these survival estimates on a test set of size 10,000 drawn from the same event time and censoring distributions as the training set. In Figure 4 we have plotted the uncensored Brier score (top left), the IPCW Brier scores (11) with Kaplan-Meier estimates of the censoring distribution (top right), the IPCW scores (6) normalized by (12) with the one-point censoring distribution $\mathbb{1}\{c_i^* > t\}$ (bottom left), and the proposed administrative Brier score given by (9) (bottom right). The uncensored Brier score is computed on an uncensored test set, while the three other scores are computed on a censored test set with the same censoring distribution as in the training set.

From the uncensored Brier score in Figure 4, it is clear that the scores of the true survival function are lower than those of the two BCE methods, and the BCE method with the censoring time covariate has the highest scores. For the true survival function and the BCE method without censoring information (blue and orange curves), we see that the IPCW

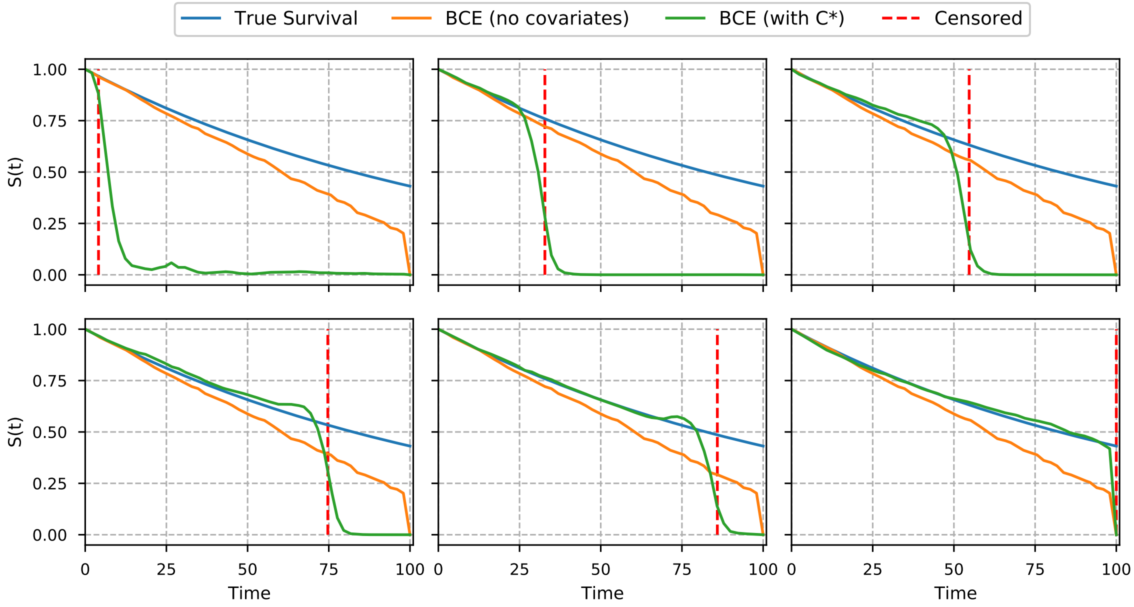

Figure 3: Survival estimates from the BCE method with and without a monotone function of the censoring time c_i^* as a covariate. The vertical dotted red line gives the censoring time c_i^*. The true survival function is plotted in blue. All event times are drawn from the same distribution.

Kaplan-Meier Brier scores and the administrative Brier scores are very close to the scores of the uncensored test set. This is expected for the IPCW Brier score in general (cf. Equation 5) and can be expected for the administrative Brier score for covariate-independent censoring (cf. Equation A.3 in Appendix A). However, for the BCE method with censoring information we see that the IPCW Kaplan-Meier Brier scores are much lower than those of the true survival function. This clearly illustrate the bias presented in Section 3.3. Namely that the IPCW Brier score will give an advantage to estimates that are biased towards zero after the administrative censoring time.

The administrative Brier score for the BCE method with censoring information is almost identical to that of the true survival function. The reason for this is that the survival estimates of the BCE method are close to the true survival function for $t \leq c_i^*$, as we saw in Figure 3. Per construction, the administrative Brier score can only penalize predictions for $t \leq c_i^*$ and disregards predictions for $t > c_i^*$.

The IPCW one-point distribution Brier score (6) yields results quite far from the other metrics. As stated in Section 3.3, this score is not sensible, but is interesting to study because the IPCW Brier score approaches this score as the estimates of the censoring distribution approach one-point distributions. So here we see that for very precise estimates of the censoring times, the IPCW Brier scores for both of the BCE methods are lower than that of the true survival function.

