Journal of Machine Learning Research 21 (2020) 1-63 Submitted 11/18; Published 6/20

Effective Ways to Build and Evaluate Individual Survival Distributions

Humza Haider HSHAIDER@UALBERTA.CA
Bret Hoehn BHOEHN@UALBERTA.CA
Sarah Davis SDAVIS1@UALBERTA.CA
Russell Greiner* RGREINER@UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8

Editor: Robert McCulloch

Abstract

An accurate model of a patient's individual survival distribution can help determine the appropriate treatment for terminal patients. Unfortunately, risk scores (for example from Cox Proportional Hazard models) do not provide survival *probabilities*, single-time probability models (for instance the Gail model, predicting 5 year probability) only provide for a single time point, and standard Kaplan-Meier survival curves provide only *population averages* for a large class of patients, meaning they are not specific to individual patients. This motivates an alternative class of tools that can learn a model that provides an individual survival *distribution* for each subject, which gives survival probabilities across all times, such as extensions to the Cox model, Accelerated Failure Time, an extension to Random Survival Forests, and Multi-Task Logistic Regression. This paper first motivates such "individual survival distribution" (ISD) models, and explains how they differ from standard models. It then discusses ways to evaluate such models — namely Concordance, 1-Calibration, Integrated Brier score, and versions of L1-loss — then motivates and defines a novel approach, "D-Calibration", which determines whether a model's probability estimates are meaningful. We also discuss how these measures differ, and use them to evaluate several ISD prediction tools over a range of survival data sets. We also provide a code base for all of these survival models and evaluation measures, at `https://github.com/haiderstats/ISDEvaluation`.

Keywords: Survival analysis, risk model, patient-specific survival prediction, calibration, discrimination

*Also: Fellow, Amii (Alberta Machine Intelligence Institute)

1. Introduction

When diagnosed with a terminal disease, many patients ask about their prognosis (Gwilliam et al., 2012): "How long will I live?", or "What is the chance that I will live for 1 year... and the chance for 5 years?". Here it would be useful to have a meaningful "survival distribution" $S(\, t \,|\, \vec{x}\,)$ that provides, for each time $t \geq 0$, the probability that this specific patient $\vec{x}$ will survive at least an additional t months. Unfortunately, many of the standard survival analysis tools cannot accurately answer such questions: (1) risk scores (*e.g.*, Cox proportional hazard (Cox, 1972)) provide only *relative* survival measures, but not the calibrated probabilities desired; (2) single-time probability models (*e.g.*, the Gail model (Costantino et al., 1999)) provide a probability value but *only for a single time point*; and (3) class-based survival curves (like Kaplan-Meier, KM (Kaplan and Meier, 1958)) are *not specific to the patient*, but rather an entire population.

To explain the last point, Figure 1[left] shows the KM curve for patients with stage-4 stomach cancer. Here, we can read off the claim that 50% of the patients will survive 11 months, and 95% will survive at least 2 months.[1] While these estimates do apply to the population, *on average*, they are not designed to be "accurate" for an individual patient since these estimates do not include patient-specific information such as age, treatments administered, or general health conditions. It would be better to directly, and correctly, incorporate these important factors $\vec{x}$ explicitly in the prognostic models.

This heterogeneity of patients, coupled with the need to provide probabilistic estimates at several time points, has motivated the creation of several *individual survival time distribution* (ISD) tools, each of which can use this wealth of healthcare information from earlier patients, to learn a more accurate prognostic model, which can then predict the ISD of a novel patient based on all available patient-specific attributes. Many (*e.g.*, Henderson and Keiding, 2005; Hollnagel, 1999) have argued that such survival *distributions* are better than simple point estimates — *e.g.*, a risk score or the probability of surviving to a single specified time — for explaining patient survival, as the distribution provides more information.

This paper considers several ISD models: the Kalbfleisch-Prentice extensions of the Cox (COX-KP) (Kalbfleisch and Prentice, 2002) and the elastic net Cox (COXEN-KP) (Yang and Zou, 2013) models, the Accelerated Failure Time (AFT) model
(Kalbfleisch and Prentice, 2002), the Random Survival Forest model with Kaplan-Meier extensions (RSF-KM) (Ishwaran et al., 2008), and the Multi-task Logistic Regression (MTLR) model (Yu et al., 2011). Figure 1[middle, right] shows survival curves (generated by MTLR) for two of these stage-4 stomach cancer patients, which incorporates other information about these individuals such as their age, gender, blood work, etc. We see that these prognoses are very different; in particular, MTLR predicts that [middle] Patient #1's median survival time is 20.2 months, while [right] Patient #2's is only 2.6 months. The blue vertical dashed-lines show the actual times of death; we see that each of these patients passed away very close to MTLR's predictions of their respective median survival times.

1. In general, a survival curve is a plot where each $[\, x,\ y\,]$ point represents (the curve's claim that) there is a y% chance of surviving at least x time. Hence, in Figure 1[left], the [11 months, 50%] point means this curve predicts a 50% chance of living at least 11 months (and hence a $100-50 = 50$% chance of dying within the first 11 months). The [2 months, 95%] point means a 95% chance of surviving at least 2 months, and the [51 months, 5%] point means a 5% chance of surviving at least 51 months.

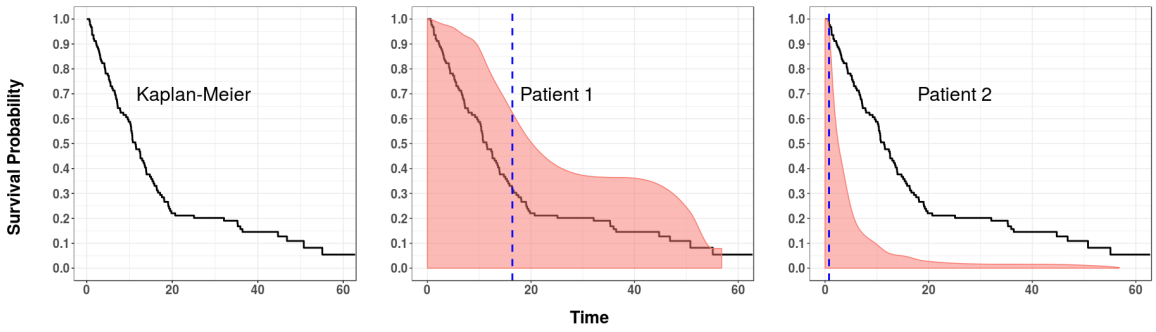
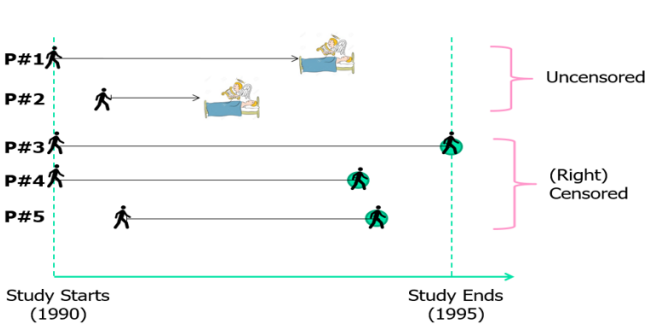


Figure 1: [left] Kaplan-Meier curve, based on 128 patients with stage-4 stomach cancer. [middle, right] Two personalized survival curves, for two patients (#1 and #2) with stage-4 stomach cancer. The blue dashed lines indicate the true time of death.

One could then use such curves to make decisions about the individual patient. Of course, these decisions will only be helpful if the model is providing accurate information — *i.e.*, only if it is appropriate to tell a patient that s/he has a 50% chance of living more than the median survival time of this predicted curve, and a 25% chance of living more than the time associated with the 25% on the curve, etc.

We focus on ways to *learn* such models from a "survival data set" (see below), describing earlier individuals. Survival prediction is similar to regression as both involve learning a model that regresses the covariates of an individual to estimate the value of a dependent real-valued response variable — here, that variable is "time to event" (where the standard event is "death"). But survival prediction differs from the standard regression task as its response variable is not fully observed in all training instances — this task allows many of the instances to be "right censored", in that we only see a *lower bound* of the response value. This might happen if a subject was alive when the study ended, meaning we only know that s/he lived *at least* (say) 5 years after the starting time (see P#3 in Figure 2), but do not know whether she actually lived 5 years and a day, or 20 years. This also happens if a subject drops out of a study, after say 4 years, and is then lost to follow-up (*e.g.*, P#4), etc. Moreover, one cannot simply ignore such instances as it is common for many (or often, *most*) of the training instances to be right-censored; see second row of Table 5.

Such "partial label information" is problematic for standard regression techniques, which assume the label is completely specified for each training instance. Fortunately, there are survival prediction algorithms that can learn an effective model, from a cohort that includes such censored data. Each such "survival data set" contains descriptions of a set of instances (*e.g.*, patients), as well as two "labels" for each: one is the time, corresponding to the *time from diagnosis to a final date* (either death, or time of last follow-up) and the other is the *status* bit, which indicates whether the patient was alive at that final date. Section 2 summarizes several popular models for dealing with such survival data. We also note that there are many survival *analysis* tools that instead use a survival data set to identify



Site	Stage	Age	Sex	..	HGB	Time	Censored
Head	3	60	m	..	71	10.1	1
Head	3	53	m	..	93	7.9	1
Neck	1	70	f	..	85	6.1	0
Head	2	57	m	..	55	5.5	0
CR	4	71	f	..	92	20.2	0
:	:	:	:	..	:	:	:

Figure 2: Example to illustrate censoring: Here, a 5-year study began in 1990. Over this time, P#1 and P#2 died, but P#3 survived until the end, and P#4 and P#5 were lost to follow-up.

biomarkers — each a feature that, by itself, is related to an individual's survival. We will later use this facility to reduce the number of features considered by our survival prediction models.

This paper provides three contributions: (1) Section 2 motivates **the need for such** ISD **models** by showing how they differ from more standard survival analysis systems. (2) Section 3 then discusses several ways to evaluate such models, including standard measures (Concordance, 1-Calibration, Brier score), variants/extensions to familiar measures (*e.g.*, L1-loss), and also **a novel approach, "D-Calibration"**, which can be used to assess the quality of the individual survival curves generated by ISD models. (3) Section 4 **evaluates several** ISD **(and related) models** (standard: KM, COX-KP, AFT and more recent: RSF-KM, COXEN-KP, MTLR) on 8 diverse survival data sets, in terms of all 5 evaluation measures. We will see that MTLR does well — typically outperforming the other models in the various measures, and often showing vast improvement in terms of calibration metrics.

The appendices provide relevant auxiliary information: Appendix A describes some important nuances about survival curves. Appendix B provides further details concerning all the evaluation metrics and in particular, how each addresses censored observations. It also contains some relevant proofs about our novel D-Calibration metric. Appendix C then explains some additional aspects of the ISD models considered in this paper. Lastly, Appendix D gives the detailed results from empirical evaluation — *e.g.*, providing detailed tables corresponding to the results shown as figures in Section 4.2.

For readers who want an introduction to survival analysis and prediction, we recommend *Applied Survival Analysis* by Hosmer et al. (2011). Wang et al. (2017) also surveyed machine learning techniques and evaluation metrics for survival analysis. However, that work primarily overviewed the standard survival analysis models, then briefly discussed some of the evaluation techniques and application areas. Our work, instead, focuses on the ISD-based models — first motivating why they are relevant for survival prediction then providing empirical results showing the strengths and weaknesses of each of the models considered.

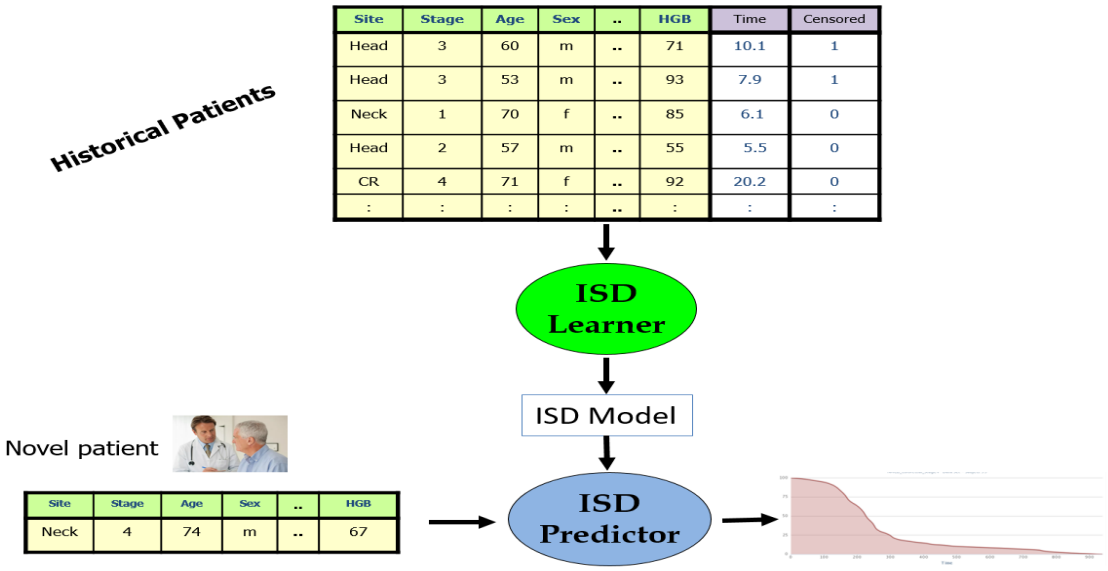


Figure 3: Machine Learning paradigm for learning, then using an ISD (Individual Survival Distribution) model.

2. Summary of Various Survival Analysis/Prediction Systems

There are many different survival analysis/prediction tools, designed to deal with various different tasks. We focus on tools that learn the model from a survival data set,

$$D \quad = \quad \{ [\vec{x}_i, t_i, \delta_i] \,|\, i = 1, \cdots, n \}_i \tag{1}$$

which provides the values for features $\vec{x}_i = [x_i^{(1)}, \cdots, x_i^{(k)}]$ for each member of a cohort of historical patients, as well as the actual time of the "event" $t_i \in \Re^{\geq 0}$ which is either death (uncensored) or the last visit (censored), and a bit $\delta \in \{0, 1\}$ that serves as the indicator for death.[2] See Figure 3, in the context of our ISD framework.

Here, we assume $\vec{x}$ is a vector of feature values describing a patient, using information that is available when that patient entered the study — *e.g.*, when the patient was first diagnosed with the disease, or started the treatment. Additionally, we assume each patient has a death time, d_i, and a censoring time, c_i, and assign $t_i := \min\{d_i, c_i\}$ and $\delta_i = \mathcal{I}[\, d_i \leq c_i \,]$ where $\mathcal{I}[\,\cdot\,]$ is the Indicator function – *i.e.*, $\delta_i := 1$ if $d_i \leq c_i$ or $\delta_i := 0$ if $d_i > c_i$. We follow the standard convention that d_i and c_i are assumed independent.

To help categorize the space of survival prediction systems, we consider 3 independent characteristics that reflect the functionality of the performance system:

2. Throughout this work we focus only on Right-Censored survival data. Additionally, we constrain our work to the standard machine-learning framework, where our predictions are based only on information available at fixed time t_0 (*e.g.*, start of treatment). While these descriptions all apply when dealing with the time to an arbitrary *event*, our descriptions will refer to "time to death".

- *[R vs P]* whether the system provides, for each patient, a risk score $r(\vec{x}) \in \Re$ versus a probabilistic value $\in [0, 1]$ (perhaps $\hat{S}(\, t \,|\, \vec{x}\,)$).
- *[1_{t^*} vs $1_\forall$ vs ∞]* whether the system returns a *single* value for each patient (associated either with a single time "1_{t^*}" or with the overall survival "$1_\forall$"), versus a range of values, one for each time. Here 1_{t^*} might refer to $\hat{S}(\, t^* \,|\, \vec{x}\,) \in [0, 1]$ for a single time t^* and $1_\forall$ if there is a single "atemporal" value (think of the standard risk score, which is not linked to a specific time), vs ∞ that refers to the set $\{\ [t,\ \hat{S}(\, t \,|\, \vec{x}\,)]\ \}_{t \geq 0}$ over all future times $t \geq 0$.
- *[i vs g]* whether the prediction is "i" meaning it is specific to a single individual patient (*i.e.*, based on a large number of features $\vec{x}$) or is "g" meaning it is general to the population. This g also applies if the model deals with a *fixed set of subpopulations* — perhaps each contains all patients with certain values of only one or two features (*e.g.*, subpopulation $p1$ is all men under 50, $p2$ are men over 50, and $p3$ and $p4$ are corresponding sets of women), or each subpopulation is a specified range of some computation (*e.g.*, $p1'$ are those with BMI<20, $p2'$ with BMI$\in$ [20, 30] and $p3'$, with BMI>30).

This section summarizes 5 (of the $2 \times 3 \times 2 = 12$) classes of survival analysis tools (see Figure 4), giving typical uses of each, then discusses how they are interrelated. Of course, there are other dimensions for structuring survival prediction models — *e.g.*, non-parametric, semi-parametric and parametric models (Wey et al., 2015). That split was more related to the assumptions of the learning process rather than the characteristics of the learned model.[3]

2.1. [R,$1_\forall$,i]: 1-value Individual Risk Models (COX)

An important class of survival analysis tools compute "risk" scores, $r(\vec{x}) \in \Re$ for each patient $\vec{x}$, with the understanding that $r(\vec{x}_a) > r(\vec{x}_b)$ corresponds to predicting that the patient described by $\vec{x}_a$ will die before the patient $\vec{x}_b$.[4] Hence, this is a *discriminative* tool for comparing pairs of patients, or perhaps for "what if" analysis of a single patient (*e.g.*, if s/he continues smoking, versus if s/he quits). These systems are typically evaluated using a discriminative measure, such as "Concordance" (discussed in Section 3.1). Notice these tools each return a single real value for each patient.

One standard generic tool here is the Cox Proportional Hazard (COX) model (Cox, 1972), which is used in a wide variety of applications. This models the hazard function[5] as

$$h_{cox}(\, t,\, \vec{x}\,) \quad = \quad \lambda_0(t)\, \exp(\vec{\beta}^T \vec{x}) \tag{2}$$

where $\vec{\beta}$ are the learned weights for the features, and $\lambda_0(t)$ is the baseline hazard function. We view this as a Risk Model by ignoring $\lambda_0(t)$ (as $\lambda_0(t)$ is the same for all patients), and

3. We will see that the RSF-KM model is non-parametric and AFT is parametric, but both are ISDs — *i.e.*, [P,∞,i].
4. To simplify notation, from here on, we will use $\vec{x}_i$ to refer to both the i^{th} patient, and the feature vector describing that patient.
5. The hazard function (also known as the failure rate, hazard rate, or force of mortality) $h(t; \vec{x}) = p(t \,|\, \vec{x})/S(\, t \,|\, \vec{x}\,)$ is essentially the chance that $\vec{x}$ will die at time t, given that s/he has lived until this time, using the survival PDF $p(t \,|\, \vec{x})$. When continuous, $h(t,\, \vec{x}) \ = \ -\frac{d}{dt} \log S(\, t \,|\, \vec{x}\,)$.

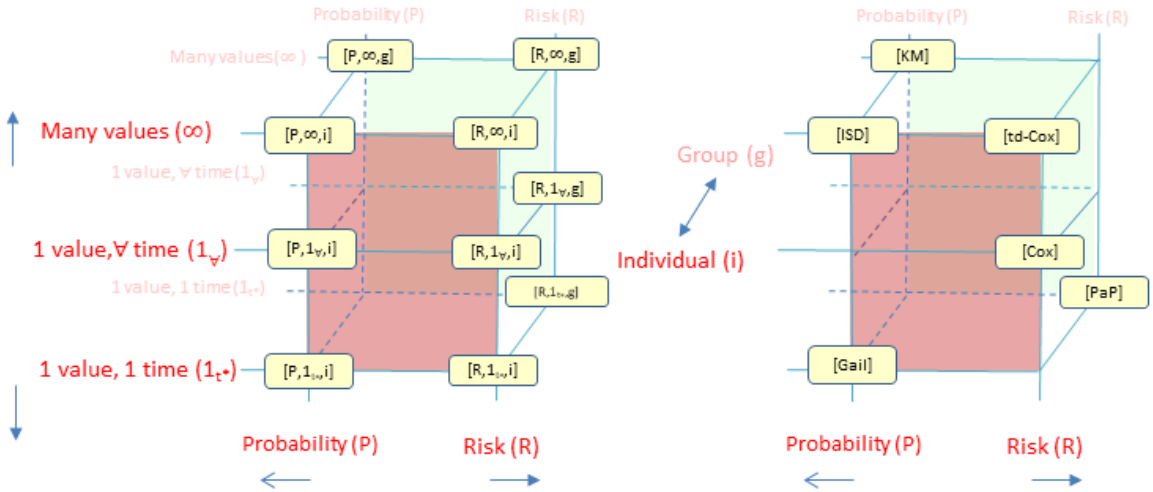


Figure 4: Dimensions for cataloging types of Survival Analysis/Prediction tools [left] — and examples of certain tools [right].

focusing on just $\exp(\vec{\beta}^T\vec{x}) \in \Re^+$. (But see the COX-KP model below, in [P,∞,i].) There are many other tools for predicting an individual's risk score, typically with respect to some disease; see for example, the Colditz and Rosner (2000) model and the myriad of others appearing on the Disease Risk Index website[6], as well as some models (like DeepSurv) that use deep nets (Katzman et al., 2018). For all of these models, the value returned is atemporal — *i.e.*, it does not depend on a specific time. There are also tools that produce [R,∞,i] models, which return a risk score associated across all time points; see Section 3.1.

2.2. [R,1_{t^*},g]: Single-time Group Risk Predictors: Prognostic Scales (PPI, PaP)

Another class of risk predictions explicitly focus on a single time, leading to prognostic scales, some of which are computed using Likert scales (Rogers et al., 2001). For example, the Palliative Prognostic Index (PPI; Morita et al., 1999) computes a risk score for each terminally ill patient, which is then used to assign that patient into one of three groups. It then uses statistics about each group to predict that patients in one group will do better at this specific time (here, 3 weeks), than those in another group. Similarly, the Palliative Prognostic Score (PaP; Pirovano et al., 1999) uses a patient's characteristics to assign him/her into one of 3 risk groups, which can be used to estimate the 30-day survival risk. (There are many other such prognostic scales, including (Chuang et al., 2004; Anderson et al., 1995; Haybittle et al., 1982).) Again, these tools are typically evaluated using Concordance.[7]

6. `https://siteman.wustl.edu/prevention/ydr`. The cancer scores, for example, use models that are based on group consensus among researchers to identify risk factors, with risk points allocated according to the strength of the causal association, obtained from population average risk of cancer and cumulative 10-year risk obtained from SEER data (Colditz et al., 2000).
7. Here, they do not compare pairs of individuals from the same group, but only patients from different groups, whose events are comparable (given censoring); see Section 3.1.