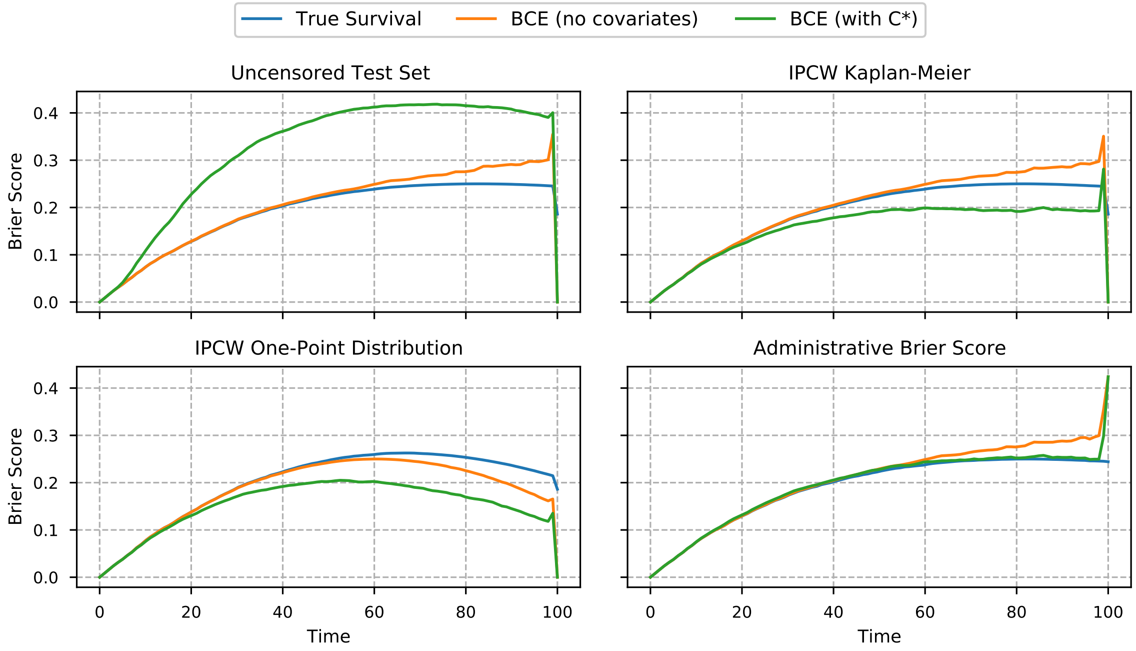

Figure 4: Brier scores from simulations with administrative censoring. The top left plot uses an uncensored test set, while the three other use the right-censored test set. The orange lines represent a BCE method without any information about the censoring times, while the green lines represent a BCE method with sufficient covariate information to identify the censoring times. The blue lines are the Brier scores of the true survival functions.

The simulations were made to illustrate a potential problem with the IPCW Brier score, and in reality it is unlikely to have a covariate that is a monotonic function of the censoring time (unless the date of entry to the experiment is included). It is however possible that some covariates related to the date of entry could be included (such as promotion campaigns or other calendar-time relates events) making a possible subset of the censoring times identifiable.

5. Simulations

The simulations in Section 4.1 illustrated the potential issues of using the IPCW Brier score on a data set with administrative censoring. However, the simulations considered a covariate independent event-time distribution and a very simple relationship between the covariates and the administrative censoring times. To investigate the issues further, we conduct a simulation study to see how they are affected by the complexity of the event-time and censoring distributions.

In the following study, we draw event times from the discrete time grid $\{0.1, 0.2, \ldots, 100\}$. The reason for considering discrete time rather than continuous time is solely that it makes it much simpler to create complicated survival distributions. Note however that the grid considered is very fine, so the simulations approximate continuous event times very well.

We will use the framework presented by Kvamme and Borgan (2019) to create the simulated data sets. This means that we draw event times by sequentially sampling from discrete-time hazards on the time grid $\{0.1, 0.2, \ldots, 100\}$. The hazards are specified through the logit-hazard function $g(t \,|\, \mathbf{x}) \in \mathbb{R}$. Note that $g(t \,|\, \mathbf{x})$ is just the notation used by Kvamme and Borgan (2019), and is not related to $G(t \,|\, \mathbf{x}_i)$. The discrete-time hazards are given by the sigmoid

$$h(t \,|\, \mathbf{x}) = \frac{1}{1 + \exp\left[-g(t \,|\, \mathbf{x})\right]},$$

and the logit hazards are defined as

$$g(t \,|\, \mathbf{x}) = \alpha_1(\mathbf{x})\, g_{\text{sin}}(t \,|\, \mathbf{x}) + \alpha_2(\mathbf{x})\, g_{\text{con}}(t \,|\, \mathbf{x}) + \alpha_3(\mathbf{x})\, g_{\text{acc}}(t \,|\, \mathbf{x}), \tag{16}$$

where

$$\begin{aligned} g_{\text{sin}}(t \,|\, \mathbf{x}) &= \gamma_1(\mathbf{x}) \sin\big(\gamma_2(\mathbf{x})[t + \gamma_3(\mathbf{x})]\big) + \gamma_4(\mathbf{x}), \\ g_{\text{con}}(t \,|\, \mathbf{x}) &= \gamma_5(\mathbf{x}), \\ g_{\text{acc}}(t \,|\, \mathbf{x}) &= \gamma_6(\mathbf{x}) \cdot t - 10, \end{aligned} \tag{17}$$

and

$$\alpha_i(\mathbf{x}) = \frac{\exp(\gamma_{i+6}(\mathbf{x}))}{\sum_{j=1}^{3} \exp(\gamma_{j+6}(\mathbf{x}))}, \quad \text{for } i = 1, 2, 3.$$