2.3. [P,1_{t^*},i]: Single-time Individual Probabilistic Predictors (Gail, PredictDepression)

Another class of single-time predictors each produce a *survival probability* $\hat{S}(\, t^* \,|\, \vec{x}\,) \in [0,1]$ for each individual patient $\vec{x}$, for a single fixed time t^* — which is the *probability* $\in [0,1]$ that $\vec{x}$ will survive to at least time t^*. For example, the Gail model [Gail] (Costantino et al., 1999)[8] estimates the probability that a woman will develop breast cancer within 5 years based on her responses to a number of survey questions. Similarly, the PredictDepression system [PredDep] (Wang et al., 2014)[9] predicts the probability that a patient will develop a major depressive episode in the next 4 years based on a small number of responses. The Apervita[10] and R-calc[11] websites each include dozens of such tools, each predicting the survival probability for 1 (or perhaps 2) fixed time points for certain classes of diseases.

Notice these probability values have semantic content, and are labels for *individual patients* (rather than risk-scores, which are only meaningful within the context of other patients' risk scores). These systems should be evaluated using a calibration measure, such as 1-Calibration or Brier score (discussed in Sections 3.3 and 3.4).

2.4. [P,∞,g]: Group Survival Distribution (KM)

There are many systems that can produce a survival distribution: a graph of $[t, \hat{S}(\, t\,)]$, showing the survival probability $\hat{S}(\, t\,) \in [0,1]$ for each time $t \geq 0$; see Figure 1. The Kaplan-Meier analytic tool (KM) is at the "class" level, producing a distribution designed to apply to everyone in a sub-population: $\hat{S}(\, t \,|\, \vec{x}\,) = \hat{S}(\, t\,)$, for every $\vec{x}$ in some class — *e.g.*, the KM curve in Figure 1[left] applies to every patient $\vec{x}$ with stage-4 stomach cancer. The SEER website[12] provides a set of Kaplan-Meier curves for various cancers. While patients can use such information to get a crude estimate of their survival probabilities, the original goal of that analysis is to better understand the disease itself, perhaps by seeing whether some specific feature made a difference, or if a treatment was beneficial. For example, we could produce one curve for all stage-4 stomach cancer patients who had treatment tA, and another for the disjoint subset of patients who had no treatment; then run a log-rank test (Harrington, 2005) to determine whether (on average) patients receiving treatment tA survived statistically longer than those who did not. Section 3 below describes various ways to evaluate [P,∞,i] models; we will use these measures to evaluate KM models as well.

2.5. [P,∞,i]: Individual Survival Distribution, ISD (COX-KP, COXEN-KP, AFT, RSF-KM, MTLR)

The previous two subsections described two frameworks:

- [P,1_{t^*},i] tools, which produce an *individualized* probability value $\hat{S}(\, t^* \,|\, \vec{x}_i\,) \in [0,1]$, but only for a single time t^*; and

8. Available at `https://bcrisktool.cancer.gov/`.
9. Available at `http://predictingdepression.com/`.
10. Available at `https://community.apervita.com/community`.
11. Available at `http://riskcalc.org/`.
12. Available at `http://seer.cancer.gov/`.

- [P,∞,g] tools, which produce the entire survival probability curve $[t, \hat{S}(t)]$ *for all points* $t \geq 0$, but are not individuated — *i.e.*, the same curve for all patients $\{ \vec{x}_i \}$.

Here, we consider an important extension: a tool that produces *the entire survival probability curve* $\{ [t, \hat{S}(t \mid \vec{x}_i)] \}_t$ *for all points* $t \geq 0$, *specific to each individual patient* $\vec{x}_i$. As noted in the previous section, this is required by any application that requires knowing meaningful survival probabilities for many time points. This model also allows us to compute other useful statistics, such as a specific patient's expected survival time.

We call each such system an "Individual Survival Distribution" model (ISD). While the Cox model is often used just to produce the risk score, it can be used as an ISD, given an appropriate (learned) baseline hazard function $\lambda_0(t)$; see Equation 2. We estimate this using the Kalbfleisch-Prentice estimator (Kalbfleisch and Prentice, 2002), and call this combination "COX-KP"; we also consider a regularized Cox model, namely the elastic net Cox with the Kalbfleisch-Prentice extension (COXEN-KP). We also explore three other models: Accelerated Failure Time model (Kalbfleisch and Prentice, 2002) with the Weibull distribution (AFT), Random Survival Forests with the Kaplan-Meier extension (RSF-KM, described in Appendix C.2) (Ishwaran et al., 2008) and Multi-task Logistic Regression system (MTLR; Yu et al., 2011). Figure 5 shows the curves from these various models, each over the same set of individuals.

Above, we briefly mentioned three evaluation methods: Concordance, 1-Calibration, and Brier score. We show below that we can use any of these methods to evaluate a ISD model. In addition, we can also use variants of "L1-loss" to see how far a predicted single-time differs from the true time of death; see Section 3.2. Each of these 4 methods considers only a single time point of the distribution, or an average of scores, each based on only a single time, or a single statistic (such as its median value). We also consider a novel evaluation measure, "D-Calibration", which uses the entire distribution of estimated survival probabilities; see Section 3.5.

2.6. Other Issues

The goal of many Survival *Analysis* tools is to identify relevant variables, which is different from our challenge here of making a prediction about an individual. Some researchers use KM to test whether a variable is relevant — *e.g.*, they partition the data into two subsets, based on the value of that variable, then run KM on each subset and declare that variable to be relevant if a log-rank test claims these two curves are significantly different (Harrington, 2005). It is also a common use of the basic Cox model — in essence, by testing if the $\hat{\beta}_i$ coefficient associated with feature x_i (in Equation 2) is significantly different from 0 (Therneau and Grambsch, 2013). We will later use this approach to select features, as a pre-processing step, before running the actual survival prediction model; see Section 4.1.

Note this "*g vs i*" distinction is not always crisp, as it depends on how many variables are involved — *e.g.*, models that "describe" each instance using no variables (like KM) are clearly "*g*", while models that use dozens or more variables — enough to distinguish each patient from one another — are clearly "*i*". But models that involve 2 or 3 variables typically will place each patient into one of a small number of "clusters", and then assign

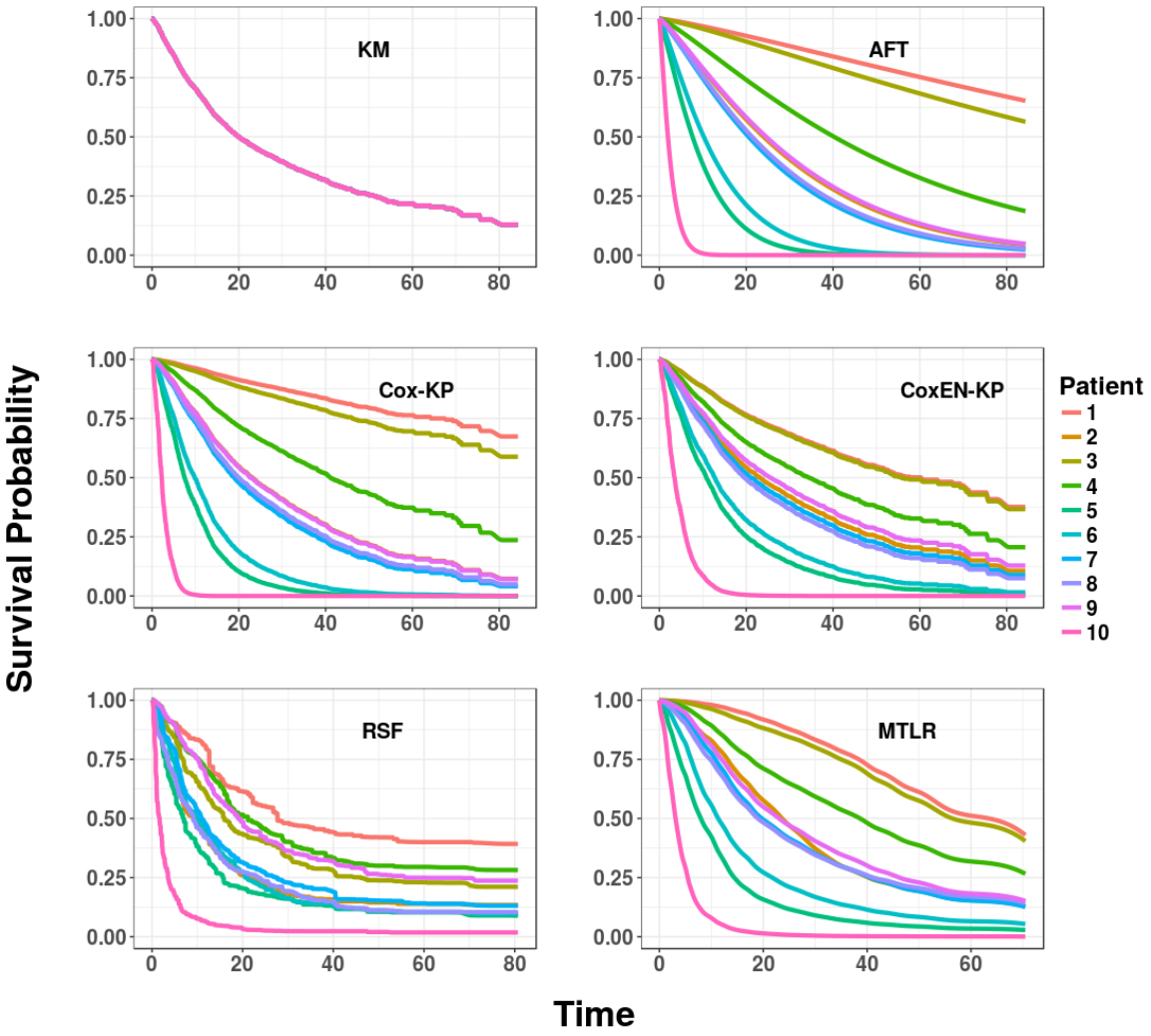


Figure 5: Survival curves (within one of 5 folds) of 10 cancer patients for the KM model and all 5 ISD models considered here, evaluated on the NACD data set (described in Section 4.1). As the KM curve (top left) is the same for all patients by definition, we provide only 1 curve — here smoothed. Note that the set of curves for AFT (with the Weibull distribution), COX-KP, and COXEN-KP each have roughly the same shape, and do not cross, due to the proportional hazards assumption, whereas the curves for RSF-KM and MTLR can cross.

the same values to each member of a cluster. By convention, we will catalog those models as "g" as the decision is not intended to be at an individual level.

The "1_{t^*}" vs "∞" distinction can be blurry, if considering a system that produces a small number $k > 1$ of predictions for each individual — *e.g.*, the Gail model provides a prediction of both 5 year and 25 year survival. We consider this system as a pair of "1_{t^*}"-predictors, as those two models are different. (Technically, we could view them as "Gail[5year]" versus "Gail[25year]" models.)

Finally, recall there are two types of frameworks that each return a single value for each instance: the single value returned by the [R,$1_\forall$,i]-model COX is *atemporal* — *i.e.*, applies to the overall model — while each single value returned by the [P,1_{t^*},i]-model Gail and the [R,1_{t^*},g]-model PaP, is for a specific time, t^*. (Note there can also be [P,$1_\forall$,i]- and [R,$1_\forall$,g]-models that are atemporal.)

2.7. Relationship of Distributional Models to Other Survival Analysis Systems

We will use the term "Distributional Model" to refer to algorithms within the [P,∞,g] and [P,∞,i] frameworks — *i.e.*, both KM and ISD models. Note that such models can match the functionality of the first 3 "personalized" approaches. First, to emulate [P,1_{t^*},i], we just need to evaluate the distribution at the specified single time t^* — *i.e.*, $\hat{S}(\, t^* \,|\, \vec{x}\,)$. So for Patient #1 (from Figure 1), for t^* = "48 months", this would be 20%. Second, to emulate [R,1_{t^*},i], we can just use the negative of this value as the time-dependent risk score — so the 4-year risk for Patient #1 would be -0.20. Third, to deal with [R,$1_\forall$,i], we need to reduce the distribution to a single real number, where larger values indicate shorter survival times. A simple candidate is the individual distribution's median value, which is where the survival curve crosses 50%.[13] So for Patient #1 in Figure 1, the median is $\hat{t}_1^{(0.5)} = 16$ months. We can then view (the negative of) this scalar as the risk score for that patient. So for Patient #1, the "risk" would be $r(\vec{x}_1) = -16$. Fourth, to view the ISD model in the [R,$1_\forall$,g] framework, we need to place the patients into a small number of "relatively homogeneous" bins. Here, we could quantize the (predicted) mean value — *e.g.*, mapping a patient to Bin #1 if that mean is in $[0, 15)$, Bin #2 if in $[15, 27)$, and Bin #3 if in $[27, 70]$. (Here, this patient would be assigned to Bin #2.) Fifth, to view the ISD model in the [R,1_{t^*},g] framework, associated with a time t^*, we could quantize the t^*-probability — *e.g.*, quantize the $\hat{S}(\, t^* = 48 \text{ months} \,|\, \vec{x}\,)$ into 4 bins corresponding to the probability intervals $[0, 0.20)$, $[0.20, 0.57)$, $[0.57, 0.83]$, and $[0.83, 1.0]$.

These simple arguments show that a distributional model can produce the scalars used by five other frameworks [P,1_{t^*},i], [R,1_{t^*},i], [R,$1_\forall$,i], [R,$1_\forall$,g], and [R,1_{t^*},g]. Of course, a distributional model can also provide other information about the patient — not just the probability associated with one or two time points, but at essentially any time in the future, as well as the mean/median value. Another advantage of having such survival curves is *visualization* (see Figure 1): it allows the user (patient or clinician) to see the *shape* of the curve, which provides more information than simply knowing the median, or the chance of surviving 5 years, etc.

13. Another candidate is the mean value of the distribution, which corresponds to the area under the survival curve; see Theorem B.1.

There are some subtle issues related to producing meaningful survival curves — *e.g.*, many curves end at a non-zero value: note the KM curve in Figure 5(top left) stops at [83, 0.12], rather than continuing to intersect the x-axis at, perhaps [103, 0.0]. This is true for many of the curves produced by the ISDs. Indeed, some of the curves do not even cross $y = 0.5$, which means the median time is not well-defined; *cf.* the top orange line on the AFT curve (top right), which stops at (83, 0.65), as well as many of the other curves throughout that figure. This causes many problems, in both interpreting and evaluating ISD models. Appendix A describes how we address this.

3. Measures for Evaluating Survival Analysis/Prediction Models

The previous section mentioned 5 ways to evaluate a survival analysis/prediction model: Concordance, 1-Calibration, Brier score, L1-loss, and D-Calibration. This section will describe these: quickly summarizing the first four (standard) evaluation measures (and leaving the details, including discussion of censoring, for Appendix B) then providing a more thorough motivation and description of the fifth, D-Calibration. The next section shows how the 6 distribution-learning models (the 5 ISD models, and KM; see Figure 5) perform with respect to these evaluations.

For notation, we will assume models were trained on a training data set, formed from the same triples as shown in Equation 1, that is $D = D_U \cup D_C$ where $D_U = \{ [\vec{x}_j, d_j, \delta_j = 1] \}_j$ is the set of *uncensored* instances (notice the event time, t_j, here is written as d_j), and $D_C = \{ [\vec{x}_k, c_k, \delta_k = 0] \}_k$ is the set of *censored* instances (t_k, here is written as c_k). Note also that this training data set D is disjoint from the validation data set, V. Since models are evaluated on V and we save discussion of censoring for Appendix B, we assume here that all of V is uncensored— *i.e.*, for now, $V = V_U = \{ [\vec{x}_j, d_j, \delta_j = 1] \}_j \approx \{ [\vec{x}_j, d_j] \}_j$ (to simplify notation).

3.1. Concordance

As noted above, each individual risk model [R,1.,-] (*i.e.*, [R,1.,i] or [R,1.,g], where "1." can be either "1_{t^*}" or "$1_\forall$") assigns to each individual $\vec{x}_i$ a "risk score" $r(\vec{x}_i) \in \Re$, where $r(\vec{x}_a) > r(\vec{x}_b)$ means the model is predicting that $\vec{x}_a$ will die before $\vec{x}_b$. Concordance (a.k.a. C-statistic, C-index) is commonly used to validate such risk models. Specifically, Concordance considers each pair of patients, and asks whether the predictor's values for those patients matches what actually happened to them. In particular, if the model gives $\vec{x}_a$ a higher score than $\vec{x}_b$, then the model gets 1 point if $\vec{x}_a$ dies before $\vec{x}_b$. If instead $\vec{x}_b$ were to die before $\vec{x}_a$, the model gets 0 points for this pair. Concordance computes this for all pairs of *comparable* patients, and returns the average.

When considering only uncensored patients, every pair is comparable, which means there are $\binom{n}{2} = \frac{n \cdot (n-1)}{2}$ pairs from $n = |V_U|$ elements. Given these comparable pairs, Concordance is calculated as:

$$\widehat{C}(\, V_U,\, r(\cdot)\,) \quad = \quad \frac{1}{\frac{|V_U| \,\cdot\, (|V_U|-1)}{2}} \cdot \sum_{[\vec{x}_i, d_i] \in V_U} \; \sum_{[\vec{x}_j, d_j] \in V_U \,:\, d_i < d_j} \mathcal{I}\left[\, r(\vec{x}_i) > r(\vec{x}_j) \,\right] \;. \qquad (3)$$

Id	d_i	Risk$_i$
1	1	6
2	3	3
3	4	5
4	6	2
5	9	4

	1	2	3	4	5
1					
2	+				
3	+	0			
4	+	+	+		
5	+	0	+	0	

Table 1: Simple example to illustrate Concordance (here, with only uncensored patients). Left: time of death, and risk score, for 5 patients. Right: "+" means the row-patient had a lower risk, and died after, the column-patient; otherwise "0".

As an example, consider the table of death times $\{d_i\}$ and risk scores for 5 patients, shown in Table 1[left]. Table 1[right] shows that these risk scores are correct in 7 of the $\binom{5}{2} = 10$ pairs, so the Concordance here is $7/10 = 0.7$.

This Concordance measure is very similar to the area under the receiver operating curve (AUC) and equivalent when d_i is constrained to values $\{0, 1\}$ (Li et al., 2016).

This Concordance measure is relevant when the goal is to *rank* or *discriminate* between patients — *e.g.*, when one wants to know who will live longer between a pair of patients. For example, if we want to transplant an available liver to the patient who will die first — this corresponds to "urgency". Concordance is the desired metric here due to its interpretation, *i.e.* given two *randomly selected* patients, $\vec{x}_a$ and $\vec{x}_b$, if a model with Concordance of 0.9 assigns a higher risk score to $\vec{x}_a$ than $\vec{x}_b$, then there is a 90% chance that $\vec{x}_a$ will die before $\vec{x}_b$.

While [R,$1_\forall$,i] models (such as COX) provide a risk score that is independent of time, there are also [R,∞,i] models that produces a risk score $r(\vec{x}, t)$ for an instance $\vec{x}$ that depends on time t, such as Aalen's additive regression model (Aalen et al., 2008) or time-dependent Cox (td-Cox; Fisher and Lin, 1999), which uses time-dependent features. These models can be evaluated using *time-dependent Concordance* (aka, "time-dependent ROC curve analysis"; Heagerty and Zheng, 2005).

Finally, the [R,−,g] systems compute a risk score, but then bin these scores into a small set of intervals. When computing Concordance, they then only consider patients in different bins. For example, if Bin1 = [0, 10] and Bin2 = [11, 20], then this evaluation would only consider pairs of patients $(\vec{x}_a, \vec{x}_b)$ where one is in Bin1 and the other is in Bin2 — *e.g.*, $r(\vec{x}_a) \in [0, 10]$ and $r(\vec{x}_b) \in [11, 20]$. Hence, it will not consider the pair $(\vec{x}_c,\ \vec{x}_d)$ if both $r(\vec{x}_c),\ r(\vec{x}_d) \in [11, 20]$.

See Appendix B.1 for more details, including a discussion of how this measure deals with censored instances and tied risk scores/death times.

3.2. L1-loss

Survival prediction is very similar to regression: given a description of a patient, predict a real number (his/her time of death). With this similarity in mind, one can evaluate a survival model using the techniques used to evaluate regression tasks, such as L1-loss — the average absolute value of the difference between the true time of death, d_i, and the

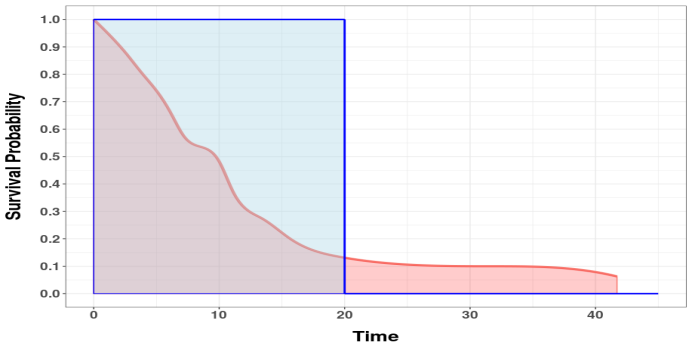


Figure 6: Example of a survival curve (in red), superimposed (in blue) with a degenerate curve that puts all of its weight on a single time point, which means it assigns 100% chance of dying at exactly this time.

predicted time $\widehat{d}_i$: $\frac{1}{n}\sum_i |d_i - \widehat{d}_i|$. We consider the L1-loss, rather than L2-loss which squares the differences, as the distribution of survival times is often right skewed, and L1-loss is less swayed by outliers than L2-loss.

One challenge in applying this measure to our [P,∞,-] models is identifying the predicted time, $\widehat{d}_i$. Here, we will use the predicted median survival time $\hat{t}_i^{(0.5)}$ — that is we set $\hat{d}_i = \hat{t}_i^{(0.5)}$ — leading to the following measure:

$$L1(\, V_U,\ \{\, \hat{S}(\,\cdot \,|\, \vec{x}_i\,)\,\}_i\,) \quad = \quad \frac{1}{|V_U|} \sum_{[\vec{x}_i, d_i] \in V_U} \left| d_i - \hat{t}_i^{(0.5)} \right|. \tag{4}$$

While we would like this value to be small, we should not expect it to be 0: if the distribution is meaningful, there should be a non-zero chance of dying at other times as well. For example, while the L1-loss is 0 for the Heaviside distribution at the time of death (shown in blue in Figure 6), this is unrealistic.

While this L1-loss is not a proper scoring rule (Gneiting and Raftery, 2007), we include this measure as it is intuitive in understanding the error of predicted survival times. However, reducing the survival distribution to a single value has shown to be fairly inaccurate (Henderson and Keiding, 2005); this finding is consistent with our empirical results in Section 4.2. Appendix B.2 discusses many other issues with the L1-loss measure, related to censored data.

3.3. 1-Calibration

The [P,1_{t^*},i] tools estimate the survival probability $\hat{S}(\, t^* \,|\, \vec{x}\,) \in [0, 1]$ for each instance $\vec{x}$ at a single time point t^*. For example, the PredictDepression system (Wang et al., 2014) predicts the chance that a patient will have a major depression episode within the next 4 years, based on their current characteristics — *i.e.*, this tool produces a single probability value $\hat{S}(\, 4\text{yr} \,|\, \vec{x}_i\,) \in [0, 1]$ for each patient described as $\vec{x}_i$. We can use 1-Calibration to measure the effectiveness of such predictors. To help explain this measure, consider the "weatherman task" of predicting, on day t, whether it will rain on day $t + 1$. Given the

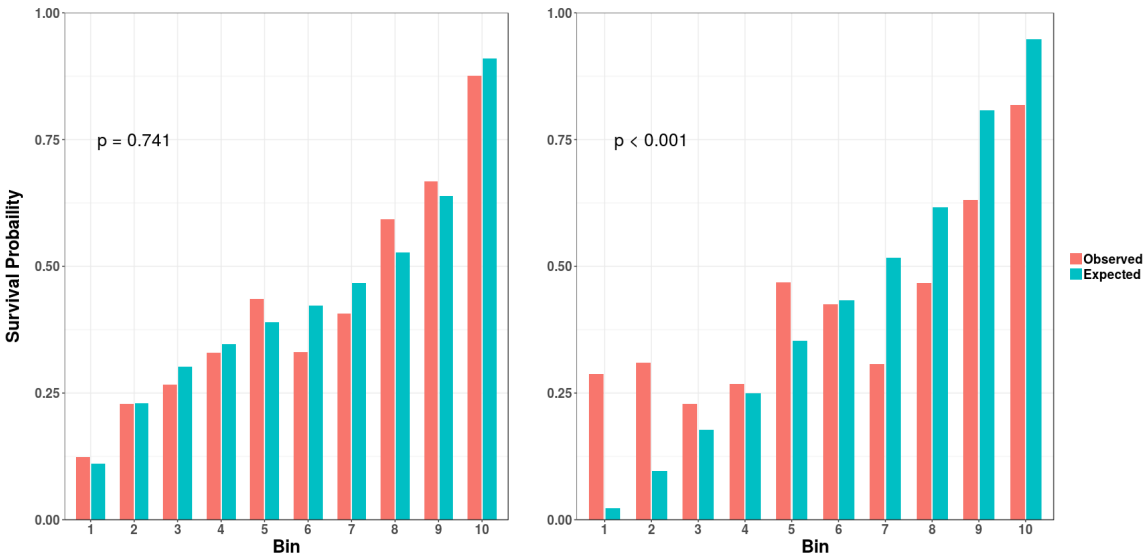


Figure 7: The binned observed versus expected probabilities associated with two 1-Calibration computations, for the MTLR [left] and the RSF-KM [right] model applied to the GBM data set for the 50th percentile of time (369.5 days).

uncertainty, forecasters provide probabilities. Imagine, for example, there were 10 times that the weatherman, Mr.W, predicted that there was a 30% chance that it would rain tomorrow. Here, if Mr.W was calibrated, we expect that it would rain 3 of these 10 times — *i.e.*, 30%. Similarly, of the 20 times Mr.W claims that there is an 80% chance of rain tomorrow, we expect rain to occur $16 = 20 \times 0.8$ of these 20 times.

Here, we have described a binary probabilistic prediction problem — *i.e.*, predicting the chance that it will rain the next day. One of the most common calibration measures for such binary prediction problems is the Hosmer-Lemeshow goodness-of-fit test (Hosmer and Lemesbow, 1980). First, we sort the predicted probabilities for this time t^* for all patients $\{\, \hat{S}(\, t^* \,|\, \vec{x}_i \,)\, \}_i$ and group them into a number (B) of "bins"; commonly into deciles, *i.e.*, $B = 10$ bins. Suppose there are 200 patients; the first bin would include the 20 patients with the largest $\hat{S}(\, t^* \,|\, \vec{x}_i \,)$ values, the second bin would contain the patients with the next highest set of values, and so on, for all 10 bins. Next, *within each bin*, we calculate the expected number of events, $\bar{p_j} = \frac{1}{|B_j|} \sum_{\vec{x}_i \in B_j} (1 - \hat{S}(\, t^* \,|\, \vec{x}_i \,))$. We also let $n_j = |B_j|$ be the size of the j^{th} bin (here, $n_1 = n_2 = \cdots = n_{10} = 200/10 = 20$), and O_j be the number of patients (in the j^{th} bin) who died before t^*. Recalling that d_i denotes Patient i's time of death and letting $o_i = \mathcal{I}\,[\, d_i \leq t^* \,]$ denote the event status of the i^{th} patient at t^*: for the jth bin, B_j, we have $O_j = \sum_{\vec{x}_i \in B_j} o_i$. Figure 7 graphs the 10 values of observed O_j and expected $n_j\, \bar{p_j}$ for the deciles, for two different tests (corresponding to two different ISD-models, on the same data set and t^* time). Additionally, see Appendix B.3 for an example walking through 1-Calibration and Algorithm 1 which summarizes this complete process (with censoring included).

For each test, we can then compute the Hosmer-Lemeshow test statistic:

$$\widehat{HL}(\, V_U,\, \hat{S}(\, t^* \,|\, \cdot\,)\,) \quad = \quad \sum_{j=1}^{B} \frac{(O_j - n_j \bar{p_j})^2}{n_j\, \bar{p_j}\, (1 - \bar{p_j})}\ . \tag{5}$$

If the model is 1-Calibrated, then this statistic follows a χ^2_{B-2} distribution, which can then be used to find a p-value. For a given time t^*, finding $p < 0.05$ suggests the survival model is <u>not</u> well calibrated at t^* — *i.e.*, the predicted probabilities of survival at t^* may not be representative of the patient's true survival probability at t^*.

Returning to Figure 7, the HL statistics are 5.99 and 321.44, for the left and right, leading to the p-values p =0.741 and p <0.001 — meaning the left one passes but the right one does not. This is not surprising, given that each pair of bars on the left are roughly the same height, while the pairs of the right are not.

Note that a [P,∞,i] model, which gives probabilities for multiple time points, may be calibrated at one time t_1, but not be calibrated at another time t_2, since O_j, as the $\bar{p_j}$ values are dependent on the chosen time point. This issue motivated us to define a notion of calibration across a distribution of time points, D-Calibration, in Section 3.5. Appendix B.3 provides further details about 1-Calibration, including ways to handle censored patients.

3.4. [Integrated] Brier Score

We often want a model to be both discriminative (high Concordance) and calibrated (passes the 1-Calibration test). While one can rank Concordance scores to compare two models' discriminative abilities, 1-Calibration cannot rank models besides suggesting one model is calibrated ($p \geq 0.05$) but another is not ($p < 0.05$) (as p-values are not intended to be ranked). The Brier score (Brier and Allen, 1951) is a commonly used metric that measures both calibration and discrimination (Murphy, 1972, 1973; DeGroot and Fienberg, 1983). Mathematically, the Brier score is the mean squared error between the {0, 1} event status at time t^* and the predicted survival probability at t^*. Given a fully uncensored validation set V_U, the Brier score, at time t^*, is

$$BS_{t^*}\left(\, V_U,\ \hat{S}(\, t^* \,|\, \cdot\,)\,\right) \quad = \quad \frac{1}{|V_U|} \sum_{[\vec{x}_i, d_i] \in V_U} \left(\, \mathcal{I}\,[\, d_i \leq t^*\,] \ -\ \hat{S}(\, t^* \,|\, \vec{x}_i\,)\,\right)^2 . \tag{6}$$

Here, a perfect model (that only predicts 1s and 0s as survival probabilities and is correct in every case) will get the perfect score of 0, whereas a reference model that gives $\hat{S}(\, t^* \,|\, \cdot\,) = 0.5$ for all patients will get a score of 0.25.

An extension of the Brier score to an interval of time points is the *Integrated* Brier score, which will give an average Brier score across a time interval,

$$\text{IBS}(\, \tau,\, V_U,\, \hat{S}(\, \cdot \,|\, \cdot\,)\,) \quad = \quad \frac{1}{\tau} \int_0^{\tau} BS_t\left(\, V_U,\, \hat{S}(\, t \,|\, \cdot\,)\,\right) dt\ . \tag{7}$$

We will use this measure for our analysis, where τ is the maximum event time of the combined training and validation data sets — this way, the interval evaluated is equivalent across cross-validation folds.

As noted above, the Brier score measures both calibration and discrimination, implying it should be used when seeking a model that must perform well on both calibration and discrimination, or when one is investigating the overall performance of survival models. Appendix B.4 shows how to incorporate censoring into the Brier score.

3.5. D-Calibration

The previous sections summarized several common ways to evaluate standard survival prediction models, that produce only a single value for each patient — *e.g.*, the patient's risk score, perhaps with respect to a single time, or the mean survival time. (Each is a [-,1.,-] model.) However, the [P,∞,-] tools produce a distribution — *i.e.*, each is a function that maps $[0, \infty]$ to $[0, 1]$ (with some constraints of course), such as the ones shown in Figure 5; see Footnote 1. It would be useful to have a measure that examines the entire distribution as a distribution.[14]

A distributional calibration (D-Calibration) (Andres et al., 2018) measure addresses the critical question:

Should the patient believe the predictions implied by the survival curve? (8)

First, consider population-based models [P,∞,g], like Kaplan-Meier curves — *e.g.*, Figure 1[left], for patients with stage-4 stomach cancer. If a patient has stage-4 stomach cancer, should s/he believe that his/her median survival time is 11 months, and that s/he has a 75% chance of surviving more than 4 months? To test this, we could take 1000 new patients (with stage-4 stomach cancer) and ask whether ≈500 of these patients lived at least 11 months, and if ≈750 lived more than 4 months.

For notation, given a data set, D, and [P,∞,g]-model Θ, and any interval $[a, b] \subset [0, 1]$, let

$$D_\Theta(\,[a,b]\,) \quad = \quad \{\, [\vec{x}_i, d_i, \delta_i = 1] \in D \mid \hat{S}_\Theta(\,d_i\,) \in [a,b]\,\} \tag{9}$$

be the subset of (uncensored) patients in D whose time of death is assigned a probability (by Θ) in the interval $[a, b]$. For example, $D_\Theta(\,[0.5,\, 1.0]\,)$ is the subset of patients who lived at least the median survival time (using $\hat{S}_\Theta(\,\cdot\,)$'s median), and $D_\Theta(\,[0.25,\, 1.0]\,)$ is the subset who died after the 25th percentile of $\hat{S}_\Theta(\,\cdot\,)$. By the argument above, we expect $D_\Theta(\,[0,\, 0.5]\,)$ to contain about 1/2 of D, and $D_\Theta(\,[0.25,\, 1.0]\,)$ to contain about 3/4 of D. Indeed, for any interval $[a,\, 1.0]$, we expect

$$\frac{|D_\Theta(\,[a,\, 1.0]\,)|}{|D|} \quad = \quad 1 - a \tag{10}$$

or in general

$$\frac{|D_\Theta(\,[a,\, b]\,)|}{|D|} \quad = \quad b - a \tag{11}$$

This leads to the idea of a survival distribution [P,∞,g] model, Θ, being D-Calibrated: for each uncensored patient $\vec{x}_i$, we can observe when s/he died d_i, and also determine the probability for that time, based on Θ: $\hat{S}_\Theta(\,d_i\,)$. If Θ is D-Calibrated, we expect roughly

14. While the Integrated Brier score does consider all the points across the distribution, it simply views that distribution as a set of (x, y) points; see Section 3.5.2 for further explanation.

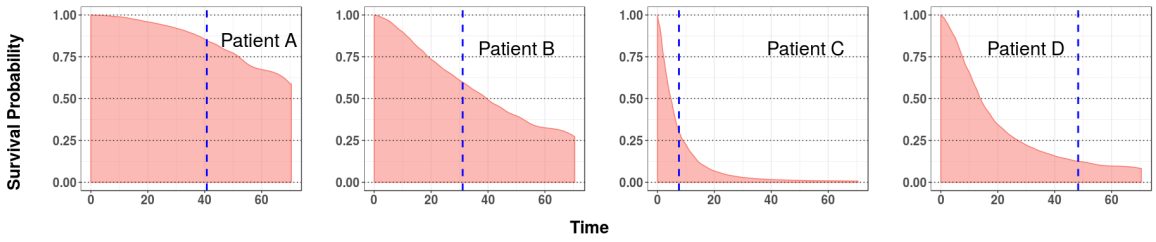


Figure 8: Four patients from the complete NACD data set. Notice each died in a different quartile (shown with a vertical dashed line); see Table 2.

10% of the patients to die in the [90%, 100%] interval — *i.e.*, $\frac{|D_\Theta([0.9,1.0])|}{|D|} \approx 1-0.9=0.1$ — and another 10% to die in the [80%, 90%) interval, and so forth for each of the 10 different 10%-intervals. More precisely, the set $\{\hat{S}_i(d_i)\}$ over all of the patients should be distributed uniformly on $[0,1]$, which means that each of the 10 buckets would contain 10% of D.

This suggests a measure to evaluate a distributional model: see how close each of these 10 buckets is to the expected 10%. The next subsection provides a statistical test to evaluate this condition.

3.5.1. Dealing with *Individual* Survival Distributions, isd

Everything above was for a *population*-based distributional model [P,∞,g]. These specific results do not apply to *individual* survival distributions [P,∞,i]: for example, consider a single patient, Patient #1, whose curve is shown in Figure 1[middle]. Should he believe this plot, which implies that his median survival time is 18 months, and that he has a 75% chance of surviving more than 13 months?

If we could observe 1000 patients exactly identical to this Patient #1, we could verify this claim by seeing their actual survival times: this survival curve is meaningful if its predictions matched the outcomes of those copies — *e.g.*, if around 250 died in the first 13 months, another ≈250 in months 13 to 18, etc.

Unfortunately, however, we do not have 1000 "copies" of Patient #1. But here we do have many other patients, each with his/her own characteristic survival curve, including the 4 curves shown in Figure 8. Notice each patient has his/her own distribution, and hence his/her own quartiles — *e.g.*, the predicted median survival times for Patient A (resp., B, C, and D), are 28.6 (resp., 65.7, 11.4, and 13.9) months; see Table 2. For these historical patients, we know the actual event time for each.[15] Here, if our predictor is working correctly, we would expect that 2 of these 4 would pass away before respective median times, and the other 2 after their median times. Indeed, we would actually expect 1 to die in each of the 4 quartiles; the blue vertical lines (the actual times of death) show that, in fact, this does happen. See also Table 2.

15. Here we just consider *uncensored* patients; Appendix B.5 extends this to deal with censoring.

Patient ID	Median Survival Time	Event time	Event Percentage	Quartile
A	85.5	43.4	84.7	#1
B	39.6	31.1	59.8	#2
C	4.7	7.5	30.4	#3
D	13.9	48.3	12.8	#4

Table 2: Description of 4 patients from the NACD data set; see also Figure 8.

With a slight extension to the earlier notation (Equation 9), for a data set D and [P,∞,i]-model Θ, and any interval $[a, b] \subset [0, 1]$, let

$$D_\Theta(\,[a,b]\,) \quad = \quad \{\; [\vec{x}_i,\, d_i,\, \delta = 1] \in D \mid \hat{S}_\Theta(\, d_i \,|\, \vec{x}_i\,) \in [a,b]\;\} \tag{12}$$

be the subset of (uncensored) patients in the data set D whose times of death are assigned a probability (based on their individual distribution, computed by Θ) in the interval $[a, b]$.

As above, we could put these $\hat{S}_\Theta(\, d_i \,|\, \vec{x}_i\,)$ into "10%-buckets", and then ask if each bucket holds about 10% of the patients. The right side of Figure 9 plots that information — for the ISD Θ learned by MTLR from the NACD data set (described in Section 4.1) — as a sideways histogram.[16] (The rust-colored intervals correspond to the censored patients; see Appendix B.5 for the details.) We see that each of these intervals is very close to 10%.

This leads to a straightforward evaluation, based on Pearson's χ^2 test: compute the χ^2-statistic with respect to the ten 10% intervals, and ask whether the buckets appear uniform at (say) the $p > 0.05$ level. Here, this χ^2 goodness-of-fit test yields $p = 0.433$, which suggests that this ISD is sufficiently uniform that we can believe that this survival model is D-calibrated. Algorithm 2 in Appendix B.5 summarizes the algorithm itself.

Theorem B.3 (in Appendix B.5) proves the appropriateness of this D-calibration test by showing that the "correct" conditional survival distribution will be D-calibrated. This addresses the question posed at the start of this subsection (Question 8):

> Yes, a patient should believe the prediction from the survival curve whenever this goodness-of-fit test reports $p > 0.05$.

That appendix also proves that the simple KM framework is asymptotically D-calibrated; see Lemma B.4.

3.5.2. Further motivating D-calibration

This subsection provides additional arguments motivating D-calibration, by showing that it is useful for evaluating if a given ISD model is meaningful, and it is fundamentally different from the other measures, especially 1-Calibration and IBS.

As noted at the start of Section 3.5, most evaluations measures deal only with survival analysis models that produce only a single number ("scalar") for each patient: *e.g.*, Concordance deals with models that produce a scalar risk score, L1-loss with models that produce

16. This figure actually shows 5-fold cross-validation results: the survival distribution for each patient was computed based on the model learned from the other 4/5 of the data, which is then applied to this patient (Witten et al., 2011).

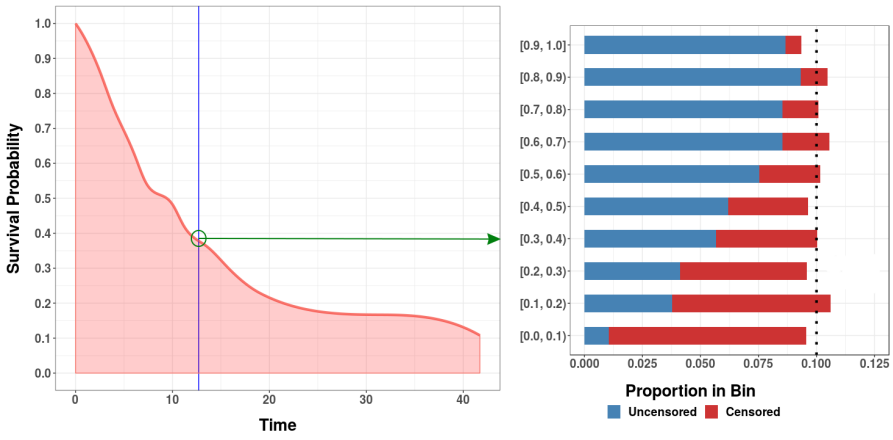


Figure 9: The right side shows the (sideways) "calibration histogram" associated with the model learned for the NACD data set. The left portion shows the survival curve for a patient $\vec{x}_{27}$ — here we see that this patient's event d_{27} =12.7 months, corresponds to $\hat{S}(\, d_{27} \,|\, \vec{x}_{27}\,)$ = 39.4%, which means the patient contributed to the [30, 40) bin. In a completely D-calibrated model, each of these horizontal bars would be 10%; here, we see that each of the 10 bars is fairly close. See also Figure 15.

an estimated scalar time-to-death, and 1-calibration (and [single-time] Brier score) with models that produce a scalar probability of surviving until a given single time t^*. While ISD-models can produce such numbers, they actually produce more: an entire distribution, giving a survival probability for each time point.

D-calibration vs IBS: The Integrated Brier Score does deal with many time points, by averaging the values of the [single time] Brier score, over the entire time interval; see Equation 7. However, this measure still does not view these curves as (conditional) survival *distributions*, in that it does not prefer models where one can interpret $S(\, t \,|\, \vec{x}\,)$ in the obvious way. In particular, as explained above, $S(\, \hat{t}_i^{(0.5)} \,|\, \vec{x}_i\,) = 0.5$ should mean there is a 50% chance that $\vec{x}_i$ will live at least $\hat{t}_i^{(0.5)}$ months, in that half of all patients $\{\vec{x}_i\}$ should die before their respective median times $\{\hat{t}_i^{(0.5)}\}$ — that is, if each patient $\vec{x}_i$ dies at d_i, then half of the $\{S(\, d_i \,|\, \vec{x}_i\,)\}$ values should be under 0.5, and the other half should be over. Recall this is what it means to claim that the set of survival curves are meaningful, when discussing how to answer Question 8, discussed above.

Now consider the plot on the left of Figure 10, where each "$*$" indicates when each person died, d_i. Here, we see $\hat{S}_{Left}(\, d_Q \,|\, \vec{x}_Q\,)$ = 1.0 and $\hat{S}_{Left}(\, d_G \,|\, \vec{x}_G\,)$ = 0.5 — *i.e.*, half of the patients had $\hat{S}_{Left}(\, d_i \,|\, \vec{x}_i\,) \in [0,\, 0.5]$ and half had $\hat{S}_{Left}(\, d_i \,|\, \vec{x}_i\,) \in (0.5,\, 1]$, as desired. Here, if a third patient $\vec{x}_3$, with predicted curve $\hat{S}_{Left}(\, \cdot \,|\, \vec{x}_3\,)$, found $\hat{S}_{Left}(\, 20 months \,|\, \vec{x}_3\,) = 0.5$, s/he would have reason to believe that s/he had a 50% chance of living 20 months.

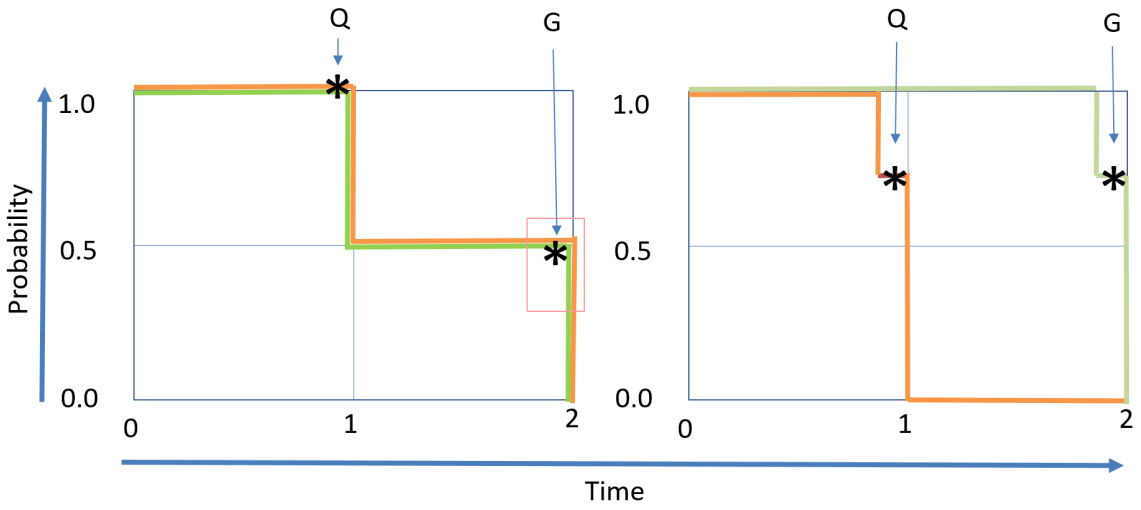


Figure 10: (left) ISD model showing the survival curves for two patients (Q and G), where the "*" shows when each died. This model is D-calibrated (with respect to two buckets), but has a relatively bad IBS score of 0.125. (right) Another ISD model, over the same two patients, that is not D-calibrated, but has a much better IBS score of 0.00521. (See Proposition B.5, in Appendix B.5.1, for details.)

Now consider the right plot, where $\hat{S}_{Right}(\, d_Q \,|\, \vec{x}_Q \,) \;=\; \hat{S}_{Right}(\, d_G \,|\, \vec{x}_G \,) \;=\; 0.75$ — that is, 100% of the patients died in the (0.5, 1] interval. Given this, it is not clear what the patient $\vec{x}_3$ can infer about $\hat{S}_{Right}(\, 20 months \,|\, \vec{x}_3 \,) = 0.5$.