Each of the three functions $g_{\text{sin}}(t \,|\, \mathbf{x})$, $g_{\text{con}}(t \,|\, \mathbf{x})$, and $g_{\text{acc}}(t \,|\, \mathbf{x})$ are constructed to give a specific contribution to the hazards: $g_{\text{con}}(t \,|\, \mathbf{x})$ gives a constant hazard for a set of covariates, $g_{\text{acc}}(t \,|\, \mathbf{x})$ allows for a hazard that increases with time, and $g_{\text{sin}}(t \,|\, \mathbf{x})$ enables periodic patterns in the hazards. With this combination, we are able to represent a variety of event-time distributions. The definitions of the functions $\gamma_k(\mathbf{x})$ are given by Kvamme and Borgan (2019, Appendix A.1) who also explain the scheme used to draw the covariates. With τ_j denoting the j'th time point in $\{0.1, 0.2, \ldots, 100\}$, the survival function is given by

$$S(\tau_j \,|\, \mathbf{x}_i) = \prod_{k=1}^{j} \left[1 - h(\tau_k \,|\, \mathbf{x}_i)\right]. \tag{18}$$

To incorporate administrative censoring in the simulations, we consider a monotonically decreasing function of time $Q(t \,|\, \mathbf{x}_i)$, and let the censoring time c_i^* be defined by a threshold ψ such that $Q(c_i^* \,|\, \mathbf{x}_i) = \psi$. This gives the one-point survival function of the censoring distribution

$$G(t \,|\, \mathbf{x}_i) = \mathbb{1}\{Q(t \,|\, \mathbf{x}_i) > \psi\}$$

Hence, the censoring is deterministic, while the complexity of $Q(t \,|\, \mathbf{x}_i)$ controls the complexity of the relationship between the covariates $\mathbf{x}_i$ and the censoring time c_i^*. We will let $Q(t \,|\, \mathbf{x}_i)$ have the same functional form as a survival function, meaning it is defined in the same manner as (18), but with its own set of covariates independent of the covariates of the event-time distribution. In all the experiments we set $\psi = 0.2$.

In the experiments, we compare the BCE method with the Logistic-Hazard method. We use the Logistic-Hazard because of its similarity to the BCE method, in that it also minimizes the binary cross-entropy, but use the discrete hazards instead of the survival estimates. We use the implementation of the Logistic-Hazard by Kvamme and Borgan (2019), but the method has been described by multiple authors (Brown, 1975; Biganzoli et al., 1998; Gensheimer and Narasimhan, 2019).

5.1 Complicated Censoring Distribution

In the first study, we draw event times using (16) with two covariates per $\gamma_k(\mathbf{x})$. The function $Q(t \,|\, \mathbf{x})$ is also defined using (16) and (18), but with an independent set of covariates, ensuring that the event times are independent of the censoring times. Note that although the covariates contain enough information to identify the censoring times, the complexity of the censoring distribution makes this a somewhat hard task.

We fit models using all the covariates from both the event-time distribution and the censoring distribution, giving a total of 36 covariates. We draw 10,000 individuals for training and testing, and 4,000 for a validation set used for early stopping of our training procedure. The networks are ReLU-nets with 4 layers and 32 nodes in each layer. Batch normalization and dropout with a rate of 0.1 are applied between each layer. The BCE and Logistic-Hazard both give estimates for 50 equidistant times points between 0 and 100, but we perform constant density interpolation (linear interpolation of survival estimates, see Kvamme and Borgan, 2019) to obtain predictions for all 1,000 time points.

In Figure 5, we have plotted the various Brier scores for the survival estimates of the Logistic-Hazard method (orange), the BCE method (green), and estimates equal to the true survival function (blue). Each panel shows a different way of computing the Brier score for these survival estimates. In the top three panels, we have the uncensored Brier score (2) of the full uncensored test set, followed by the administrative Brier score (9), and the IPCW Brier score (11) with Kaplan-Meier weights. In the three lower panels of the figure, we have the IPCW Brier score (6) normalized by (12) with the one-point censoring distribution, and two IPCW Brier scores using (11) with different weights (see below). Only the uncensored Brier score uses the full uncensored test set, while the other metrics use the censored test set. The latter two IPCW scores in Figure 5 are calculated with censoring distributions estimated with Logistic-Hazard and max weights of 100 and 1000. Recall from Section 3.5 that when we estimate the censoring distribution with other methods than Kaplan-Meier, the weights can become very large, resulting in unstable results. Hence we set a max weight, which is given above the respective plots.

From Figure 5, we see that both methods perform rather poorly compared to the true survival function, and the Logistic-Hazard obtains smaller scores than the BCE for all scores except for the IPCW with the one-point censoring distribution. The problems with IPCW do not appear here because the functional form of the censoring distribution is quite complicated. Hence, the BCE method is not able to identify the censoring times to the extent that it can take advantage of the bias of the IPCW scores.