This argues that a measure that regards the survival curves as *distribution* should prefer $\hat{S}_{Left}(\cdot\,|\,\cdot)$ over $\hat{S}_{Right}(\cdot\,|\,\cdot)$. However, Table 3 shows that this is not true for the IBS score, as the IBS score of the D-calibrated $\hat{S}_{Left}(\cdot\,|\,\cdot)$ is significantly worse than IBS score of the non-D-calibrated $\hat{S}_{Right}(\cdot\,|\,\cdot)$. This pair of examples also illustrates that D-calibation is fundamentally different from IBS as they are evaluating significantly different criteria, and so can have different preferences between models. Another distinction is interpretation: we explained above how to interpret the claim that a model is D-calibrated, however, it is not clear how one can interpret the claim that, say, a model's IBS=0.2.[17]

Indeed, it was these arguments that motivated us to introduce D-Calibration, to determine whether a proposed ISD-model produces meaningful distributions — *i.e.*, that the probabilities produced truly reflect the likelihood of death over the population.

Finally, to see that these two metrics are measuring different aspects, we will later see that the IBS scores for the {AFT, COX-KP, COXEN-KP, MTLR} models are all well within

17. We also note an IBS score is (near) 0 only if each of the curves is essentially a Heaviside function — like Figure 6 (or a variant, as shown in Figure 10) — where the (near) vertical transition occurs precisely at the time of death. Of course, this model is not robust, as the IBS score would be significantly larger if the Heaviside function was even slightly off — with a per-patient error of $|\eta_i|$ for each patient whose Heaviside was at $d_i + \eta_i$, while the time of death was actually d_i. There are similarly large penalties associated with censored instances.

model⇓	D-calibrated ?	IBS
Left: $\hat{S}_{Left}(\cdot \mid \cdot)$	**Yes**	0.125
Right: $\hat{S}_{Right}(\cdot \mid \cdot)$	No	**0.00521**

Table 3: Comparing models, from Figure 10.

1 standard error of one another for the GBM data set, but only coxen-kp and mtlr are D-Calibrated. (This is also true for the GLI data set.)

D-Calibration vs 1-Calibration: This standard notion of 1-Calibration appears similar to D-Calibration, as both involve binning probability values and applying a goodness-of-fit test. However, 1-Calibration involves a single prediction time — here $\hat{S}(\, t^* \,|\, \vec{x}_i \,)$, which is the probability that the patient $\vec{x}_i$ will survive at least to the specified time, t^*. Patients are then sorted by these probabilities, partitioned into equal-size bins, and assessed as to whether the observed survival rates for each bin match the predicted rates using a Hosmer-Lemeshow test. By contrast, D-Calibration considers the entire curve, $\hat{S}(\, t \,|\, \vec{x}_i \,)$ over all times t, producing curves like the ones shown in Figures 1, 5, and 8. Each curve corresponds to a patient who has an associated time of death, d_i. Here, we are considering the model's (estimated) probability of the patient's survival at his/her time of death, given by $\hat{S}_i(\, d_i \,|\, \vec{x}_i \,)$. These patients are then placed into $B = 10$ buckets,[18] based on the values of their associated probabilities, $\hat{S}_i(\, d_i \,|\, \vec{x}_i \,)$. Here the goodness-of-fit test measures whether the resulting buckets are approximately equal-sized, as would be expected if this model accurately estimated the true survival curves (argued further in Appendix B.5).

Note D-Calibration tests the proportion of instances in buckets across the entire $[0, 1]$ interval, but this is not required for the "single probability" 1-Calibration. For example, the single probability estimates for the rsf-km curve in Figure 4, at time 20, ranges only from 0.05 to 0.62. That is, the distribution calibration $\{\, \hat{S}_i(\, d_i \,|\, \vec{x}_i \,) \,\}$ should match the uniform distribution over [0,1], while the single probability calibration $\{\, \hat{S}_i(\, t^* \,|\, \vec{x}_i \,) \,\}$ is instead expected to match the empirical percentage of deaths. Table 4 summarizes the differences between D-Calibration and 1-Calibration.[19]

To see that they are measuring different aspects, Figure 11 presents one simple model that is perfectly D-Calibrated but not 1-Calibrated, and another model that is perfectly 1-Calibrated but not D-Calibrated; see proof of Proposition B.6, in Appendix B.5.1. In addition, we will see several examples below of this difference — *e.g.*, coxen-kp is D-Calibrated for the GLI data set, but it is not 1-Calibrated at any of the 5 time points considered, and aft is 1-Calibrated for the 50th and 75th percentiles of GBM but is not D-Calibrated.

18. Note the number of buckets does not have to be 10 — we chose 10 to match the typical value chosen for the 1-Calibration test.
19. Further differences occur when considering how censored patients are handled; see Appendices B.3 and B.5.

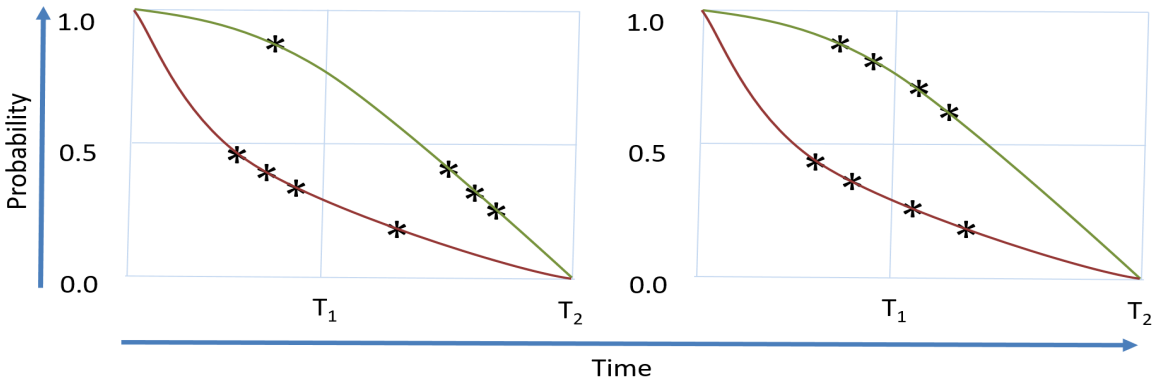


Figure 11: Simple examples to illustrate: [left] a model can have perfect 1-Calibration for a time, but not be D-Calibrated, and [right] a model can have perfect D-Calibration, but not be 1-Calibrated for a time. Each curve represents four apparently-identical patients, whose deaths are indicated by "*"s. (See Proposition B.6, in Appendix B.5.1, for details.)

	1-Calibration	D-Calibration
Objective	Evaluate Single Time Probabilities	Evaluate Entire Survival Curve
Values considered	$\{\ \hat{p}(\ t^* \,|\, \vec{x}_i)\ \}$	$\{\ \hat{p}(\ d_i \,|\, \vec{x}_i)\ \}$
Should match	Empirical number of deaths	Uniform
Statistical Test	Hosmer-Lemeshow test	Pearson's χ^2 test

Table 4: Summary of differences between 1-Calibration and D-Calibration.

4. Evaluating ISD Models

Sections 2.4 and 2.5 listed several distributional models (KM, and the ISDs: COX-KP, COXEN-KP, AFT, MTLR, and RSF-KM), and Section 3 provided 5 different evaluation measures: Concordance, L1-loss, 1-Calibration, Integrated Brier score, and D-Calibration. This section provides an empirical comparison of these 6 models, with respect to all 5 of these measures, over 8 data sets.

Of course, these 6 models do not include all possible survival models; they instead serve as a sample of the types of models available. The KM, COX-KP, and AFT model are all very common — these are standard approaches used throughout survival analysis and represent non-parametric, semi-parametric, and parametric models, respectively. As our preliminary studies with COX-KP suggested it was overfitting, we also included a regularized extension, using elastic net, COXEN-KP. Since Random Survival Forests (RSF) were introduced in 2008, they have had a large impact on the survival analysis community. However, as the Kaplan-Meier extension to transform RSF into an ISD is not well known, it is summarized in Appendix C.2. More recent still is the MTLR technique (Yu et al., 2011) that directly learns a survival distribution, by essentially learning the associated probability mass function (whose sequential right-to-left sum, when smoothed, is the survival distribution). We found some subsequent similar models, including "Multi-Task Learning for Survival Analysis" (MTLSA; Li et al., 2016), some deep learning variants (Ranganath et al., 2016; Katzman et al., 2018; Luck et al., 2017), and a computationally demanding Bayesian regression trees model (Sparapani et al., 2016), but for brevity, we focused on just the first such model, MTLR.[20]

Note the distribution class $\mathcal{D}$ chosen for AFT certainly influences its performance — *e.g.*, it is possible that AFT[Weibull] on a data set may fail D-Calibration whereas AFT[Log-Logistic] may pass; similarly for 1-Calibration at some time t^*, and the scores for Concordance, L1-loss and Integrated Brier score will depend on that distribution class. This paper will focus on AFT[Weibull] because, while still being parametric, the Weibull distribution is versatile enough to fit many data sets.

4.1. Datasets and Evaluation Methodology

There are many different survival data sets; here, we selected 8 publicly available medical data sets in order to cover a wide range of sample sizes, number of features, and proportions of censored patients. We excluded small data sets (with fewer than 150 instances) to reduce the variance in the evaluation metrics. Our data sets ranged from 170 to 2402 patients, from 12 to 7401 features, and percentage of censoring from 17.23% to 86.21%; see Table 5.

Note that we have not included extremely high-dimensional data (with tens of thousands of features, often found in genomic data sets), as such data raises additional challenges beyond the scope of standard survival analysis; see Witten and Tibshirani (2010) and Kumar and Greiner (2019) for methods to handle such extremely high-dimensional data.

The Northern Alberta Cancer Dataset (NACD), with 2402 patients and 53 features, is a conglomerate of many different cancer patients, including lung, colorectal, head and neck,

20. In addition, our empirical analysis in Haider's MSc thesis (Haider, 2019) found that DeepHit (Katzman et al., 2018) performs comparably to MTLR in many data sets, for these measures, but worse on the smaller data sets.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL
Number of patients: N	592	1105	950	2402	170	1095	295	240
% Censored	17.23	44.34	51.89	36.59	84.12	86.21	73.22	42.50
Maximum Follow-up time	3881 d	6432 d	83 m	84.3 m	3932 d	8605 d	18.34 y	21.8 y
# Features Originally: f_{raw}	12	13	53	53	18	61	4921	7401
# Features Post-Processing: f_{proc}	9	10	45	53	13	59	4921	7401
# Features Selected: f_{final}	6	10	34	46	8	28	2330	1771
f_{final} / N	0.010	0.009	0.036	0.019	0.047	0.026	7.898	7.379

Table 5: Overview of data sets used for empirical evaluations. From top to bottom: (1) the number of patients in each data set, (2) the percent of patients censored, (3) the "Maximum Follow-up time", (4) the number of features contained in the original data set, (5) the number of features after removal of features containing over 25% missing data or only 1 unique value, (6) the number of features after univariate Cox selection, and (7) the feature-to-sample_size ratio.

esophagal, stomach, and other cancers. In addition to using the complete NACD data set, we considered the subset of 950 patients with colorectal cancer (Nacd-Col) with the same 53 features.

Another four data sets were retrieved from data generated by The Cancer Genome Atlas (TCGA) Research Network (Genome Data Analysis Center, 2016): Glioblastoma multiforme (GBM; 592 patients, 12 features), Glioma (GLI; 1105 patients, 13 features), Rectum adenocarcinoma (READ; 170 patients, 18 features), and Breast invasive carcinoma (BRCA; 1095 patients, 61 features). To ensure a variety of feature/sample-size ratios, we consider only the clinical features in our experiments.

Lastly, we included two high-dimensional data sets: the Dutch Breast Cancer Dataset (DBCD; van Houwelingen et al., 2006) contains 4919 microarray gene expression levels for 295 women with breast cancer, and the Diffuse Large B-Cell Lymphoma (DLBCL; Li et al., 2016) data set contains 7401 features focusing on Lymphochip DNA microarrays for 240 biopsy samples.

We applied the following pre-processing steps to each data set: we first removed any feature that was missing over 25% of its values, as well as any features containing only 1 unique value. For the remaining features, we "one-hot encoded" each nominal feature and then passed each feature to a univariate Cox filter, and removed any feature that was not significant at the $p \leq 0.10$ level. Following feature selection, we replaced any missing value with the respective feature's mean value. (Note this feature selection was found to benefit all isd models across all performance metrics; data not shown.) Table 5 provides the data set statistics and a full breakdown of feature numbers in each step.

Following feature selection, the data was partitioned into 5 disjoint folds by first sorting the instances by time and censorship, then placing each censored (resp., uncensored) instance sequentially into the folds — meaning all folds had roughly the same distribution of times and censorships. The values of each feature were then normalized (transformed to zero mean with unit variance) within each fold.

For COXEN-KP, RSF-KM, and MTLR, we used an internal 5CV for hyper-parameter selection. There were no hyper-parameters to tune for the remaining models: COX, KM, and AFT.

As 1-Calibration required specific time points, and as models might perform well on some survival times but poorly on others, we chose five times to assess the calibration results of each model: the 10th, 25th, 50th, 75th, and 90th percentiles of survival times for each data set. Here, we used the D'Agostino-Nam translation to include censored patients for these evaluation results — see Appendix B.3. Appendix D.4 presents all 240 values (6 models $\times$ 8 data sets $\times$ 5 time-points); here we instead summarize the number of data sets that each model passed as 1-Calibrated (at $p \geq 0.05$) for each percentile.

For all evaluations, we report the averaged 5CV results for Concordance, Integrated Brier score, and L1-loss. As Concordance requires a risk score, we use the negative of the median survival time and similarly use the median survival time for predictions for the L1-loss. To adjust for presence of censored data, we used the L1-Margin loss, given in Appendix B.2, which extends the "Uncensored L1-loss" given in Section 3.2 (which considers only uncensored patients). Additionally, as 1-Calibration (resp., D-Calibration) results are reported as p-values, and it is not appropriate to average over the folds, so we combined the predicted survival curves from all cross-validation folds for a single evaluation, and report the resulting p-value.

Empirical evaluations were completed in R version 3.4.4. The implementations of KM, AFT, and COX-KP can all be found in the *survival* package (Therneau, 2015) whereas COXEN-KP uses the *cocktail* function found in the *fastcox* package (Yang and Zou, 2017). Both RSF and RSF-KM come from the *randomForestSRC* package (Ishwaran and Kogalur, 2018). An implementation of MTLR can be found in the *MTLR* package and all code used in this analysis is publicly available on the GitHub account[21] of the lead author.

4.2. Empirical Results

Below, we consider a data set to be "Nice" if its feature-to-sample-size ratio was less than 0.05 (for the final feature set) and its censoring was less than 55%; this includes four of the 8 data sets: GBM, Nacd-Col, GLI, NACD, which are shown first in all of our empirical studies. We let "High-Censor" data sets refer to READ and BRCA and "High-Dimensional" data sets refer to the other two (DLBCL and DBCD).

4.2.1. Concordance, Integrated Brier score, and L1-loss Results

Figures 12, 13 and 14 give the empirical results for Concordance, Integrated Brier score, and L1-Margin loss respectively, where each circle is the score of the associated model on the data set, and lines correspond to one standard deviation (over the 5 cross-validation folds). Appendix D provides the exact empirical results for these measures.

Best Performance: The blue circles represent the best performing models, for each data set; here we find that MTLR performs best on a majority of data sets: six of eight for Concordance and L1-loss, and seven of eight for the Integrated Brier score.

21. https://github.com/haiderstats/ISDEvaluation

Nice Data Sets: Recall that the first 4 data sets are Nice. Here, we find that most models performed comparably — in particular, aft and cox-kp perform nearly as well as the other, more complex, models. aft even performs best in terms of L1-loss on GBM. The only exception was rsf-km, which did much worse on GBM and GLI, in all three measures.

km was worse than the various isd-models for all 3 measures. (The only exception was rsf-km, which was worse on for the data sets GLI and GBM for Integrated Brier score, and for those data sets and also Nacd-Col for L1-loss.)

High-Censor data sets — READ and BRCA: Note first that the variance in the evaluation metrics is generally higher on READ for all models (except km) due to the small number of uncensored patients within each test fold — this is not present in BRCA due to the larger sample size (1095). Again we find that coxen-kpand mtlr are similar for all three measures, but rsf-km performs consistently worse across all three metrics for both READ and BRCA. aft and cox-kp are either comparable (or inferior) to the other three isd-models: Concordance: worse performance but within error-bars for READ and BRCA; Integrated Brier score: similar for both READ and BRCA; L1-loss: slightly worse for READ and BRCA. Additionally, aft and cox-kp tend to show higher variance in evaluation estimates on READ than other models for all three measures.

km is worse than all 5 isd-models for Concordance, but comparable to the best for Integrated Brier score and L1-loss (actually scoring better than cox-kp and aft for L1-loss on READ and beating cox-kp on BRCA).

High-Dimensional Data sets — DBCD and DLBCL: There are no entries for cox-kp for these two data sets as it failed to run on them, likely due to the large number of features. As aft is unregularized, it is not surprising that it does poorly across all measures for these high-dimensional data sets — indeed, even worse than km, which did not use any features! We see that the other three isd-models — coxen-kp, mtlr and rsf-km — perform similarly to one another here, and km also achieves similar results (ignoring Concordance where km always achieves 0.5, as it gives identical predictions for all patients).

4.2.2. 1-Calibration Results

Table 6 gives the number of data sets each model passed for 1-Calibration, for each time of interest. We see that mtlr is typically 1-Calibrated across the percentiles of survival times. Specifically, mtlr is 1-Calibrated for at a minimum of four of eight data sets for the 10th, 25th, 50th, and 90th percentiles, outperforming all other models considered. The 90th percentile appear to be the most challenging in general, as some models (aft, cox-kp, rsf-km) are not 1-Calibrated for any data sets, coxen-kp is 1-Calibrated for two, and mtlr is 1-Calibrated for four. The 75th percentile was also challenging; however aft, cox-kp, and rsf-km were 1-Calibrated for one, coxen-kp is 1-Calibrated for two, and mtlr is 1-Calibrated for three. The most challenging data sets for rsf-km once again were GBM, GLI, BRCA, and READ, for which rsf-km was 1-Calibrated only at the 10th percentile for READ – see Appendix D.4. Additional challenging data sets include the complete NACD and DBCD, which were challenging for all models. As km assigns an identical prediction for all patients, it cannot partition patients into different bins, meaning it cannot be evaluated by 1-Calibration.

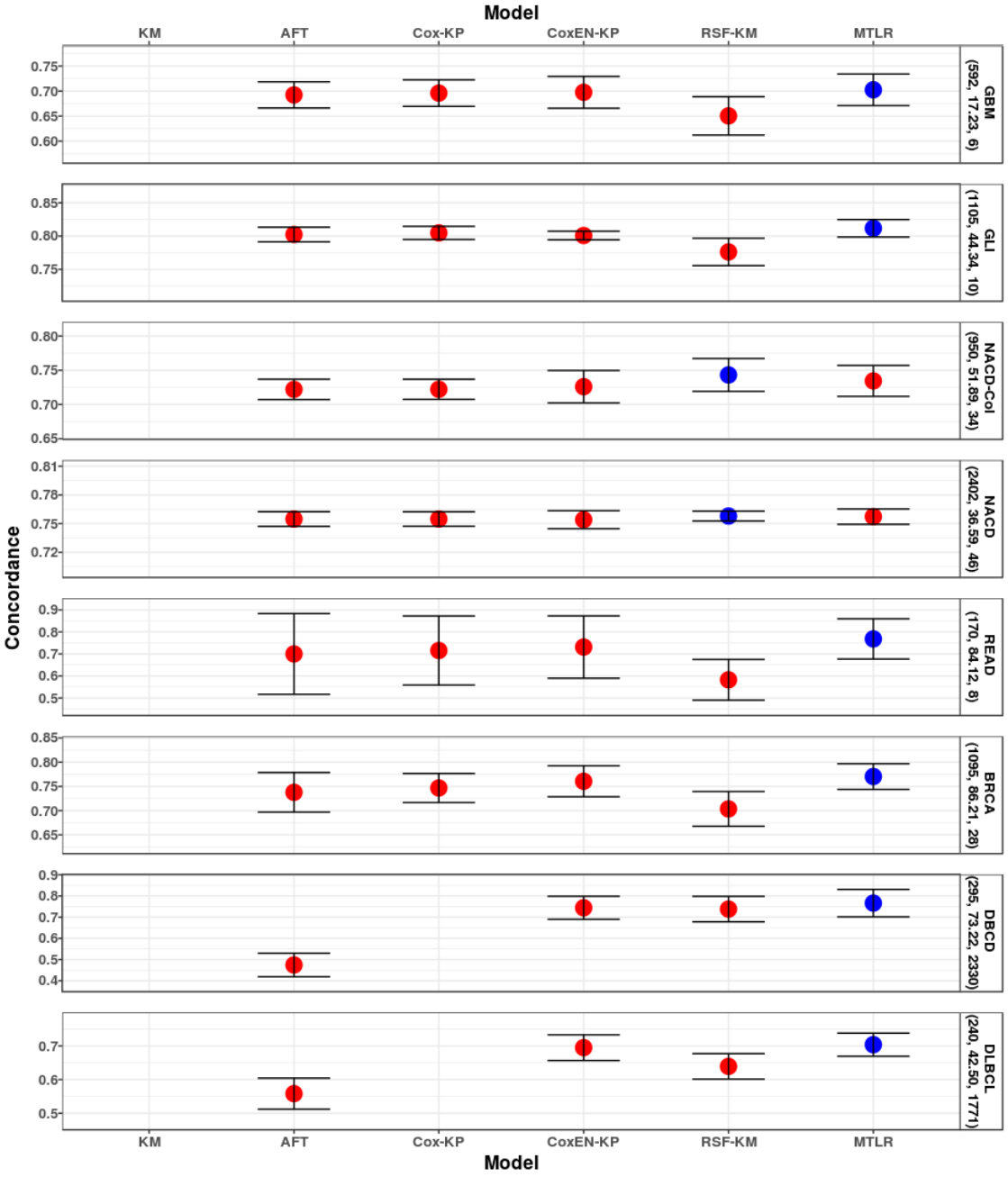


Figure 12: Concordance means and one standard deviation are given by circles and error bars, respectively. The best (highest Concordance) scoring model is given in blue; all other models are in red. We included KM so this figure would "line up" with Figures 13 and 14, but left the value blank, as the Concordance scores for KM are always 0.5. For these 3 figures: As COX-KP failed to run for data sets DBCD and DLBCL, those entries are blank. summarizes the 1-Calibration algorithm. The description at the right gives the name of the data set, and the 3 numbers "under" each data set name are (f_{final}, % Censored, N).

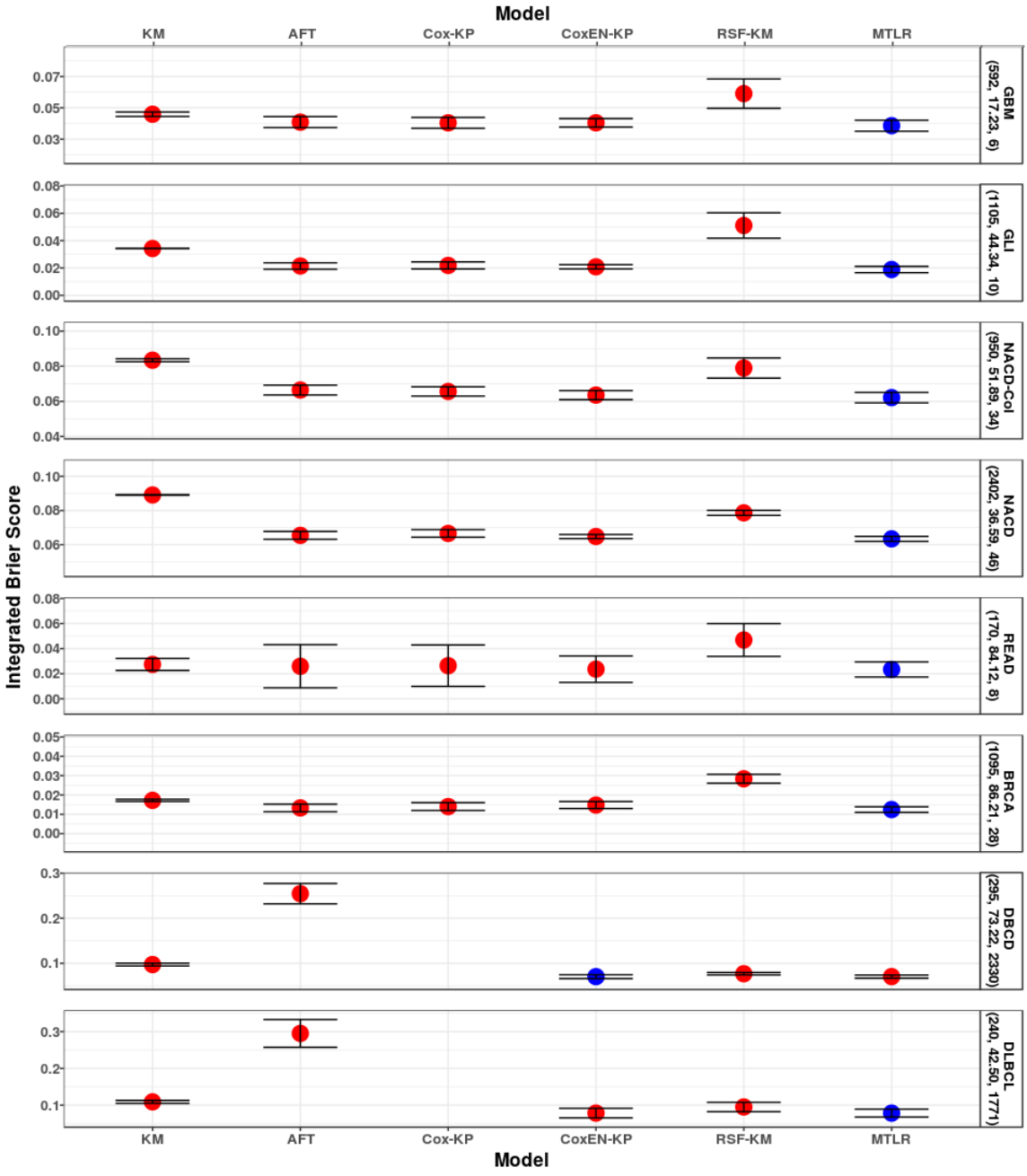


Figure 13: Integrated Brier score means and one standard deviation are given by circles and error bars, respectively. The best (lowest Integrated Brier score) scoring model is given in blue; all other models in red.

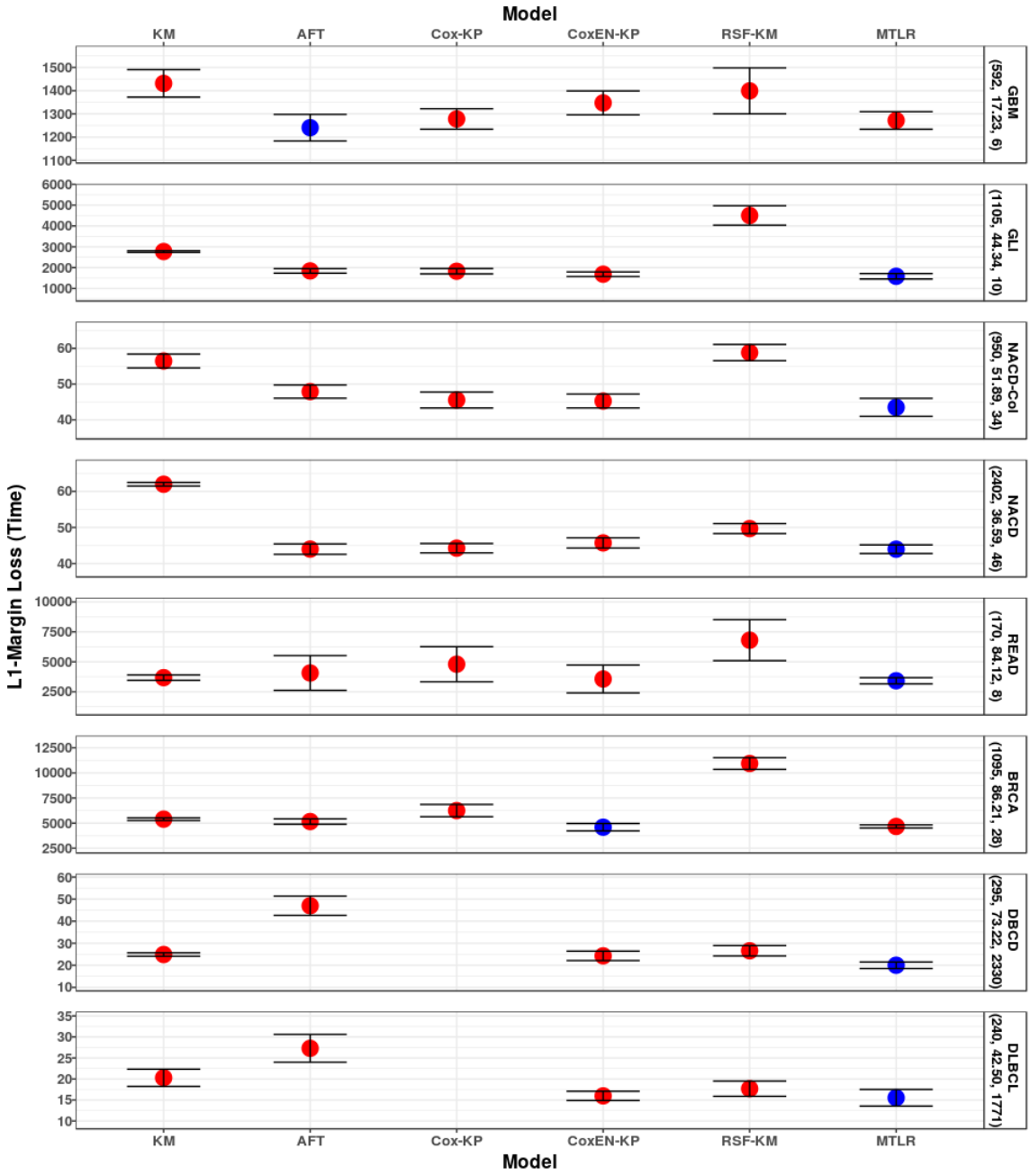


Figure 14: L1-loss means and one standard deviation are given by circles and error bars, respectively. The best (lowest L1-loss) scoring model is given in blue, all other models in red. As different data sets use different time units, we simply give the units of L1-loss as "Time" rather than days/months/years.

	10th	25th	50th	75th	90th
AFT	4	2	1	1	0
COX-KP	4	2	2	1	0
COXEN-KP	4	3	1	2	2
RSF-KM	4	2	2	1	0
MTLR	**6**	**8**	**6**	**3**	**4**

Table 6: Results from 1-Calibration evaluations. Columns represent percentiles used for each time point and rows indicate the model used. Recall there are 8 data sets — meaning no model performed perfectly for any of the percentiles.

	GBM	GLI	NACD-COL	NACD	READ	BRCA	DBCD	DLBCL	Total
KM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	8
AFT	0.000	0.017	**0.290**	0.000	**0.807**	**0.988**	0.000	0.000	3
COX-KP	0.046	0.049	**0.107**	0.000	**0.939**	**0.995**	-	-	3
COXEN-KP	**0.447**	**0.128**	**0.691**	0.000	**1.000**	**1.000**	**0.430**	**0.758**	7
RSF-KM	0.000	0.000	**0.403**	0.000	**0.974**	**0.757**	**0.911**	**0.574**	5
MTLR	**0.688**	**0.883**	**0.656**	**0.411**	**1.000**	**0.995**	**0.994**	**0.755**	8

Table 7: Results for D-Calibration evaluations. Columns correspond to the data set and rows to the model. Results are the p-value from the goodness-of-fit test. **Bold** values indicate that a model passed D-Calibration, *i.e.*, $p \geq 0.05$; and "-" means the algorithm did not return an answer.

4.2.3. D-Calibration Results

Table 7, which gives the D-Calibration p-values for each model and data set, shows that both KM and MTLR pass D-Calibration for every data set, with KM receiving the highest possible p-value, p =1.000, for each. (In fact, Lemma B.4 in Appendix B.5 proves that KM is asymptotically D-Calibrated.) While KM will tend to be D-Calibrated, it is also the *least* informative model, since it assigns all patients the same survival curve. MTLR is also D-Calibrated for all data sets, but in addition, it also provides each patient with his/her own survival curve.

Following KM and MTLR, COXEN-KP performed next best, only failing to be D-Calibrated for one data set: NACD. RSF-KM followed closely behind, being D-Calibrated for five of eight data sets, failing on GBM, GLI, and NACD. AFT performed similarly to COX-KP, each of which were D-Calibrated on three of eight data sets.

Figure 15 provides (sideways) histograms, to help visualize D-calibration. For each subfigure, each of the 10 horizontal bars should be 10%; we see a great deal of variance for the not-D-Calibrated COX-KP [left], a small (but acceptable) variability for the D-Calibrated MTLR [middle], and essentially perfect alignment for the D-Calibrated KM [right]. See also Figure 9.

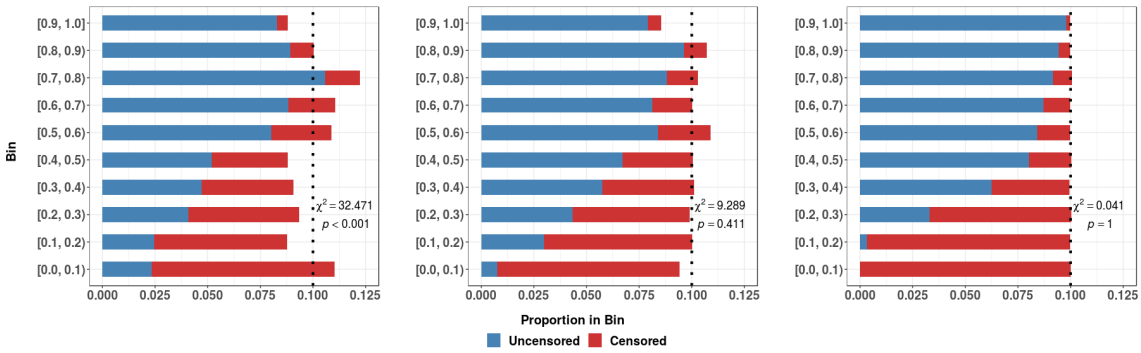


Figure 15: These figures show the (sideways) decile histogram used for the D-Calibration test. Each of these is run on the NACD data set; from left to right: running cox-kp, mtlr and km.

5. Discussion

Comparing different isd-models: Steyerberg et al. (2010) noted two different types of performance measures of a survival analysis model — calibration and discrimination — each of which can be assessed separately:

Calibration: "Of 100 patients with a risk prediction of x%, do close to x experience the event?"

Discrimination: "Do patients with higher risk predictions experience the event sooner than those who have lower risk predictions?"

Discrimination is a very important measure for some situations — *e.g.*, if we have 2 patients who each need a kidney transplant, but there is only a single kidney donor, then we want to know which patient will die faster *without* the transplant (Kamath and Kim, 2007). As discussed in Section 3.1, Concordance measures how well a predictor does, in terms of this discrimination task.

This paper, however, motivates and studies models that produce an individual survival curve for a specific patient. Such isd tools may not be optimal for maximizing discrimination (and therefore Concordance); even tools like cox and rsf, that were originally developed for discrimination, were then extended to produce these individual survival curves. Given this qualifier, we see (over the set of isd tools tested), mtlr scored best on Concordance for six of the eight data sets tested and rsf-km scored the best on the other two. (The relatively low performance of cox-kp is unexpected given the claim that "a method designed to maximize the Cox's partial likelihood also ends up (approximately) maximizing the [Concordance]" (Steck et al., 2008).) However, when we look at the Nice data sets, 4 of the 5 isd-models give nearly identical results (rsf-km differs by giving noticeably lower performance on GBM and GLI). These findings suggest that, for Nice data sets, more complex models (mtlr, rsf-km, and coxen-kp) do not offer large benefits in terms of

Concordance. For the High-Dimensional data sets, MTLR and COXEN-KP performed only marginally better than RSF-KM for DBCD but noticeably better than RSF-KM on DLBCL. Although these are only two data sets, this suggests that RSF-KM may not be optimal for these high-dimensional data sets, in terms of Concordance. For the High-Censor data sets, RSF-KM saw much worse performance for Concordance (among other metrics) suggesting RSF-KM may not be suitable for data sets with a high proportion of censored data.

As noted above, Concordance is only one measure for an ISD tool. Given that an ISD tool can produce a survival curve for each patient (and not just a single real-valued score), it can be used for various tasks, with various associated evaluations. For example, consider patients who are deciding whether to undergo an intensive medical procedure. Using the plots from Figure 8, note that Patient C has a very steep survival curve with a low median survival time, while Patient A has a shallow survival curve with a large median survival time. If we were to use this to predict the outcome of a procedure, we might expect Patient C to opt-out of the procedure, but Patient A to go through with it. Note the decision for Patient C is completely independent of Patient A, in that we could give the procedure to one, both, or neither of them. As these patients are not being ranked for a limited procedure, Concordance is not an appropriate metric; instead we need to evaluate such predictors using a calibration score — perhaps 1-Calibration or D-Calibration, as discussed in Sections 3.3 and 3.5.

As discussed in Section 3.3, 1-Calibration is particularly relevant for [P,1_{t^*},i] models — *i.e.*, models that produce a probability score for only 1 time point (for each patient). We also noted that ISD models, which produce individual survival curves, can also be evaluated using 1-Calibration, once the evaluator has identified the relevant specific time t^*. Here, we evaluated a variety of time points: the 10th, 25th, 50th, 75th, and 90th percentiles of survival times for each data set. We found MTLR to be superior to all the models considered here for all percentiles. The observation that MTLR was 1-Calibrated for a range of time points, across a large number of diverse data sets, suggests that the probabilities assigned by MTLR's survival curves are representative of the patients' true survival probabilities; the observation that the other models were not 1-Calibrated as often, calls into question their effectiveness here.

Of course, our analysis is performing the 1-Calibration test for 5 models (KM is excluded) across 8 data sets and 5 percentiles, meaning we are performing 200 statistical tests. We considered applying some p-value corrections — *e.g.*, the Bonferroni correction — to reduce the chance of "false-positives", which here would mean declaring a model that was truly calibrated as not. However, the actual p-values (see Appendix D.4) show that including these corrections would actually benefit MTLR the most, further strengthening the claim that MTLR has excellent 1-Calibration performance.

Our D-Calibration results further support the use of MTLR's individual survival curves over other ISD-models, by showing that MTLR was the only ISD-model to be D-Calibrated for all data sets. (Recall that KM is technically not an ISD since it gives one curve for all patients.) We see that different ISD-models are quite different for this measure — *e.g.*, AFT and COX-KP produce significantly worse performance for D-Calibration, being D-Calibrated for only three data sets. As discussed in Section 4.2, AFT is a completely parametric model, which means it cannot produce different shapes (see Figure 5[top-right]), likely impacting its ability to be D-Calibrated. (Our analysis showed only that AFT[Weibull] is not D-Calibrated

for these data sets; AFT[χ] for some other distribution class χ, might be D-Calibrated for more data sets.)

In addition to discussing discrimination (Concordance) and calibration (1-Calibration, D-Calibration) separately, we can also consider a hybrid evaluation metric — the Integrated Brier score — which measures a combination of both calibration and discrimination — see Section 3.4.We see MTLR performing the best for seven of the eight data sets, however, MTLR is no longer superior for DBCD, one of the high-dimensional data sets, even though it was superior for Concordance. Instead, COXEN-KP, RSF-KM, and MTLR all perform nearly identical for these High-Dimensional data sets.

The Integrated Brier scores, along with 1-Calibration and D-Calibration results, collectively show MTLR outperforms other models (for calibration), and is followed by COXEN-KP and RSF-KM. Specifically, COXEN-KP and RSF-KM are competitive to MTLR for High-Dimensional data sets — the 1-Calibration metric shows that both COXEN-KP and RSF-KM match the performance of MTLR for DLBCL: COXEN-KP and MTLR are 1-Calibrated across all percentiles and RSF-KM is 1-Calibrated across three of five, though p-values are very close to the 0.05 threshold for the other two. DBCD appeared to be the more challenging High-Dimensional data set — MTLR and COXEN-KP were 1-Calibrated for only two of five percentiles and RSF-KM was 1-Calibrated for one. This, coupled with the findings for Integrated Brier Score and D-Calibration, suggest that COXEN-KP, RSF-KM and MTLR are equally competitive for modeling individual patients' survival probabilities *when dealing with a large number of features*. However, this does not apply to smaller-dimensional data sets.

RSF-KM was not 1-Calibrated across any percentiles for GBM, GLI, or BRCA, and only 1-Calibrated at the 10th percentile for READ, and was not D-Calibrated for GBM and GLI. This, along with the poor performance of RSF-KM for all measures of GBM, GLI, READ, and BRCA suggests that RSF-KM does not produce effective individual survival curves for low-dimensional data sets. Other experiments (not shown) suggest that RSF-KM tends to overfit to the training set when given too few features. Additional meta-parameter tuning in these experiments was unable to correct for overfitting.

Given that survival prediction looks very similar to regression, it is tempting to evaluate such models using measures like L1-loss (which can lead to models like censored support vector regression (Shivaswamy et al., 2008)). A small L1-loss shows that a model can help with many important tasks, such as decisions about hospice, and for deciding about various treatments, based on their predicted survival times. However, simply because a model has the best performance for L1-loss does not mean the estimates are useful — consider the complete NACD data set, where the best performing model, MTLR, still had an average L1-loss of 43.97 months. While this is the lowest average error, predicting the time of death with an expected error of 43.97 months is likely not helpful to a patient, especially as the maximum follow-up time was 84.3 months. Indeed, for many of the data sets, many of the best models had L1-Losses that were *half* of the maximum follow-up; this disappointing performance is consistent with previous findings (Henderson and Keiding, 2005).

While the best model may not represent a "good" model, our empirical results still showed MTLR had the lowest L1-loss on six of eight data sets, although all ISD models performed comparably for the four Nice data sets (with the exception of RSF-KM). We see that KM is also competitive for the High-Censor data sets, but given the construction of the L1-

Characteristic of Dataset		Applicable Datasets	Evaluation	
%Censored	#Dimensionality	Name	Calibration	Discrimination
Low	Low	GBM, GLI, Nacd-Col, NACD	mtlr	cox-kp/aft
High	Low	READ, BRCA	mtlr/coxen-kp	mtlr/coxen-kp
Low	High	DLBCL	mtlr/coxen-kp/rsf-km	mtlr/coxen-kp
High	High	DBCD	mtlr/coxen-kp/rsf-km	mtlr/coxen-kp/rsf-km

Table 8: Our recommendation for isd models, for different types of data sets. The first row [Low, Low] corresponds to the Nice data sets. We also have divided the High-Dimensional set into Low versus High censoring. (DBCD is 73.22% censored.)

Margin loss, this is not surprising; see Appendix B.2. Moreover, the three complex models (coxen-kp, rsf-km, mtlr) appear comparable for the High-Dimensional data sets.

Which isd-Model to Use?: As shown above, which isd-model works best depends on properties of the data set, and on what we mean by "best". Table 8 summarizes our results here.

In general, for Nice data sets, mtlr was superior for calibration but for discrimination, all isd-models were equivalent, leading us to recommend using the simplest models: cox-kp, aft. As we found that rsf-km would overfit to the training data when the number of features was small (here, less than 34), we recommend avoiding rsf-km when there are so few features.

For High-Censor data sets, we recommend mtlr or coxen-kp when there are not many features (*e.g.*, READ, BRCA) for both calibration and discrimination. Typically cox-kp and aft had poor performance and high variability for High-Censor data sets. For High-Dimensional data sets with low censoring (less than 70%; *e.g.*, DLBCL), mtlr, coxen-kp, and rsf-km had the best performance for calibration. For discrimination, rsf-km seemed slightly worse for Concordance and Brier score, suggesting it may be a weaker model.

To explore whether these findings hold in general, we examined 33 other public data sets — 16 (Low Dimension, Low Censoring), 12 (Low Dimension, High Censoring), 4 (High Dimension, Low Censoring) and 1 (High Dimension, High Censoring) where High Censoring is $\geq$ 70%. Note that all Low Dimensional data sets were taken from the TCGA website whereas the other (High Dimensional) data sets arise from a variety of sources. The results from these 33 data sets are consistent with the findings reported here; specific results can be found on the lead author's RPubs site[22]. Given the low overall number of High-Dimensional data sets, these findings should be examined on further data sets.

Why use isd-Models?: As noted above, this paper considers only models that generate isds (*i.e.*, [P,∞,i]). This is significantly different from models that only generate risk scores ([R,$1_\forall$,i]), as those models can only be evaluated using a discriminatory metric. While this discrimination task (and hence evaluation) is helpful for some situations (*e.g.*, when deciding which patients should receive a limited resource), it is not helpful for others (*e.g.*, deciding whether a patient should go to a hospice, or terminate a treatment). A patient's

22. See http://rpubs.com/haiderstats/ISDEvaluationSupplement

primary focus will be on his/her own survival, not how they rank among others; hence the risk score such models produce do not meaningfully inform individual patients.

The single point probability models, [P,1_{t^*},i], are a step in the direction for benefiting patients, but they are still often inadequate as they apply only to a single time-point. While hospital administrators may want to know about specific time intervals (*e.g.*, t^* ="30-day readmission" probabilities), medical conditions seldom, if ever, are so precise. This is problematic as these probabilities can change dramatically over a short time interval — *i.e.*, whenever a survival curve has a very steep drop. For example, consider Patient #5 ($P5$) in Figure 5 for the MTLR model. Here, we would be optimistic about this patient if we considered the single point probability model at $t^* = 6$months, as $\hat{S}_{MTLR}(\, 6\text{months} \,|\, P5\,) = 0.8$, but very concerned if we instead used $t^* = 12$months, as $\hat{S}_{MTLR}(\, 12\text{months} \,|\, P5\,) = 0.3$. Note this trend holds for the other ISD-models shown, and also for many of the patients, including $P6,\ P7,\ P10$.