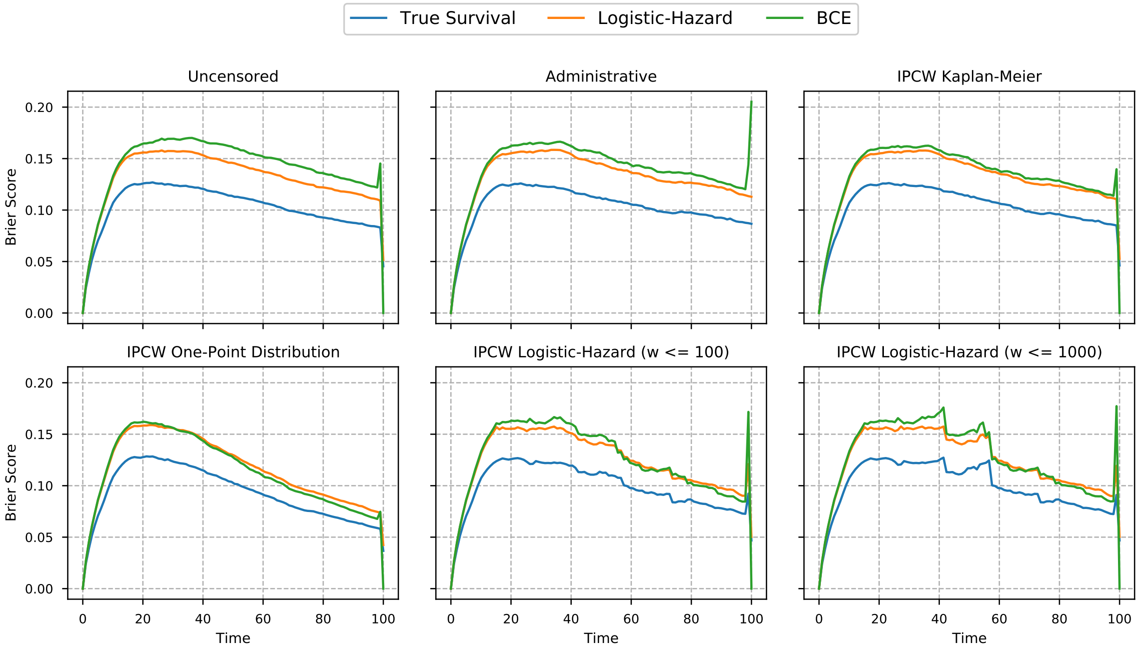

Figure 5: Brier scores from simulations with complicated administrative censoring.

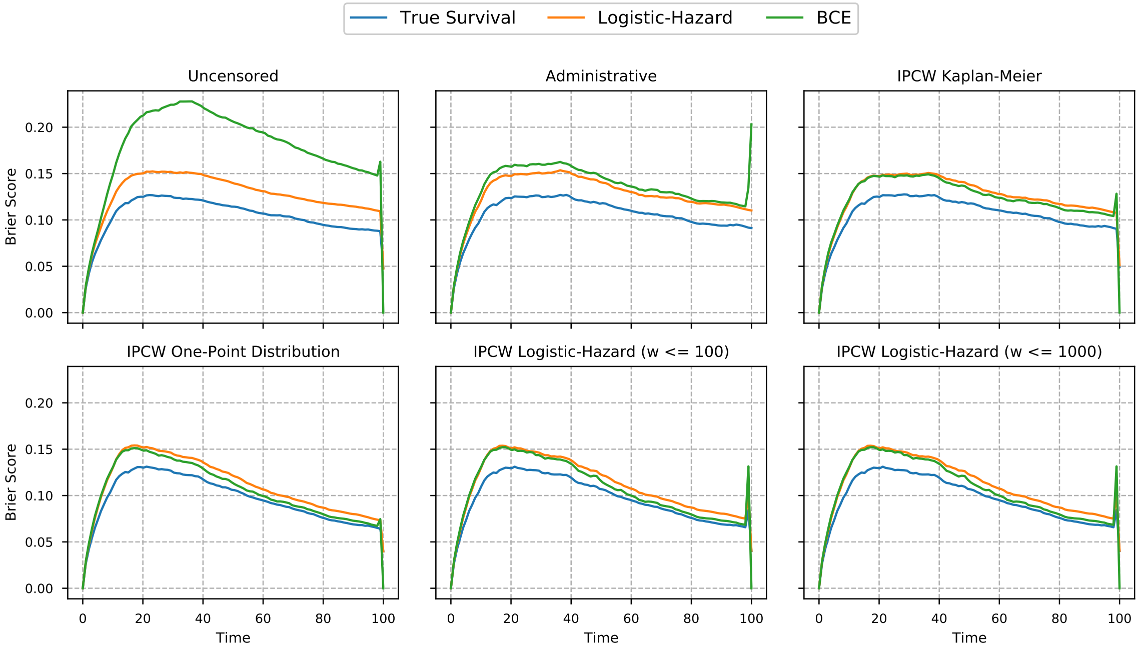

Figure 6: Brier scores from simulations with simple administrative censoring.

5.2 Simple Censoring Distribution

In the second simulations study, we let $Q(t \mid \mathbf{x})$ be defined by (18) with $g(t \mid \mathbf{x}) = \gamma_5(\mathbf{x})$ from (17) and let $\gamma_5(\mathbf{x})$ be a function of 5 covariates. This gives a fairly simple censoring distribution, meaning that the censoring times c_i^* should be possible to predict quite accurately. We repeat the simulations with the event-time distribution unchanged, and the results are displayed in Figure 6. First, we note that the performance of the BCE method on the uncensored test set is worse than before. This is expected because it is now easier for the BCE method to identify the censoring time c_i^*, so the survival estimates are likely more biased. As a consequence of this, we see that the IPCW Brier scores of the BCE method are smaller than those of the Logistic-Hazard. The difference is quite small though. The administrative Brier score is the only metric that obtains the same ordering of the predictions as in the uncensored test set.

In these simulations we have shown that the issues of the IPCW Brier score depends on how easily the administrative censoring times can be predicted. For a complicated relationship between covariates and the censoring distribution, the IPCW Brier score will likely work fine, but for a simpler relationship the IPCW might result in a faulty ranking of model predictions.

6. KKBox Churn Prediction

Finally, we revisit the KKBox data set discussed in Section 2. We fit the BCE method from Section 4, corresponding to multiple binary classifiers, and the Logistic-Hazard method. The training set is of size 100,000 and we use a validation set of size 20,000 for early stopping. We use 6-layer ReLU networks with 128 nodes in each layer, and with batch normalization and a dropout rate of 0.1 between each layer. Entity embeddings are used for the categorical covariates (Guo and Berkhahn, 2016) with embedding sizes that are the square root of the number of categories. The outputs of the networks are of size 150 representing an equidistant grid of the full time-scale of the training set. Constant density interpolation (Kvamme and Borgan, 2019) is applied to the survival estimates to obtain predictions for all the time-points in the test set. We use the AdamWR (Loshchilov and Hutter, 2019) optimizer with a cycle length of 1 epoch, but we double the cycle length and multiply the learning rate by 0.8 after each cycle. Also, we do not include weight decay. The data set is, essentially, the data set presented by Kvamme et al. (2019), but including all administrative censoring times and an extra categorical covariate stating the payment method. The code for obtaining the data set is available at `github.com/havakv/pycox`.

The empirical marginal survival distribution of the administrative censoring times is quite linear, suggesting that the recruitment rate is constant over time. At times 200, 400 and 600 the proportions of administratively censored individuals in the test set are 0.25, 0.43 and 0.67, and the proportions of individuals still at risk are 0.40, 0.26 and 0.10.

The censoring distribution is estimated in two ways. The first is with the Kaplan-Meier estimator and the second is with a Logistic-Hazard with the same hyperparameters and covariates as the Logistic-Hazard used to estimate the churn distribution. The Brier scores are computed on a test set of size 100,000 and displayed in Figure 7. The figure contains the two plots from Figure 2, the administrative Brier scores (9), and three other IPCW scores (11) with different max weights. Note that a maximum weight of 1 is proportional