This suggests a model based on only a single time point may lead to inappropriate decisions for a patient. Note also that such a model might not even provide consistent relative rankings over a pair of patients — *i.e.*, it might provide different discriminative conclusions. Consider patients $P2$ and $P9$ in Figure 5[MTLR]. Here, at $t^* = 20$months, we would conclude that the purple $P9$ is doing worse (and so should get the available liver), but at $t^* = 30$months, that the orange $P2$ is more needy. We see similar inversions for a few other pairs of patients in MTLR, and also for several pairs in the RSF model.

Of course, one could argue that we just need to use multiple single-time models. Even here, we would need to *a priori* specific the set of time points — should we use 6 months and 12 months, and perhaps also 30 months? And maybe 20 months?

This becomes a non-issue if we use individual survival distribution (ISD; [P,∞,i]) models, which produce an entire survival curve, specifying probability values for every future time point. Moreover, while risk score models can only be evaluated using a discrimination metric, these ISD models can be evaluated using all metrics, making them an overall more versatile method for survival analysis.

Bottom line: In general, a survival task is based on both a data set, and an objective, corresponding to the associated evaluation measure. Our ISD framework is an all-around more flexible approach, as it can be evaluated using any of the 5 measures discussed here (Section 3) — both commonly-used and alternative. Importantly, when evaluating ISD models discriminatively (using Concordance), the risk scores we advocate (mean/median survival time) have meaning to clinicians and patients, whereas a general risk score, in isolation, has no clinical relevance. Moreover, the resulting survival curves are easy to visualize, which adds further appeal.

6. Conclusion

Future Work: This paper has focused on the most common situation for survival analysis: where (1) all instances in the training data are described using a fixed number of features (see the matrix in Figure 3), (2) there are no missing values, and (3) each instance either has a specified time of death, or is right-censored — *i.e.*, we have a lower bound on that patient's time of death. There are many techniques for addressing the first two issues — such as ways to "encode" a time series of EMRs (Electronic Medical Records) as a fixed number

of features, or using mean imputations. There are also relatively easy extensions to some of the models (*e.g.*, MTLR) to address the third point: handle left-censored instances (where the data set specifies an upper-bound on the patient's time of death), or interval-censored. These extensions, however, are beyond the scope of the current paper.

Contributions: This paper has surveyed several different approaches to survival analysis, including assigning individualized risk scores [R,$1_\forall$,i], assigning individualized survival probabilities for a single time point [P,1_{t^*},i], modeling a population-level survival distribution [P,∞,g], and computing individual survival distributions (ISD) [P,∞,i]. We discussed the advantages of computing an individual survival distribution for each patient, as this can help patients and clinicians make informed decisions about treatments, lifestyle changes, and end-of-life care. We discussed how ISD models can be used to compute Concordance measures for discrimination and L1-loss, but noted they should primarily be evaluated using calibration metrics (Sections 3.3 and 3.5), as these measure how well the individual survival curves represent the "true" survival of patients.

Next, we identified various types of ISD-models, and empirically evaluated them over a wide range of survival data sets — over a range of #features, #instances and %censoring. This analysis showed that MTLR was typically superior for the L1-loss, Integrated Brier score, and Concordance, but most importantly, showed it outperformed or matched all other models for the calibration metrics.

We also provide formal analyses related to D-calibration, proving that the true survival distribution is D-calibrated (Theorem B.3), that the Kaplan-Meier distribution is asymptotically D-calibrated (Lemma B.4) and that D-calibration is fundamentally different from IBS (Proposition B.5) and from 1-Calibration (Proposition B.6). To be self-contained, this paper provides high-level descriptions of the survival analysis algorithms, and also of the evaluation measures. (Note that the code base is also available, from `https://github.com/haiderstats/ISDEvaluation`.)

In conclusion, this paper explains why we encourage researchers and practitioners to use ISD-models (and especially ones similar to MTLR) to produce meaningful survival analysis tools, by showing how this can help patients and clinicians make informed healthcare decisions.

Acknowledgements

We gratefully acknowledge funding from NSERC (Discovery Grant), Amii, and Borealis AI, through an NSERC Engage Grant. We also thank Chun-Nam Yu, Ping Jin and Vickie Baracos for insights leading to this investigation, and Adam Kashlak for his insightful discussions regarding D-Calibration.

Appendix A. Extending Survival Curves to 0

In practice, survival curves often stop at a non-zero probability — see Figure 5 and Figure 16[left] below. This is problematic as it means they do not correspond to complete distribution (recall a survival curve should be "1−CDF(t)", where CDF is the Cumulative Distribution Function) which leads to problems for many of the metrics, as it is not clear how to compute the mean or the median value of the distribution. One approach is to extend each of the curves, horizontally, to some arbitrary time and then drop each to zero (the degenerate case being dropping the survival probability to zero at the last observed time point). This approach has downsides: Dropping the curve to zero at the last observed time point produces curves whose mean survival times are actually a lower bound on the patient's mean survival time, which is likely too small. In the event that the last survival probability is above 0.5 (as is often the case for highly censored data sets) this may bias our estimate of the L1-loss, which is based on the median value. Alternatively, if we instead extend each curve to some arbitrary time and then drop the curve to zero, we need to decide on that extension, which also could bias the L1-loss.

Since both standard approaches have clear downsides (and there is no way of knowing how the survival curves act beyond the sampled survival times), we chose to simply extrapolate survival curves using a simple linear fit: for each patient $\vec{x}_i$, draw a line from $(0, 1)$ — *i.e.*, time is zero and survival probability is 1 — to the last calculated survival probability, $(t_{max}, \hat{S}(\, t_{max} \,|\, \vec{x}_i \,))$, then extend this line to the time for which survival probability equals 0 — *i.e.*, $(t^0(\vec{x}_i), 0)$ — see Figure 16[right]. Note that curves cannot cross within the extended interval, which means this extension will not change the discriminatory criteria.

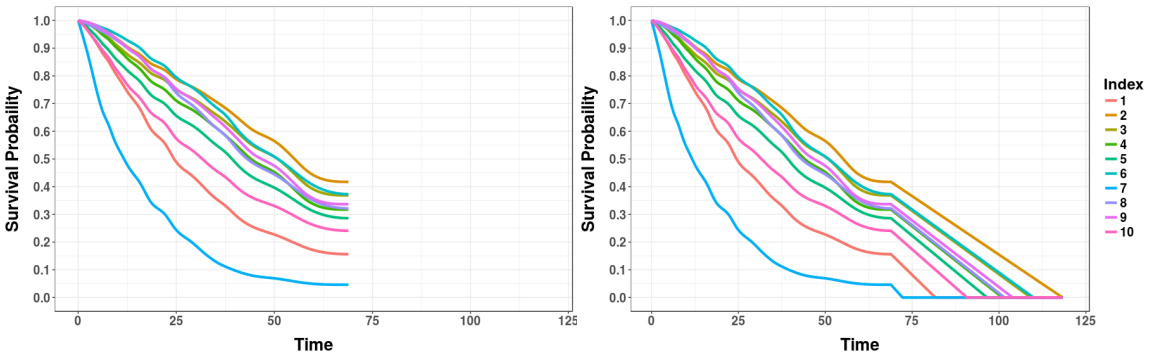


Figure 16: On left, survival curves generated from MTLR for the NACD-COL data set. Left shows this model's survival curves end at 68.9 months. On right, linear extensions of those survival curves go as far as 118 months.

There are extreme cases where a survival model will predict a survival curve with survival probabilities of 1 (up to machine precision) for all survival times (think "a horizontal line, at $p = 1$") — this occurred for unregularized models on high-dimensional data sets. In these cases, this linear extrapolation will never reach 0. To address this, we fit the Kaplan-Meier curve with the linear extension described above to compute t^0_{KM}; we then replace any

infinite prediction with this value. Additionally, as the Kaplan-Meier curve is to represent the survival curve on a *population* level, we also truncated any patient's median survival time by t^0_{KM}.

Appendix B. Evaluation Measures Supplementary Information

This appendix provides additional information about the various evaluation measures.

B.1. Concordance

As discussed in Section 3.1, Concordance is designed to measure the discriminative ability of a model. This is challenging for censored data. For example, suppose we have two patients who were censored at t_1 and t_2, respectively. Since both patients were censored, there is no way of knowing which patient died first, and hence, the risk scores for these patients are incomparable. However, if one patient's censored time is later than the death time of another patient, we do know the true survival order of this pair: the second patient died before the first.

To be precise, we first need to define the set of *comparable pairs* $\mathrm{CP}(V)$, which is the subset of pairs of indices (here using the validation data set V and recalling that $\delta = 1$ indicates a patient who died (uncensored)) containing all pair of instances when we know which patient died first:

$$\mathrm{CP}(V) \quad = \quad \{\, [i,j] \;\in\; V \times V \mid t_i < t_j \text{ and } \delta_i = 1 \,\} \tag{13}$$

Notice when the earlier event is uncensored (a death), we know the ordering of the deaths (whether the second time is censored or not); see Figure 17. The $t_i < t_j$ condition is to prevent double-counting such that $|\mathrm{CP}(V)| \leq \binom{|V|}{2}$.

We then consider how many of the possible pairs our predictor put in the correct order: That is, of all $[i,j]$ pairs in $\mathrm{CP}(V)$, we want to know how often $r(\vec{x}_i) > r(\vec{x}_j)$ given that $t_i < t_j$. Hence, the Concordance score of V, with respect to the risk scores, $r(\cdot)$, is

$$\hat{C}(V,\, r(\cdot)) \quad = \quad \frac{1}{|\mathrm{CP}(V)|} \sum_{i:\delta_i=1} \sum_{j:\, t_i<t_j} \mathcal{I}\,[\, r(\vec{x}_i) > r(\vec{x}_j)\,]\,. \tag{14}$$

One issue is how to handle ties, in either risk scores or death times – *i.e.*, for two patients, Patient A and Patient B, consider either $r(\vec{x}_A) = r(\vec{x}_B)$ or $d_A = d_B$. The two standard approaches are (1) to give the model a score of 0.5 for ties (of either risk scores or death times), or (2) to remove tied pairs entirely (Yan and Greene, 2008). The first option is equivalent to Kendall's tau, while the second leads to the Goodman-Kruskal gamma. The empirical evaluations (given in Section 4.2) use the first, as this gives Kaplan-Meier a Concordance index of 0.5 for all models. If we use the second option (excluding ties), then the Concordance for the Kaplan-Meier model is not well-defined.

B.2. L1-loss, and variants

As discussed in Section 3.2, survival analysis can be viewed as a regression problem that is attempting to minimize the difference between an estimated time of death and the true

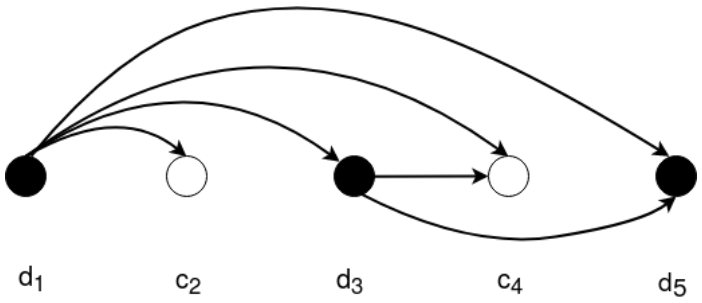


Figure 17: Depiction of Concordance comparisons, including censored patients. Black and white circles indicate uncensored and censored patients, respectively. Each d_i is the death time for an uncensored patient, and each c_j is the censoring time for a censored patient. We can only compare: uncensored patients who died *prior* to a censored patient's censoring time, or an uncensored patient's death time. Here, time increases as we go left-to-right; hence $d_1 < c_2 < d_3 < c_4 < d_5$. Here, we can compare 6 of the $\binom{5}{2} = 10$ pairs of patients. Figure adapted from (Wang et al., 2017).

time of death. However, typical regression problems require having precise target values for each instance; here, many instances are censored — *i.e.*, providing only lower bounds for the target values. One option is to simply remove all the censored patients and use the L1-loss given by Equation 4 (which we call "Uncensored L1-Loss"); however, this will likely bias the true loss.

One way to incorporate censoring is to use the Hinge loss for censored patients, which assigns 0 loss to any patient whose censoring time c_k is prior to the estimated median survival time, $\hat{t}_k^{(0.5)}$ — *i.e.*, a loss of 0 if $c_k < \hat{t}_k^{(0.5)}$ — and a loss of $c_k - \hat{t}_k^{(0.5)}$ if the censoring time is greater than $\hat{t}_k^{(0.5)}$. That is:

$$L1_{hinge}(\, V,\, \{\hat{t}_j^{(0.5)}\}_j\,) \;=\; \frac{1}{|V|}\left[\sum_{j\in V_U} |d_j - \hat{t}_j^{(0.5)}| \;+\; \sum_{k\in V_C} [c_k - \hat{t}_k^{(0.5)}]_+\right]. \qquad (15)$$

where V_U is the subset of the validation data set that is uncensored, and V_C is the censored subset, and $[a]_+$ is the positive part of a, *i.e.*,

$$[a]_+ \;=\; \max\{a, 0\} \;=\; \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

This formulation is an optimistic lower bound on the L1-loss for two reasons: (1) it gives a loss of 0 if the censoring occurs prior to the estimated survival time, implying that $d_k = \hat{t}_k^{(0.5)}$, and (2) it gives a loss of $c_k - \hat{t}_k^{(0.5)}$ if the censoring time occurs after the estimated survival time, which assumes that $d_k = c_k$. Both are the best possible values for the unknown d_k, given the constraints.

One weakness of the L1-Hinge loss is that if a model predicts very large survival times for all patients (both censored and observed), the hinge loss will give 0 loss for the censored

patients; in data sets with a large proportion of censored patients, this leads to an optimistic score overall. Thus the hinge loss will favor models that tend to largely overestimate survival times as opposed to those models underestimating survival time.

A third variant of L1-loss, the *L1-Margin loss*, assigns a "Best-Guess" value to the death time corresponding to c_k, which is the patient's conditional expected survival time given they have survived up to c_k — given by

$$BG(c_k) \quad = \quad c_k \; + \; \frac{\int_{c_k}^{\infty} S(t)\, dt}{S(c_k)} \tag{16}$$

where $S(\cdot)$ is the survival function; Theorem B.1 proves this value corresponds to the conditional expectation. In practice we use Kaplan-Meier estimate, $\hat{S}_{KM}(\,\cdot\,)$, generated from the training data set (disjoint from the validation data set) as our estimate of $S(\cdot)$ in Equation 16.

We also realized that these $BG(c_k)$ estimates are more accurate for some patients than for others. If $c_k \approx 0$ — that is, if the patient was censored near the beginning time — then we know very little about the true timing of when the death occurred, so the estimate $BG(c_k)$ is quite vague, which suggests we should give very little weight to the associated loss, $|BG(c_k) - \hat{t}_k^{(0.5)}|$. Letting α_k be the weight associated with these terms, we would like $\alpha_k \approx 0$. On the other hand, if c_r is large — towards the longest survival time observed (call it d_{max}) — then there is a relatively narrow gap of time where this $\vec{x}_r$ could have died (probably within the small interval (c_r, d_{max})); here, we should give a large weight to loss associated with this estimate.

This motivates us to define

$$L1_{margin}(V, \hat{t}_j^{(0.5)}) = \frac{1}{|V_U| + \sum_{k \in V_C} \alpha_k} \left[\sum_{j \in V_U} |d_j - \hat{t}_j^{(0.5)}| \; + \; \sum_{k \in V_C} \alpha_k |BG(c_k) - \hat{t}_k^{(0.5)}| \right] \tag{17}$$

where α_k reflects the confidence in each Best-Guess estimate. To implement this, we set $\alpha_k = 1 - \hat{S}_{KM}(c_k)$, which gives little weight to instances with early censor times but considers late censor times to be almost equivalent to an observed death time. Note this is the version of L1-loss we presented in Figure 14, with details in Table 12.

For completeness, we prove Equation 16. (This claim is also proven by Gupta and Bradley (2004), which uses *mean residual life* rather than *expected total life*.)

Theorem B.1. *The conditional expectation of time of death, D, given that a patient was censored at time c, is given by: $E[D \,|\, D > c] \;=\; c + \frac{\int_c^{\infty} S(x)\, dx}{S(c)}$.*

Proof. Let D be the r.v. for the time when a patient dies, and define

$$S(c) \quad = \quad P(D > c) \quad = \quad \int_c^{\infty} P(D = t)\, dt$$

as the survival function — *i.e.*, the probability that the patient dies after time c. Given this, the conditional probability is

$$\begin{aligned} P(\,D = t \,|\, D > c\,) &= \frac{P(\,D = t,\ D > c\,)}{P(\,D > c\,)} \\ &= \frac{P(\,D = t,\ D > c\,)}{S(\,c\,)} \\ &= \begin{cases} 0 & \text{if } t < c \\ \frac{P(\,D=t\,)}{S(\,c\,)} & \text{otherwise} \end{cases}. \end{aligned}$$

Hence,

$$\begin{aligned} E[\,D \,|\, D > c\,] &= \int_0^c t \times 0\, dt \quad + \quad \int_c^\infty t\, \frac{P(\,D = t\,)}{S(\,c\,)}\, dt \\ &= \frac{1}{S(\,c\,)} \left[\int_c^\infty c\ P(\,D = t\,)\, dt \quad + \quad \int_c^\infty (t - c)\, P(\,D = t\,)\, dt \right] \\ &= \frac{1}{S(\,c\,)} \left[c\, S(\,c\,) \quad + \quad \int_c^\infty \left(\int_c^t dx \right) P(D = t)\, dt \right] \\ &= c \ + \quad \frac{1}{S(\,c\,)} \left[\int_c^\infty \left(\int_x^\infty P(D = t)\, dt \right) dx \right] \qquad (18) \\ &= c \ + \quad \frac{\int_c^\infty S(\,x\,)\, dx}{S(\,c\,)}. \end{aligned}$$

□

Step 18 is an application of Tonelli's theorem (Saks, 1937), which lets us swap the order of integration for a non-negative function. As desired, this quantity, $E[\,D \,|\, D > c\,]$, is always at least c. Moreover, when $c = 0$, this is

$$0 \ + \ \frac{\int_0^\infty S(\,t\,)\, dt}{1} \quad = \quad \int_0^\infty S(\,t\,)\, dt \quad = \quad E[\,D\,]$$

which is the expected value of the distribution for this survival curve (and exactly the claim of the Theorem).

A further discussion of L1-Loss (and variants such as log L1-Loss) is available in Haider's MSc thesis (Haider, 2019).

B.3. 1-Calibration

To demonstrate the description from Section 3.3, consider the following example: If there are $n = 50$ patients, then $50/10 = 5$ will be in each bin, and the first bin B_1 will contain the 5 with lowest predicted probability values, and the second bin B_2 will contain the next

smallest 5 values, and so forth — *e.g.*,

$$\begin{aligned} B_1 &= \{0.32,\ 0.34,\ 0.43,\ 0.43,\ 0.48\} \\ B_2 &= \{0.55,\ 0.56,\ 0.61,\ 0.61,\ 0.72\} \\ &\vdots \\ B_{10} &= \{0.85,\ 0.85,\ 0.86,\ 0.87,\ 0.87\} \end{aligned}$$

Now consider the 5 patients who belong to B_1. As the average of their probabilities is $\frac{0.32+0.34+0.43+0.43+0.48}{5} = 0.4$, we should expect 40% of these 5 individuals to die in the next 5 years — that is, 2 should die. We can then compare this prediction ($0.40 \times 5 = 2$) with the actual number of these B_1 patients who died. We can similarly compare the number of B_2 patients who actually died to the number predicted (based on the average of these 5 probability values, which here is $0.61 \times 5 = 3.05$), and so forth.

In general, we say that the predictor is 1-Calibrated if these B predictions, for the $B = 10$ bins, are sufficiently close to the actual number of deaths with respect to these bins. Here, we use the Hosmer–Lemeshow statistical test (given in Section 3.3) to see if the observed results were significant, repeating Equation 5:

$$\widehat{HL}\,(\,V_U,\ \hat{S}(\,t^*\,|\,\cdot\,)\,) \quad = \quad \sum_{j=1}^{B} \frac{(O_j - n_j\,\bar{p_j})^2}{n_j\,\bar{p_j}\,(1-\bar{p_j})},$$

where O_j is the number of observed events, n_j is the number of patients, $\bar{p_j}$ is the average predicted probability, and subscript j refers to within the jth of B bins.

B.3.1. Incorporating Censoring into 1-Calibration

Survival data typically contains some amount of censoring, making the exact number of deaths for the jth bin, O_j, unobservable when the bin contains patients censored before t^*. That is, given a censored patient whose censoring time occurred before the time of interest ($c_i < t^*$), the patient may or may not have died by t^*. There are many standard techniques for incorporating censoring (Guffey, 2013); we use the D'Agostino-Nam translation (d'Agostino and Nam, 2003), which uses the *within bin* Kaplan-Meier curve in place of O_j. Specifically, the test statistic is given by,

$$\widehat{HL}_{DN}\,(\,V,\ \hat{S}(\,t^*\,|\,\cdot\,)\,) \quad = \quad \sum_{j=1}^{B} \frac{(\ n_j\ (1 - KM_j(t^*))\ -\ n_j\,\bar{p}_j)^2}{n_j\,\bar{p_j}\,(1-\bar{p_j})}, \tag{19}$$

where $KM_j(t^*)$ is the height of the Kaplan-Meier curve generated by the patients in the jth bin, evaluated at t^*. We use $1 - KM_j(t^*)$ as we are predicting the *number of deaths* and not $KM_j(t^*)$, which instead gives the probability of *survival* at t^*. Note also that $\widehat{HL}_{DN}$ follows a χ^2_{B-1} distribution, as opposed to the χ^2_{B-2} distribution for Equation 5. Algorithm 1 summarizes the 1-Calibration algorithm.

Algorithm 1: 1-Calibration Process (Demonstrated with $B = 10$ bins)

Input: ISD model θ that generates $\hat{S}_\theta(\, t \,|\, x\,)$, dataset $D = \{\ [x_i, e_i, \delta_i] \mid i = 1..n\ \}$, specific time t^*, P-value p^*

Result: *Success* or *Failure*

---

for $i = 1..n$ **do**

  $p_i \ \leftarrow\ 1 - S_\theta(\, t^* \,|\, x_i\,)$ `// predicted probability of event before` t^* `for instance with feature values` x_i

end

Set

  $B_1 \ \leftarrow\ \{\, k \mid p_k$ in the 10% of p_i values that are the lowest$\}$

  ... `// ties broken randomly`

  $B_{10} \ \leftarrow\ \{\, k \mid p_k$ in the 10% of p_i values that are the largest$\}$

for $j = 1..10$ **do**

  $\bar{p}_j \ \leftarrow\ \frac{1}{|B_j|} \sum_{i \in B_j} p_i$ `// average of the` p_i`'s within` B_j

  $KM_j \ \leftarrow\ KM(\{\ [x_i, e_i, \delta_i] \mid i \in B_j\ \})$ `// Kaplan Meier distribution based on the elements of` B_j

  $O_j \ \leftarrow\ 1 - KM_j(\, t^*\,)$ `// probability that` KM_j `assigns to event occurring by time` t^*

end

$h \leftarrow \sum_{j=1}^{10} \frac{n}{10} \frac{(O_j - \bar{p}_j)^2}{\bar{p}_j(1-\bar{p}_j)}$ `// HL statistic (Equation 19) based on the predicted counts` $\frac{n}{10} \times \{\bar{p}_1, ..., \bar{p}_{10}\}$ `vs observed` $\frac{n}{10} \times \{O_1, ..., O_{10}\}$

$p \leftarrow$ P-value corresponding to h from χ^2_9 dist'n `//` $B - 1$ `degrees of freedom`

if $p > p^*$ **then**

  return *Success*

else

  return *Failure*

end

B.4. Brier Score

This section supplements the description of the Brier score given in Section 3.4, discussing the incorporation of censoring into the Brier score.