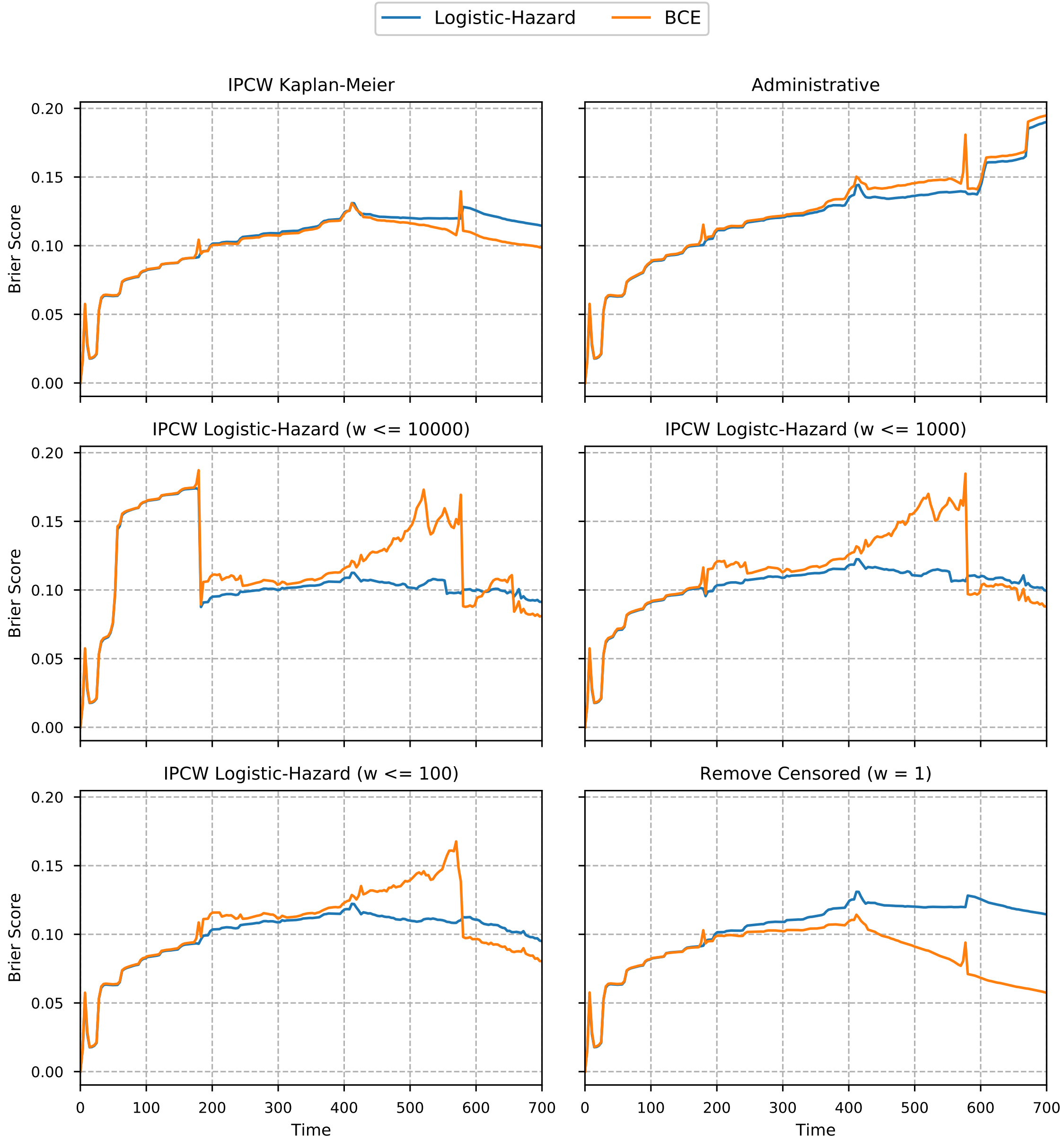

Figure 7: Brier scores on the KKBox data set.

to the BS_C weighted by the one-point censoring distribution (6) where the set of censored individuals $\{i : T_i \leq t, D_i = 0\}$ are removed. This means the score removes censored individual in the same manner as the BCE method does.

As we recall from Section 2, the BCE method obtains lower scores than the Logistic-Hazard method for the IPCW Brier scores with Kaplan-Meier estimates (top left Figure 7). We now see from the administrative scores (top right) that this is probably a wrong conclusion, as the BCE method has higher administrative Brier scores than the Logistic-Hazard. We also see that the IPCW scores obtained with Logistic-Hazard are very dependent on the maximum weight allowed in the scores. For higher allowed weights, we obtain higher scores with BCE than Logistic-Hazard, but note that for the highest times this is not the case.

In summary, we have found that the various Brier scores do not agree on the ranking of the predictions. The fact that the administrative Brier score is partly in disagreement with the other scores, means that the covariates might contain information concerning the administrative censoring times. This means that the IPCW Brier scores can give misleading results, and the administrative Brier score is likely a safer choice of metric. A reasonable next step would be to analyze the covariates in an attempt to identify which covariates contain censoring information, and see if they can be removed without an extensive impact on the predictive performance. We do, however, consider this outside the scope of this paper.

7. Discussion

In this paper, we have addressed potential issues of the inverse probability of censoring weighted (IPCW) Brier score, in particular for administrative censoring. If the covariates have sufficient information to predict the administrative censoring times, the IPCW Brier score will not be minimized by the true survival functions, but instead by a function that falls to zero after the censoring time. We have also shown that a binary classifier that disregards censored individuals will approach this minimizer. As a consequence, the binary classifier might actually get lower IPCW Brier scores than the true survival function. Due to this bias, we argue that the IPCW Brier score needs to be applied with care.

The IPCW Brier score can be substantially affected by the estimated censoring distribution. If this is the case, the validity of the scores can be questioned. In regards to the issues with the IPCW scheme, we propose the *administrative Brier score* that works for administrative right-censored event times and does not require estimation of the censoring distribution. The score requires the administrative censoring times to be known for every individual in the data set (also when the censoring time is larger than the event time), and there can be no other form of censoring present. These requirements mean that the score might not be applicable to most medical data sets, and is more relevant for industrial applications.

We simulate examples where the IPCW score fails to identify the best survival estimates, but the administrative Brier score still provides reasonable scores. We also investigate a real-world churn data set with administrative censoring and find that it exhibits some of the same behavior as our simulations. This shows that the proposed administrative Brier score can be a useful evaluation metric.

We note that the issues of the IPCW scores are addressed by the administrative Brier score $\mathrm{BS}_\mathrm{A}(t, \pi)$ by removing the problematic subset $\{i : c_i^* < t\}$. This does, however, not penalized poor predictions on this subset. So while the administrative scores ensure that the true survival functions are optimal (in expectation), this solution is not longer unique and any functions equal to the true survival functions up to the censoring times are considered optimal. We would, therefore, argue that the administrative Brier score is most useful as an addition to the IPCW Brier score, as we do not consider it to improve on every aspect of the IPCW Brier score. Most notably, if the administrative and the IPCW scores disagree on the ranking of a set of predictions, this is likely a sign that the data set contains some covariates closely connected to the censoring times.

One should also note that the issues of the IPCW Brier score can be controlled by considering a test set that occurs later in (calendar) time than the training set. By ensuring that the training period and test period are disjoint, it is highly unlikely that a model will be able to predict the administrative censoring times. Whether or not such a split is practically possible depends on the data set.

In this paper, we have only investigated the IPCW Brier score, but note that there are multiple other IPCW scores that might suffer from the same drawbacks as the Brier score. For example the IPCW Binomial log-likelihood (Graf et al., 1999) and the IPCW concordance index (Uno et al., 2011; Gerds et al., 2013). Preliminary investigations of the IPCW Binomial log-likelihood (Graf et al., 1999) suggest that it has the same issues as the IPCW Brier score while an administrative version of the Binomial log-likelihood behaves in the same manner as the administrative Brier score.

The issues discussed in this paper are mostly relevant for machine learning methods, such as neural networks, as the issues are only notable for quite precise estimates of the censoring times C_i^*. Hence, we are unlikely to encounter such issues with classical statistical models.

Large parts of the machine learning literature rely on empirical evaluation of predictive methodology. We, therefore, believe that more research on the evaluation metrics for right-censored survival data is needed.

Acknowledgments

This work was supported by The Norwegian Research Council 237718 through the Big Insight Center for research-driven innovation.

Appendix A. The Administrative Brier Score

The administrative Brier score is given by

$$\mathrm{BS}_{\mathrm{A}}(t, \pi) = \frac{1}{\tilde{n}_{\mathrm{A}}(t)} \sum_{i=1}^{n} [\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{c_i^* \geq t\},$$

where

$$\tilde{n}_{\mathrm{A}}(t) = \sum_{i=1}^{n} \mathbb{1}\{c_i^* \geq t\},$$

cf. formula (9). In Section 3.4 we derive an expression for the conditional expectation of the administrative Brier score given that the administrative censoring times C_i^* take the values c_i^*. We will here consider its unconditional expected value, where we also average over the distribution of the C_i^*'s.

Under some fairly mild regularity conditions, one may approximate the expected value of a ratio of two means by the ratio of the expected values of the means. Thus we have that

$$\mathbb{E}[\mathrm{BS_A}(t,\pi)] \approx \frac{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n [\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{C_i^* \geq t\}\right)}{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{C_i^* \geq t\}\right)}. \tag{A.1}$$

Now we find that

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{C_i^* \geq t\}\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{P}(C_i^* \geq t \,|\, \mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^n G(t- \,|\, \mathbf{x}_i).$$

Further, if we consider the $\pi(t \,|\, \mathbf{x}_i)$'s to be known functions, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n [\mathbb{1}\{T_i^* > t\} - \pi(t \,|\, \mathbf{x}_i)]^2 \mathbb{1}\{C_i^* \geq t\}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\left[\mathbb{1}\{T_i^* > t\} - 2\pi(t \,|\, \mathbf{x}_i)\mathbb{1}\{T_i^* > t\} + \pi(t \,|\, \mathbf{x}_i)^2\right] \mathbb{1}\{C_i^* \geq t\}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left[\mathrm{P}(T_i^* > t \,|\, \mathbf{x}_i) - 2\pi(t \,|\, \mathbf{x}_i)\mathrm{P}(T_i^* > t \,|\, \mathbf{x}_i) + \pi(t \,|\, \mathbf{x}_i)^2\right] \mathrm{P}(C_i^* \geq t \,|\, \mathbf{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^n \left[S(t \,|\, \mathbf{x}_i) - 2\pi(t \,|\, \mathbf{x}_i)S(t \,|\, \mathbf{x}_i) + \pi(t \,|\, \mathbf{x}_i)^2\right] G(t- \,|\, \mathbf{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^n \left\{\pi(t \,|\, \mathbf{x}_i)^2[1 - S(t \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 S(t \,|\, \mathbf{x}_i)\right\} G(t- \,|\, \mathbf{x}_i).
\end{aligned}$$

By similar calculations, it is straightforward to show that we obtain the same expected value if we instead use the predictors $\pi_{c^*}(t \,|\, \mathbf{x}_i)$ given by (8).

If we insert the above expressions in (A.1), we obtain the following approximation for the expected value of the administrative Brier score

$$\mathbb{E}[\mathrm{BS_A}(t,\pi)] \approx \frac{\frac{1}{n}\sum_{i=1}^n \left\{\pi(t \,|\, \mathbf{x}_i)^2[1 - S(t \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 S(t \,|\, \mathbf{x}_i)\right\} G(t- \,|\, \mathbf{x}_i)}{\frac{1}{n}\sum_{i=1}^n G(t- \,|\, \mathbf{x}_i)}. \tag{A.2}$$

This result holds for the predictors $\pi(t \,|\, \mathbf{x}_i)$ as well as the predictors $\pi_{c^*}(t \,|\, \mathbf{x}_i)$. Thus the (approximate) expected value of the administrative Brier score does not change if we set the survival predictions to zero after the administrative censoring times. This means that, contrary to the IPCW Brier score, there is no advantage in having survival predictions that use (partial) information on the administrative censoring times.

For the special case where the censoring distributions are the same for all i, i.e., $G(t \,|\, \mathbf{x}_i) = G(t)$, the approximation may be written as

$$\mathbb{E}[\mathrm{BS_A}(t,\pi)] \approx \frac{1}{n}\sum_{i=1}^n \left\{\pi(t \,|\, \mathbf{x}_i)^2[1 - S(t \,|\, \mathbf{x}_i)] + [1 - \pi(t \,|\, \mathbf{x}_i)]^2 S(t \,|\, \mathbf{x}_i)\right\}, \tag{A.3}$$

which is the same as the expected value of the Brier score when there is no censoring and the expected value of the IPCW Brier score; cf. (3) and (5)

Appendix B. The BCE Survival Estimates

To better understand the survival estimates of the binary classifiers in Section 4, we investigate the minimizers of the expected loss. Again, we stress that the BCE method corresponds to a set of binary classifiers in the manner given by the loss (13) and (14).