In 1999, Graf et al. (1999) proposed a way to compute the Brier Score for censored data, by using *inverse probability of censoring weights* (IPCW), which requires estimating the censoring survival function, denoted as $\hat{G}(t)$ over time points t. We can estimate $\hat{G}(t)$ by the Kaplan-Meier curve of the *censoring distribution* — *i.e.*, swapping those who are censored with those who are not ($\delta_i^{Cens} = 1 - \delta_i$), and building the standard Kaplan-Meier model. Intuitively, this IPCW weighting counteracts the sparsity of later observations — if a patient dies early, there is a good chance that $d_i < c_i$, meaning the event is observed, but if the patient survives for a long time, it becomes more likely that $c_i < d_i$, meaning this patient will be censored. Gerds *et al.* (2008; 2006) formalizes and proves this intuition.

The censored version of the Brier score for a given time, t^*, is calculated as

$$BS_{t^*}\left(V,\ \hat{S}(t^*|\cdot)\right) = \frac{1}{|V|}\sum_{i=1}^{|V|}\left[\frac{\mathcal{I}\left[\,t_i \le t^*, \delta_i = 1\,\right]\left(0 - \hat{S}(t^*|\vec{x}_i)\right)^2}{\hat{G}(t_i)} + \frac{\mathcal{I}\left[\,t_i > t^*\,\right]\left(1 - \hat{S}(t^*|\vec{x}_i)\right)^2}{\hat{G}(t^*)}\right], \quad (20)$$

where $t_i = \min\{d_i, c_i\}$. The first part of Equation 20 considers only uncensored patients whereas the second part counts all patients whose event time is greater than t^*. The patients who were censored *prior* to t^* are not explicitly included, but contribute based on their influence in $\hat{G}(\cdot)$.

As $\hat{G}(t)$ is a decreasing step function of t, $\frac{1}{\hat{G}(t)}$ is increasing, which means that patients who survive longer than t^* have larger weights than patients that died earlier, since the longer surviving patients were more likely to become censored.

B.5. D-Calibration

We begin this section by (1) showing that, given only uncensored patients, the distribution of the true survival function, $S(t|x)\}_t$, should follow a uniform distribution, then (2) showing how to incorporate censored patients into the D-Calibration estimate, and then (3) proving that this combination of censored and uncensored patients should produce a uniform distribution, motivating a goodness-of-fit test. We then show that the Kaplan-Meier distribution is asymptotically D-calibrated, then prove two propositions (discussed earlier) that show that D-calibration is fundamentally different from two other measures: IBS and 1-Calibration.

For this analysis, we assume each patient $\vec{x}_i$ has a true survival function, $S(\,t\,|\,\vec{x}_i\,)$, which is the probability that this patient will die after time t. Assume each patient has a time of death, d_i and a censoring time, c_i, and $t_i = \min\{d_i, c_i\}$ is the observed event time. Given a validation set V, we first examine the case of all uncensored patients — *i.e.*, $t_i = d_i$ for $i = 1, \ldots, |V|$.

Lemma B.2. *The distribution of a patient's survival probability at the time of death $S(\,d_i\,|\,\vec{x}_i\,)$ is uniformly distributed on [0,1].*

Proof. The probability integral transform (Angus, 1994) states that, for any random continuous variable, X, with cumulative distribution function given by $F_x(\cdot)$, the random variable

$Y = F_x(X)$ will follow a uniform distribution on [0,1], denoted as $U(0,1)$. Thus, given randomly sampled event times, d_i, we have $F(\, d_i \,|\, \vec{x}_i \,) \sim U(0,1)$. As the survival function is simply $S(\, d_i \,|\, \vec{x}_i \,) = 1 - F(\, d_i \,|\, \vec{x}_i \,)$, its distribution is $1 - U(0,1)$, which also follows $U(0,1)$ and hence, $S(\, d_i \,|\, \vec{x}_i \,) \sim U(0,1)$. □

This Lemma shows that, given the true survival model, producing $S(\, \cdot \,|\, \vec{x}_i \,)$ curves, the distribution of $S(\, d_i \,|\, \vec{x}_i \,)$ should be uniform over event times. Thus if a learned model accurately learns the true survival function, $\hat{S}_\Theta(\, \cdot \,|\, \cdot \,) \approx S(\cdot|\cdot)$, we will expect the distribution across event times to be uniform. This is then tested using the goodness-of-fit test assuming each bin contains an equal proportions of patients.

Of course, conditions become more complicated when considering censored patients. Here, we employ the standard assumption that censoring time is independent of death time, $c_i \perp d_i \mid \vec{x}_i$ (Kaplan and Meier, 1958). Suppose we have a censored patient — *i.e.*, $t_i = c_i$ — such that $S(\, c_i \,|\, \vec{x}_i \,) = 0.25$. Since the censoring time is a lower bound on the true death time, we know that $S(\, d_i \,|\, \vec{x}_i \,) \leq 0.25$, since $c_i < d_i$ and survival functions monotonically decrease as event time increases. If we are using deciles, we would like to know the probability that the time of death occurred in the [0.2,0.3) bucket — *i.e.*, $P(\, S(\, d_i \,|\, \vec{x}_i \,) \in [0.2, 0.3) \;\mid\; S(\, d_i \,|\, \vec{x}_i \,) \leq 0.25 \,)$. Using the rules of conditional probability, this is computationally straightforward[23]:

$$\begin{aligned}
P(\, S(d_i) \in [0.2, 0.3) \,|\, S(d_i) < 0.25 \,) &= \frac{P(\, S(d_i) \in [0.2, 0.3),\, S(d_i) < 0.25 \,)}{P(\, S(d_i) < 0.25 \,)} \\
&= \frac{P(\, S(d_i) \in [0.2, 0.25)) \,)}{P(\, S(d_i) < 0.25 \,)} \\
&= \frac{0.05}{0.25} \qquad \text{(as } S(\cdot) \sim U(0,1)\text{)} \\
&= 0.2
\end{aligned}$$

Similarly, we can use the same logic as above to compute these probabilities for the other two buckets, $[0.1, 0.2)$ and $[0.0, 0.1)$:

$$\begin{aligned}
P(\, S(d_i) \in [0.1, 0.2) \;\mid\; S(d_i) < 0.25 \,) &= \frac{P(\; S(d_i) \in [0.1, 0.2),\, S(d_i) < 0.25 \;)}{P(\; S(d_i) < 0.25 \;)} \\
&= \frac{P(\; S(d_i) \in [0.1, 0.2) \;)}{P(\; S(d_i) < 0.25 \;)} \\
&= \frac{0.1}{0.25} \qquad \text{(as } S(\cdot) \sim U(0,1)\text{)} \\
&= 0.4
\end{aligned}$$

and similarly for the $[0.0, 0.1)$ bucket. Note that these probabilities sum to one, $(0.2 + 0.4 + 0.4) = 1$, as desired.

23. To simplify notation, we drop the conditioning on $\vec{x}_i$ of $S(\cdot|\cdot)$.

This example motivates the following procedure to incorporate censored patients into the D-Calibration process: Given B buckets that equally divide [0,1] into intervals of width $W = 1/B$, suppose a patient is censored at time c with associated survival probability $S(c)$. Let ℓ_1 be the infimum probability of the bucket that contains $S(c)$ — *e.g.*, 0.2 for the example above where $S(c_i) = 0.25 \in [0.2, 0.3)$. Then we assign the following weights to buckets:

(α) Bucket $[\ell_1, \ell_2)$ (which contains $S(c)$): $\frac{S(c)-\ell_1}{S(c)} = 1 - \frac{\ell_1}{S(c)}$

(β) All following buckets (*i.e.*, the buckets whose survival probabilities are all less than ℓ_1): $\frac{W}{S(c)} = \frac{1}{B \cdot S(c)}$,

Algorithm 2 summarizes the D-calibration process.

Note this formulation follows directly from the example above. This weight assignment effectively "blurs" censored patients across the buckets following the bucket where the patient's learned survival curve, $\hat{S}_\Theta(\, c_i \,|\, \vec{x}_i \,)$, placed the censored patient.

To further illustrate this concept of blurring a patient across buckets, consider a patient who is censored at $t = 0$ with $S(c_i) = 1$. This patient is then blurred across all ($B = 10$) buckets, adding a weight of 0.1 to all 10 buckets. Alternatively, if a patient is censored very late, with $S(c_i) \leq 0.1$ then the patient is not blurred at all — only a weight of 1 is added to the last bucket.[24]

To perform the goodness-of-fit test, we must first calculate the observed proportion of patients within each bucket. Let b_k represent the observed proportion of patients within the interval $[p_k, p_{k+1})$ — *e.g.*, $[p_k, p_{k+1}) = [0.2, 0.3)$ in the example above. There are six categories of patients; see Table 9. We focus on the 3 cases that contribute a non-0 amount to the value of b_k. We can formally calculate:

$$
\begin{aligned}
b_k \quad = \quad & \frac{1}{|V|} \sum_{i=1}^{|V|} \Bigg[\qquad 1 \qquad \cdot \; \mathcal{I}\left[\, S(d_i) \in [p_k, p_{k+1}) \;\wedge\; d_i \leq c_i \,\right] && (21) \\
& + \quad \frac{S(c_i) - p_k}{S(c_i)} \cdot \; \mathcal{I}\left[\, S(c_i) \in [p_k, p_{k+1}) \;\wedge\; c_i < d_i \,\right] && (22) \\
& + \quad \frac{(p_{k+1} - p_k)}{S(c_i)} \cdot \; \mathcal{I}\left[\, S(c_i) \geq p_{k+1} \qquad \wedge \; c_i < d_i \,\right] \Bigg]. && (23)
\end{aligned}
$$

Above, (21) refers to the weight that the patients with observed events contribute to the k^{th} bucket — *i.e.*, each uncensored patient whose survival probability at time of death lands in $[p_k, p_{k+1})$ contributes a value of 1. Here, we consider $d_i = c_i$ to be an uncensored event. Next, (22) gives the weight from the censored patients whose survival probability at time of censoring is within the k^{th} bucket (item (α) above). Lastly, (23) gives the weights from

24. This identifies a weakness of D-Calibration: if a validation set contains N_0 patients censored at time 0, then all buckets are given an equal weight of N_0/B; if N_0 is large relative to the total number of patients, then the buckets may appear uniform, no matter how the other patients are distributed, which means any model based on such heavily "time 0 censored" data would be considered to be D-Calibrated.

Algorithm 2: D-Calibration Process (demonstrated with $B = 10$ buckets)

Input: ISD model θ that generates $\hat{S}_\theta(\, t \,|\, x\,)$, dataset $D = \{\ [x_i, e_i, \delta_i] \mid i = 1..n\ \}$, P-value p^*

Result: *Success* or *Failure*

---

for $j = 1..10$ **do**
  $b_j \leftarrow 0$ `// initialize buckets`
end

for $i = 1..n$ **do**
  $s_i \leftarrow \max\{\, \hat{S}_\theta(\, e_i \,|\, x_i\,),\ 10^{-5}\}$ `// predicted probability that instance with features` x_i `survives beyond event time` e_i
  $j^* \leftarrow \lceil 10 \times s_i \rceil$ `// index of bucket associated with` s_i
  if $\delta_i = 1$ **then** `// instance is uncensored ⇒ credit bucket` b_{j^*}
    $b_{j^*} \leftarrow b_{j^*} + 1$
  else `// instance is censored ⇒ spread credit among possible buckets`
    for $k = 1\,..\,(j^*-1)$ **do**
      $b_k \leftarrow b_k + \frac{1}{10 s_i}$
    end
    $b_{j^*} \leftarrow b_{j^*} + [1 - \frac{(j^*-1)}{10 s_i}]$
  end
end

$s \leftarrow \frac{10}{n} \sum_{j=1}^{10} (b_j - \frac{n}{10})^2$ `// Compute` χ^2 `statistic based on observed counts` $\{b_1, ..., b_{10}\}$ `vs` $\{\frac{n}{10}, ..., \frac{n}{10}\}$

$p \leftarrow$ P-value corresponding to s from χ^2_9 dist'n `//` $B - 1$ `degrees of freedom`

if $p > p^*$ **then**
  return *Success*
else
  return *Failure*
end

δ_i	$\hat{S}_\Theta(\, d_i \,|\, \vec{x}_i \,)$	$\hat{S}_\Theta(\, c_i \,|\, \vec{x}_i \,)$	Value	Eqn
1	$< p_k$		0.0	
1	$\in [p_k, p_{k+1})$		1.0	(21)
1	$\geq p_{k+1}$		0.0	
0		$< p_k$	0.0	
0		$\in [p_k, p_{k+1})$	$\frac{S(\, c_i \,) - p_k}{S(\, c_i \,)}$	(22)
0		$\geq p_{k+1}$	$\frac{p_{k+1} - p_k}{S(\, c_i \,)}$	(23)

Table 9: Summary of 6 types of patients, wrt each interval $I_k = [p_k, p_{k+1})$, using $S(c_i)$ to abbreviate $\hat{S}_\Theta(\, c_i \,|\, \vec{x}_i \,)$

censored patients whose survival probability was contained in a previous bucket (item (β) above).

Theorem B.3 below proves that the expected value of b_k for the correct distribution is equal for all buckets — *i.e.*, $\mathbb{E}[b_k] = p_{k+1} - p_k$ — which allows us to apply the goodness-of-fit test with uniform proportions.

We assume that all survival curves are *strictly* monotonically decreasing, meaning we have the equivalence, $d_i \leq c_i \iff S(d_i) \geq S(c_i)$, which allows us to replace $d_i \leq c_i$ with $S(d_i) \geq S(c_i)$, within the indicator functions in b_k. To simplify notation, we define $I_k := [p_k, p_{k+1})$, $S_c := S(\, c \,|\, \vec{x}\,)$, and $S_d := S(\, d \,|\, \vec{x}\,)$. The proof below shows that the expected value of the summand within Equations (21) – (23) above is equal to $p_{k+1} - p_k$ for the true survival function — *i.e.*, we ignore $\frac{1}{|V|} \sum_{i=1}^{|V|} [\cdot]$ and take the expected value of the term inside the summation.

Theorem B.3. *Given the formula for b_k (Equations (21) – (23)), if the true survival function $S(\cdot|\cdot)$ is strictly monotonically decreasing then the proportions are equal across all bins* — i.e., $\mathbb{E}[b_k] = p_{k+1} - p_k$.

Proof.

$$\begin{aligned}
\mathbb{E}[\, b_k \,] = \mathbb{E}\Big[\quad & 1 \quad \cdot \quad \mathcal{I}\,[\, S_d \in I_k \,\wedge\, S_d \geq S_c \,] \\
& + \frac{S_c - p_k}{S_c} \quad \cdot \quad \mathcal{I}\,[\, S_c \in I_k \;\wedge\; S_c > S_d \,] \\
& + \frac{(p_{k+1} - p_k)}{S_c} \cdot \mathcal{I}\,[\, S_c > S_d \,\wedge\, S_c \in [p_{k+1}, 1] \,] \Big]
\end{aligned}$$

$$\begin{aligned}
= \quad & \mathbb{E}\Big[\quad \mathcal{I}\,[\, S_d \in I_k \;\wedge\; S_d \geq S_c \,] \quad \Big] \\
& + \mathbb{E}\left[\quad \frac{S_c - p_k}{S_c} \;\cdot\; \mathcal{I}\,[\, S_c \in I_k \;\wedge\; S_c > S_d \,] \quad \right] \\
& + \mathbb{E}\left[\quad \frac{(p_{k+1} - p_k)}{S_c} \;\cdot\; \mathcal{I}\,[\, S_c > S_d \,\wedge\, S_c \geq p_{k+1} \,] \quad \right]
\end{aligned}$$

$$\begin{aligned}
= \quad & \Pr[\, S_d \in I_k \;\wedge\; S_d \geq S_c \,] \\
& + \Pr[S_c \in I_k \;\wedge\; S_c > S_d] \; - p_k\, \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\,[\, S_c > S_d \,\wedge\, S_c \in I_k \,] \right] \\
& + (p_{k+1} - p_k)\; \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\,[\, S_c > S_d \,\wedge\, S_c \geq p_{k+1} \,] \right]
\end{aligned}$$

$$\begin{aligned}
= \Pr[\; S_d \in I_k \;\wedge\; S_d \geq S_c \;] \quad + \quad & \Pr[\; S_c \in I_k \;\wedge\; S_c > S_d \;] && \text{(I)} \\
- p_k\; \mathbb{E}\left[\;\frac{1}{S_c} \;\cdot\; \mathcal{I}\,[\, S_c > S_d \;\wedge\; S_c \geq p_k \,] \;\right] & && \text{(II)} \\
+ p_{k+1}\; \mathbb{E}\left[\;\frac{1}{S_c} \;\cdot\; \mathcal{I}\,[\, S_c > S_d \,\wedge\, S_c \geq p_{k+1} \,] \;\right] & && \text{(III)}
\end{aligned}$$

Focusing on the second probability in line (I), note $S_c \in I_k = [p_k, p_{k+1})$ and $S_c > S_d$ imply that $S_d \in [0, p_{k+1})$, which can be expanded to the cases for $S_d < p_k$ and $S_d \in I_k$. Using this, we reformulate the probability by noting the equivalence of the event space,

$$\Pr[\; S_c \in I_k \;\wedge\; S_c > S_d] \;=\; \Pr[S_c \in I_k \;\wedge\; S_d < p_k] \;+\; \Pr[(S_c \wedge S_d) \in I_k \;\wedge\; S_c > S_d \,].$$

Combining the second addend above with the first probability in line (I), we again simplify by noting these probabilities bound $S_c < p_{k+1}$,

$$\Pr[S_d \in I_k \;\wedge\; S_d \geq S_c] \;+\; \Pr[(S_c \wedge S_d) \in I_k \;\wedge\; S_c > S_d] \;=\; \Pr[S_d \in I_k \;\wedge\; S_c < p_{k+1}].$$

Using this simplification we can rewrite the entirety of line (I),

$$\begin{aligned}
& \Pr[\; S_d \in I_k \;\wedge\; S_d \geq S_c \;] \quad + \quad \Pr[\; S_c \in I_k \;\wedge\; S_c > S_d \;] \\
= \quad & \Pr[\; S_d \in I_k \;\wedge\; S_c < p_{k+1} \;] \quad + \quad \Pr[\; S_c \in I_k \;\wedge\; S_d < p_k \;]
\end{aligned}$$

Recalling the independence assumption, $c \perp d$ (recall everything is conditioned on the instance $\vec{x}$), we have the following equalities:

$$\begin{aligned}
\Pr[S_d \in I_k \ \wedge \ S_c < p_{k+1}] &= \Pr[S_d \in I_k] \cdot \Pr[S_c < p_{k+1}] &&= (p_{k+1} - p_k)\ \Pr[S_c < p_{k+1}], \\
\Pr[S_c \in I_k \ \wedge \ S_d < p_k] &= \Pr[S_c \in I_k] \cdot \Pr[S_d < p_k] &&= \Pr[S_c \in I_k]\, p_k\,,
\end{aligned}$$

where the final equalities are due to the uniformity of the survival function on d, $S(d) \sim U(0,1)$ — see Lemma B.2. This then leaves the final simplification of line (I) as,

$$\begin{aligned}
\Pr[S_d \in I_k \ \wedge \ S_d \geq S_c] + \Pr[S_c \in I_k \ \wedge \ S_c > S_d] \ =& (p_{k+1} - p_k)\ \Pr[S_c < p_{k+1}] \\
&+\ p_k\ \Pr[S_c \in I_k].
\end{aligned}$$

Now we address line (II):

$$\begin{aligned}
-p_k\, \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\left[\, S_c > S_d \ \wedge\ S_c > p_k \,\right]\right] &= -p_k\ \left(\int_{p_k}^{1}\int_{0}^{S_c} \frac{1}{S_c}\, f(S_c)\, dS_d\, dS_c\right) \quad \text{(Def. of } \mathbb{E}[\cdot]) \\
&= -p_k\ \left(\int_{p_k}^{1} \frac{S_c}{S_c}\, f(S_c)\, dS_c\right) \\
&= -p_k\ \Pr[S_c > p_k]
\end{aligned}$$

where $f(\,\cdot\,)$ is the probability distribution function (PDF) for the distribution generated by the survival function applied to a *censored* observation. As the censoring distribution is unknown, $f(\,S_c\,)$ is also unknown, whereas $f(\,S_d\,)$ is the PDF of the uniform distribution.

Following the steps above for line (III) analogously gives us

$$p_{k+1}\, \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\left[\, S_c > S_d \ \wedge\ S_c > p_{k+1} \,\right]\right] \quad = \quad p_{k+1}\ \Pr[S_c > p_{k+1}]$$

Combining the simplifications of lines (I), (II) and (III), we have the following,

$$\begin{aligned}
\mathbb{E}[b_k] \ =& (p_{k+1} - p_k)\ \Pr[S_c < p_{k+1}] \ +\ p_k\ \Pr[S_c \in I_k] && \text{(I)} \\
&-\ p_k\ \Pr[S_c > p_k] && \text{(II)} \\
&+\ p_{k+1}\ \Pr[S_c > p_{k+1}] && \text{(III)} \\
\\
=& p_{k+1}\ (\Pr[S_c < p_{k+1}] \ +\ \Pr[S_c > p_{k+1}]) \\
&-\ p_k\, (\Pr[S_c < p_{k+1}] - \Pr[S_c \in [p_k, p_{k+1}) + \Pr[S_c > p_k]) \\
\\
=& p_{k+1} - p_k
\end{aligned}$$

□

This proof requires the assumption that survival curves are *strictly* monotonically decreasing on [0,1]. This means survival curves will not contain any large flat areas; this means there will not be non-zero probability mass for $S(c_i) = S(d_i)$ when $c_i \neq d_i$, which means certain terms in the proof below would fail to cancel with one another, leaving us with non-equivalent proportions within each bucket (specifically higher proportions within buckets that contain these flat lines).

A natural corollary of Theorem B.3 is that all consistent estimators of the true survival distribution will be D-Calibrated (if the true survival distribution is strictly monotonic). Furthermore, if survival time is independent and identically distributed (i.i.d.) across patients, then there will only be one true survival curve for all patients, and thus, Kaplan-Meier is uniformly consistent (Breslow and Crowley, 1974; CsörgŐ and Horváth, 1983):

Lemma B.4. *The Kaplan-Meier distribution is asymptotically D-Calibrated.*

This is consistent with the results given in Section 4.2, which showed that KM always passed the D-Calibration test with a p-value 1.000, in all 8 data sets. If all of the data was uncensored, we would expect the typical 5% Type I error rate for claiming p <0.05 as significant; however, in the presence of censored data, the proportion of patients within buckets become smoothed, boosting the p-value.