The expected loss of the binary classifier (13) is

$$\begin{aligned}
\mathbb{E}\left[\text{loss}_{\text{BCE}}(t, \pi)\right] &= -\sum_{i=1}^{n} \Big(\mathrm{P}(T_i > t) \log\left[\pi(t \,|\, \mathbf{x}_i)\right] + \mathrm{P}(T_i \leq t,\, D_i = 1) \log\left[1 - \pi(t \,|\, \mathbf{x}_i)\right] \Big) \\
&= \;\; -\sum_{i=1}^{n} \Big(\mathrm{P}(T_i^* > t,\, C_i^* > t) \log\left[\pi(t \,|\, \mathbf{x}_i)\right] + \mathrm{P}(T_i^* \leq t,\, C_i^* \geq T_i^*) \log\left[1 - \pi(t \,|\, \mathbf{x}_i)\right] \Big) \\
&= \;\; -\sum_{i=1}^{n} \Big(S(t \,|\, \mathbf{x}_i) G(t \,|\, \mathbf{x}_i) \log\left[\pi(t \,|\, \mathbf{x}_i)\right] + \left[\int_0^t G(u- \,|\, \mathbf{x}_i) f(u \,|\, \mathbf{x}_i) du\right] \log\left[1 - \pi(t \,|\, \mathbf{x}_i)\right] \Big).
\end{aligned}$$

The minimizers of this expectation with respect to $\pi(t \,|\, \mathbf{x}_i)$ can be found by equating the partial derivatives with zero,

$$\frac{\partial \mathbb{E}\left[\text{loss}_{\text{BCE}}(t, \pi)\right]}{\partial \pi(t \,|\, \mathbf{x}_i)} = -\frac{S(t \,|\, \mathbf{x}_i)\, G(t \,|\, \mathbf{x}_i)}{\pi(t \,|\, \mathbf{x}_i)} + \frac{\int_0^t G(u- \,|\, \mathbf{x}_i) f(u \,|\, \mathbf{x}_i) du}{1 - \pi(t \,|\, \mathbf{x}_i)} = 0,$$

This gives the minimizers

$$\begin{aligned}
\pi(t \,|\, \mathbf{x}_i) &= \frac{S(t \,|\, \mathbf{x}_i) G(t \,|\, \mathbf{x}_i)}{S(t \,|\, \mathbf{x}_i) G(t \,|\, \mathbf{x}_i) + \int_0^t G(u- \,|\, \mathbf{x}_i) f(u \,|\, \mathbf{x}_i) du} \\
&\leq \frac{S(t \,|\, \mathbf{x}_i) G(t \,|\, \mathbf{x}_i)}{S(t \,|\, \mathbf{x}_i) G(t \,|\, \mathbf{x}_i) + \int_0^t G(t \,|\, \mathbf{x}_i) f(u \,|\, \mathbf{x}_i) du} \\
&\leq \frac{S(t \,|\, \mathbf{x}_i)}{S(t \,|\, \mathbf{x}_i) + \int_0^t f(u \,|\, \mathbf{x}_i) du} \\
&= S(t \,|\, \mathbf{x}_i).
\end{aligned} \tag{B.4}$$

We see that the $\pi(t \,|\, \mathbf{x}_i)$'s are underestimating the true survival as long as there is censoring present. This is expected as the binary classifiers remove censored individuals, decreasing the proportion of survived individuals in the data set.

The situation where the censoring times c_i^* can be identified by the covariates, corresponds to using the one-point censoring distribution $\mathbb{1}\{c_i^* > t\}$ for $G(t \,|\, \mathbf{x}_i)$ in (B.4). The minimizer can then be written as

$$\pi(t \,|\, \mathbf{x}_i) = \frac{S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}}{S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\} + \int_0^t \mathbb{1}\{c_i^* \geq u\} f(u \,|\, \mathbf{x}_i) du}.$$

If $c_i^* > t$, we have

$$\pi(t \,|\, \mathbf{x}_i) = \frac{S(t \,|\, \mathbf{x}_i)}{S(t \,|\, \mathbf{x}_i) + \int_0^t f(u \,|\, \mathbf{x}_i) du} = S(t \,|\, \mathbf{x}_i),$$

and if $c_i^* \leq t$, we have $\pi(t \,|\, \mathbf{x}_i) = 0$. The minimizer can, therefore, be written as

$$\pi(t \,|\, \mathbf{x}_i) = S(t \,|\, \mathbf{x}_i) \mathbb{1}\{c_i^* > t\}.$$

We recognize this as the minimizer of the IPCW Brier score with administrative censoring from Section 3.3. So if there is sufficient information in the covariates to identify the administrative censoring times c_i^*, the binary classifiers will approach the minimizers of the IPCW Brier score.

References

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

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

Charles C. Brown. On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics*, 31(4):863–872, 1975.

Travers Ching, Xun Zhu, and Lana X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4):e1006076, 2018.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48 (6):1029–1040, 2006.

Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

Frank E. Harrell Jr, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247 (18):2543–2546, 1982.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*, 2019.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*, 2017.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E. Lewis, Congzheng Song, David A. Gutman, Sameer H. Halani, Jose Enrique Velazquez Vega, Daniel J. Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(11707), 2017.