B.5.1. D-Calibration is Different from IBS and 1-Calibration

The next two propositions prove that D-calibration is different from other, more standard models.

Proposition B.5. *Given a single set of survival data subjects, it is possible for one* ISD *model $\hat{S}_1$ to be perfectly D-calibrated but have a worse IBS score than another model $\hat{S}_2$ that is not D-calibrated.*

Proof. Figure 10 shows two ISD models, for two patients: $\vec{x}_Q$ who died at time 1 and $\vec{x}_G$ who died at time 2. The left model $\hat{S}_{Left}$ is clearly D-calibrated, for 2 buckets, as it includes 1 patient whose probability at the time of death $\hat{S}_{Left}(\, d_i \,|\, \vec{x}_i \,)$ is in each of [0, 0.5], and (0.5, 1.0]. However the right model $\hat{S}_{Right}(\, \cdot \,|\, \cdot \,)$ is not D-calibrated, as both patients are in the single bucket (0.5, 1.0].

To compute the IBS scores: We first compute the relevant "$(\ \hat{S}(\, t \,|\, x \,) - \mathcal{I}[t; x]\,)^2$" integrals for each patient, then divide by the total time t_{max} considered (which here is $t_{max} = 2$), then take the average over the 2 patients. First, for the $\hat{S}_{Left}(\, \cdot \,|\, \cdot \,)$ model on the left: the IBS integrals for these 2 patients are

$$\begin{array}{rcccccccc} V(\hat{S}_{Left},\ \vec{x}_Q) & = & \int_0^1 (1-1)^2 dt & + & \int_1^2 (\frac{1}{2}-0)^2 dt & = & 0 + \frac{1}{4} & = & \frac{1}{4} \\ V(\hat{S}_{Left},\ \vec{x}_G) & = & \int_0^1 (1-1)^2 dt & + & \int_1^2 (\frac{1}{2}-1)^2 dt & = & 0 + \frac{1}{4} & = & \frac{1}{4} \end{array}$$

Hence, the IBS score, for this $\hat{S}_{Left}$ model over these 2 patients, is

$$\text{IBS}(\hat{S}_{Left}, \{\vec{x}_Q, \vec{x}_G\}\,) \quad = \quad \frac{1}{t_{max}}\ \frac{1}{2} \sum_i V(\hat{S}_{Left}, \vec{x}_i) \quad = \quad \frac{1}{8} \quad = \quad 0.125\ .$$

Now consider the $\hat{S}_{Right}$ model, on the right. Here, we assume the horizontal bar, at height $\hat{S}_{Right}(\cdot \,|\, \cdot) = 3/4$, has width $\epsilon = 1/6$. (Of course, the patient died at the right end.) Here,

$$\begin{array}{rcccccccccc} V(\hat{S}_{Right}, \vec{x}_Q) & = & \int_0^{1-\epsilon}(1-1)^2 dt & + & \int_{1-\epsilon}^{1}(\frac{3}{4}-1)^2 dt & + & \int_1^2 (0-0)^2 dt & = & 0 + \frac{1}{16}\epsilon + 0 & = & \frac{\epsilon}{16} \\ V(\hat{S}_{Right}, \vec{x}_G) & = & \int_0^{2-\epsilon}(1-1)^2 dt & + & \int_{2-\epsilon}^{2}(\frac{3}{4}-1)^2 dt & & & = & 0 + \frac{1}{16}\epsilon & = & \frac{\epsilon}{16} \end{array}$$

Here,

$$\text{IBS}(\hat{S}_{Right}, \{\vec{x}_Q, \vec{x}_G\}) \;=\; \frac{1}{t_{max}} \, \frac{1}{2} \sum_i V(\hat{S}_{Right}, \vec{x}_i) \;=\; \frac{\epsilon}{32} \;\approx\; 0.00521$$

for this value of ϵ. Notice this is a factor of 24 smaller ("better") than the IBS score for $\hat{S}_{Left}$. See Table 3. □

Proposition B.6. *It is possible for an* ISD *model to be perfectly D-calibrated but not 1-calibrated at a time t^*, and for (another)* ISD *model to be perfectly 1-calibrated at time t^* but not D-calibrated.*

Proof. **"1-Calibration $\not\Rightarrow$ D-Calibration":** Consider the model shown in Figure 11[left]. Here, the green curve corresponds to 4 apparently-identical patients $\{\vec{x}_{g,1}, \ldots, \vec{x}_{g,4}\}$, and the red curve, to apparently-identical $\{\vec{x}_{r,1}, \ldots, \vec{x}_{r,4}\}$. The "$*$"s mark the time when each patient died, denoted as $d_{\vec{x}}$ for $\vec{x}$. We intentionally use simple examples, with no censored patients, with curves that go to 0. Note this model assigns $\hat{S}(\, T_1 \,|\, \vec{x}_{g,i}\,) \;=\; 0.75$ for each of the 4 green patients, and $\hat{S}(\, T_1 \,|\, \vec{x}_{r,j}\,) \;=\; 0.25$ for each of the 4 red patients.

To show that this model is 1-Calibrated, with respect to T_1: Recall we first sort the set of all 8 $\{\, \hat{S}(\, T_1 \,|\, \vec{x}\,)\,\}$ values, then partition them into k bins. Here, we consider $k = 2$, rather than the deciles earlier. The first bin contains the 4 patients with $\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.75$ (note this includes exactly the 4 "green" patients $\vec{x}_{g,i}$); and the second, the 4 patients with $\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.25$ (corresponding to the 4 "red" patients $\vec{x}_{g,i}$). Now note that 3 of the 4 "$\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.75$ patients" are alive at T_1; and 1 of the 4 "$\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.25$ patients" are alive at T_1 — which means this model is perfectly 1-Calibrated at T_1.

However, this model is not D-Calibrated: to be consistent with the earlier 1-Calibration analysis, we split the [0,1] probability interval into 2 buckets (not 10), as shown in Figure 11. Here, $\hat{S}(\, d_{\vec{x}} \,|\, \vec{x}\,) \in [0.5, 1]$ holds for only 1 patient (just $\vec{x}_{r,1}$), and $\hat{S}(\, d_{\vec{x}} \,|\, \vec{x}\,) \in [0, 0.5]$ holds for 7; if the model was D-Calibrated, each of these two buckets should contain 4 patients.

"D-Calibration $\not\Rightarrow$ 1-Calibration": See Figure 11[right], where again, each curve represents 4 distinct, but indistinguishable patients; notice the outcomes are different from those on the left. To see that this model is D-Calibrated, note there are 4 patients with $\hat{S}(\, d_{\vec{x}} \,|\, \vec{x}\,) \in [0.5, 1]$ (which happen to be the green patients), and 4 with $\hat{S}(\, d_{\vec{x}} \,|\, \vec{x}\,) \in [0, 0.5)$ (the red patients). However, the model is not 1-Calibrated, at T_1: Of the 4 patients with $\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.75$, 2 are alive at T_1 (these are the 2 right-most green patients); and of the 4 patients with $\hat{S}(\, T_1 \,|\, \vec{x}\,) = 0.25$, 2 are alive at T_1 (these are the 2 right-most red patients). To be 1-Calibrated, there should be 3 living patients in the first bin, and 1 in the second; hence, this model is not 1-Calibrated at T_1. □

Appendix C. Comments about Various ISD's

C.1. Overview of MTLR

Consider[25] modeling the probability of survival of patients at each of a vector of time points $\tau = [t_1, t_2, \ldots, t_m]$ — *e.g.*, τ could be the 60 monthly intervals from 1 month up to 60 months. We can set up a series of logistic regression models: For each patient, represented as $\vec{x}$,

$$S_{\vec{\theta}_i}(\, T \geq t_i \,|\, \vec{x}\,) \quad = \quad \Big(1 + \exp(\vec{\theta}_i \cdot \vec{x})\Big)^{-1}, \qquad 1 \leq i \leq m, \tag{24}$$

where $\vec{\theta}_i$ are the time-specific parameter vectors. While the input features $\vec{x}$ stay the same for all these classification tasks, the binary labels $y_i = [T \geq t_i]$ can change depending on the threshold t_i. We encode the survival time d of a patient as a sequence of binary values: $y = y(d) = [y_1, y_2, \ldots, y_m]$, where $y_i = y_i(d) \in \{0, 1\}$ denotes the survival status of the patient at time t_i, so that $y_i = 0$ (no death event yet) for all i with $t_i < d$, and $y_i = 1$ (death) for all i with $t_i \geq d$. Here there are $m + 1$ possible legal sequences of the form[26] $[0, 0, \ldots, 1, 1, \ldots, 1]$, including the sequence of all '0's and the sequence of all '1's. Our MTLR model extends Equation 24 by computing the probability of observing the survival status sequence $y = [y_1, y_2, \ldots, y_m]$ as:

$$S_{\mathbf{\Theta}}(\, Y{=}[y_1, y_2, \ldots, y_m] \,|\, \vec{x}\,) \quad = \quad \frac{\exp(\sum_{i=1}^{m} y_i \times \vec{\theta}_i \cdot \vec{x})}{\sum_{k=0}^{m} \exp(f_{\mathbf{\Theta}}(\vec{x}, k))},$$

where $\mathbf{\Theta} = [\vec{\theta}_1, \ldots, \vec{\theta}_m]$, and $f_{\mathbf{\Theta}}(\vec{x}, k) = \sum_{i=k+1}^{m} (\vec{\theta}_i \cdot \vec{x})$ for $0 \leq k \leq m$ is the score of the sequence with the event occurring in the interval $[t_k, t_{k+1})$ before taking the logistic transform, with the boundary case $f_{\mathbf{\Theta}}(\vec{x}, m) = 0$ being the score for the sequence of all '0's. Given a data set of n patients $\{\vec{x}_r\}$ with associated time of deaths $\{d_r\}$, we find the optimal parameters (for the MTLR model) $\mathbf{\Theta}^*$ as

$$\mathbf{\Theta}^* \;=\; \arg\max_{\mathbf{\Theta}} \sum_{r=1}^{n} \left[\sum_{i=1}^{m} y_j(d_r)(\vec{\theta}_i \cdot \vec{x}_r) - \log \sum_{k=0}^{m} \exp f_{\mathbf{\Theta}}(\vec{x}_r, k) \right] - \frac{C}{2} \sum_{j=1}^{m} \|\vec{\theta}_j\|^2 \tag{25}$$

where the C (for the regularizer) is found by an internal cross-validation process. (The original Yu et al. (2011) paper also included a smoothing regulatrizer, $\sum_{j=1}^{m-1} \|\vec{\theta}_{j+1} - \vec{\theta}_j\|^2$; however, Jin (2015, Appendix) proved this was not necessary.)

There are many details here — *e.g.*, to insure that the survival function starts at 1.0, and decreases monotonically and smoothly until reaching 0.0 for the final time point; to deal appropriately with censored patients; to decide how many time points to consider (m); and to minimize the risk of overfitting (by regularizing), and by selecting the relevant features. The paper by Yu et al. (2011) provides the details.

Afterwards, the ISD-Predictor can use the learned MTLR-model $\mathbf{\Theta}^* = [\vec{\theta}_1, \ldots, \vec{\theta}_m]$ to produce a curve for a novel patient, who is represented as the vector of his/her covariates $\vec{x}_j$. This involves computing the m values, $[f_1(\vec{x}_j, \vec{\theta}_1), \ldots,\ f_r(\vec{x}_j, \vec{\theta}_m)]$; the running sum of

25. This paragraph is paraphrased from (Yu et al., 2011); reprinted with permission of publisher/author.
26. Notice there are no '0's after a '1'. This is the 'no zombie' rule: once someone dies, that person stays dead.

these values is essentially the survival curve. We then use splines to produce a smooth, monotonically decreasing curve — such as the 10 such curves shown in Figure 5 [bottom-right].

C.2. Extension to Random Survival Forests (RSF-KM)

Given a labeled data set, a random survival forest learner will produce a set of T decision trees from T different bootstrapped samples of the training data. It grows each tree recursively, starting from the root — identifying each position with the set of patients who arrive there. For each position, the growth stops if there are fewer than d_0 deaths (where d_0 is chosen via cross-validation). Otherwise, it identifies an appropriate feature for this node: it first randomly draws a small random subset of the features to consider, then selects the feature (from that subset) that maximizes the difference in survival between two daughter nodes, based on the lo-grank test statistic (or some other chosen splitting rule). This becomes the "rule" of that node; the learner then considers its two daughters, by splitting on the node's feature.

Each leaf node in each tree corresponds to the set of training instances that reached that node. Given these learned trees, to classify a novel instance $\vec{x}$, the random forest performance system will drop $\vec{x}$ into each of the trees, which will lead to T different leaf nodes; it will then use the T subsets of training instances to make a decision. Since each terminal node in the random survival forest contains a set of instances, we can use these instances to produce a Kaplan-Meier curve.[27]

Once the survival forest has been learned (with T trees), a patient is dropped into each of the T survival trees, leading to T leaf nodes, which produces T Kaplan-Meier curves. The RSF-KM implementation then "averages" these curves, by taking a point-wise average across the curve for all time points; see Figure 18.[28]

Note that the risk score generated by the median of the individual survival curves (produced here) does not necessarily result in the same ordering of patients as the risk scores of the original RSF implementation, which uses averaged cumulative hazards as a risk score. For this reason, we also applied the original RSF process to the data sets presented in the paper. We found that the Concordance scores were similar to that of RSF-KM; MTLR still outperformed RSF on the data sets where MTLR outperformed RSF-KM (data not shown).

Appendix D. Detailed Empirical Results

This sub-appendix includes the tables that correspond to the figures given in Section 4.2. Further, Appendix D.4 provides the all p-values for the 1-Calibration tests.

D.1. Concordance

See Table 10 for the results corresponding to Figure 12.

27. While the original paper does not consider survival curves, documentation `https://kogalur.github.io/randomForestSRC/theory.html#section8.1` describing the inner workings of the R package states that survival curves in terminal nodes are created via the Kaplan-Meier estimator.

28. The method for generating individual survival curves could not be found in any of the literature by the authors of random survival forests. Survival curves were reverse-engineered by the authors of this paper — all survival curves tested matched the methodology explained here.

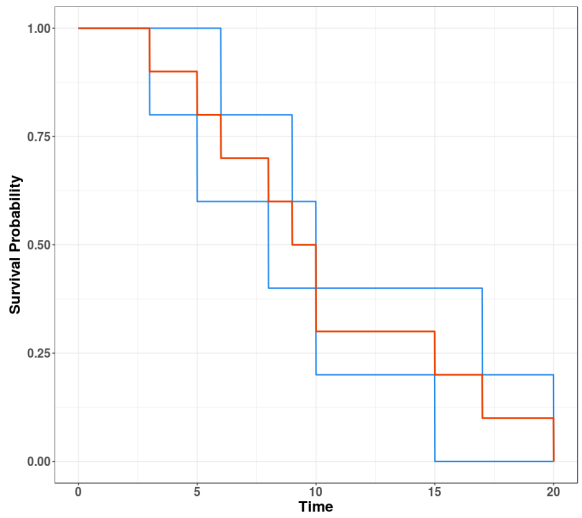


Figure 18: This figure illustrates how to combine two different survival curves to produce a new one. (RSF-KM uses this idea to "merge" the curves obtained from the various leaf nodes reached by a novel instance.) Here, two survival curves, given in blue, are averaged to produce the survival curve shown in dark orange. Note that the averaged curve is generated from a point-wise average, *i.e.*, new calculations must only be computed at each death time — *i.e.*, , whenever there is a drop in either (blue) Kaplan-Meier curve.

	GBM	GLI	NACD-COL	NACD	READ	BRCA	DBCD	DLBCL
KM	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000))
AFT	0.692 (0.026)	0.802 (0.011)	0.722 (0.015)	0.755 (0.008)	0.700 (0.183)	0.738 (0.041)	0.474 (0.056)	0.558 (0.046)
COX-KP	0.696 (0.026)	0.805 (0.01)	0.722 (0.015)	0.755 (0.008)	0.716 (0.157)	0.747 (0.030)	-	-
COXEN-KP	0.698 (0.032)	0.801 (0.006)	0.726 (0.024)	0.754 (0.009)	0.731 (0.141)	0.761 (0.032)	0.744 (0.055)	0.695 (0.038)
RSF-KM	0.650 (0.038)	0.776 (0.021)	**0.743 (0.024)**	**0.758 (0.005)**	0.582 (0.093)	0.704 (0.036)	0.738 (0.060)	0.639 (0.038)
MTLR	**0.703 (0.032)**	**0.812 (0.013)**	0.734 (0.023)	0.757 (0.008)	**0.768 (0.091)**	**0.770 (0.027)**	**0.766 (0.065)**	**0.704 (0.034)**

Table 10: Concordance results corresponding to Figure 12. **Bold** values indicate the best (highest Concordance) performing model, for each data set.

D.2. Brier Score

See Table 11 for the results corresponding to Figure 13.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL
KM	0.046 (0.001)	0.034 (0.000)	0.083 (0.001)	0.089 (0.000)	0.027 (0.005)	0.017 (0.001)	0.097 (0.003)	0.109 (0.004)
AFT	0.041 (0.003)	0.021 (0.002)	0.066 (0.003)	0.065 (0.002)	0.026 (0.017)	0.0133 (0.002)	0.254 (0.023)	0.295 (0.038)
COX-KP	0.040 (0.003)	0.022 (0.003)	0.066 (0.003)	0.067 (0.002)	0.026 (0.017)	0.014 (0.002)	-	-
COXEN-KP	0.040 (0.003)	0.021 (0.002)	0.064 (0.003)	0.065 (0.001)	0.024 (0.011)	0.015 (0.002)	**0.070 (0.004)**	0.078 (0.013)
RSF-KM	0.059 (0.009)	0.051 (0.009)	0.079 (0.006)	0.079 (0.001)	0.047 (0.013)	0.028 (0.002)	0.077 (0.003)	0.095 (0.013)
MTLR	**0.039 (0.004)**	**0.019 (0.002)**	**0.062 (0.003)**	**0.063 (0.001)**	**0.023 (0.006)**	**0.012 (0.001)**	0.070 (0.003)	**0.078 (0.011)**

Table 11: Integrated Brier score results corresponding to Figure 13. **Bold** values indicate the best performing model for each data set — with the lowest Integrated Brier score. Note that this table (and all following tables) may show ties (up to three digits), but will only bold the one with the best performance, even if based on the additional digits not shown.

D.3. Empirical Values of L1-Loss, and Variants

Table 12 provides the results for the Margin-L1-loss, shown in Figure 14.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL
KM	1431.31 (59.25)	2746.70 (91.85)	56.45 (1.95)	61.97 (0.50)	3677.90 (222.77)	5392.04 (128.19)	24.88 (0.78)	20.28 (2.06)
AFT	**1240.60 (57.38)**	1838.20 (105.23)	47.89 (1.85)	43.99 (1.44)	4068.72 (1451.14)	5156.1 (264.50)	47.01 (4.38)	27.29 (3.31)
COX-KP	1278.18 (44.02)	1824.06 (127.49)	45.53 (2.23)	44.26 (1.29)	4799.91 (1460.52)	6247.01 (612.3)	-	-
COXEN-KP	1347.36 (51.6)	1683.04 (110.82)	45.25 (1.94)	45.72 (1.43)	3564.01 (1163.92)	**4593.52 (370.75)**	24.28 (2.14)	15.97 (1.10)
RSF-KM	1399.17 (99.06)	4503.49 (465.47)	58.81 (2.28)	49.69 (1.38)	6805.00 (1710.50)	10934.45 (579.44)	26.58 (2.35)	17.68 (1.81)
MTLR	1271.73 (37.71)	**1582.72 (131.1)**	**43.48 (2.52)**	**43.97 (1.20)**	**3417.49 (256.83)**	4669.55 (153.50)	**20.01 (1.47)**	**15.52 (1.97)**

Table 12: Margin-L1-loss results corresponding to Figure 14. **Bold** values indicate the best performing model for each data set — with the lowest Margin-L1-loss.

D.4. 1-Calibration

Each table in this subappendix corresponds to a different percentile of event times for each data set, showing the 10th, 25th, 50th, 75th, and 90th percentiles. **Bolded** values indicate models that passed 1-Calibration ($p > 0.05$). The "Total" column of each table gives the total number of data sets passed by each model — that is, the values in those columns correspond to Table 6.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	Total
AFT	0.001	**0.159**	**0.794**	0.012	**1.000**	**0.919**	0.000	0.000	4
COX-KP	0.001	**0.140**	**0.794**	0.008	**0.999**	**0.782**	-	-	4
COXEN-KP	0.000	0.033	0.043	0.000	**0.999**	**0.561**	**0.454**	**0.646**	4
RSF-KM	0.000	0.000	**0.078**	0.016	**0.998**	0.000	**0.164**	**0.273**	4
MTLR	**0.908**	**0.450**	**0.440**	0.047	**1.000**	**0.929**	0.000	**0.177**	6

Table 13: 1-Calibration Results at $t^* = 10$th Percentile of Event Times

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	**Total**
AFT	0.000	0.040	**0.586**	0.009	0.000	**0.205**	0.000	0.000	2
COX-KP	0.000	0.008	**0.379**	0.003	0.000	**0.535**	-	-	2
COXEN-KP	0.000	0.002	0.003	0.000	**0.238**	0.044	**0.436**	**0.547**	3
RSF-KM	0.000	0.000	**0.312**	0.006	0.000	0.000	0.042	**0.227**	2
MTLR	**0.963**	**0.312**	**0.645**	**0.254**	**0.449**	**0.448**	**0.177**	**0.052**	8

Table 14: 1-Calibration Results at t^* = 25th Percentile of Event Times

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	**Total**
AFT	**0.117**	0.030	0.035	0.043	0.000	0.000	0.000	0.000	1
COX-KP	**0.495**	0.005	0.038	**0.124**	0.000	0.017	-	-	2
COXEN-KP	0.019	0.000	0.000	0.000	0.049	0.000	0.025	**0.822**	1
RSF-KM	0.000	0.000	**0.761**	0.001	0.000	0.000	0.000	**0.068**	2
MTLR	**0.796**	**0.306**	**0.813**	**0.112**	**0.995**	0.013	0.041	**0.262**	6

Table 15: 1-Calibration Results at t^* = 50th Percentile of Event Times

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	**Total**
AFT	**0.378**	0.000	0.002	0.002	0.000	0.000	0.000	0.000	1
COX-KP	0.008	0.000	0.003	0.004	**0.087**	0.016	-	-	1
COXEN-KP	**0.338**	0.000	0.000	0.000	0.001	0.000	0.003	**0.436**	2
RSF-KM	0.000	0.000	**0.070**	0.003	0.000	0.000	0.002	0.038	1
MTLR	**0.140**	**0.565**	0.044	0.045	0.026	0.000	0.036	**0.218**	3

Table 16: 1-Calibration Results at t^* = 75th Percentile of Event Times

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	**Total**
AFT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0
COX-KP	0.000	0.000	0.000	0.000	0.000	0.000	-	-	0
COXEN-KP	**0.050**	0.000	0.004	0.000	0.000	0.000	0.010	**0.112**	2
RSF-KM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0
MTLR	**0.109**	**0.148**	0.000	0.001	0.000	0.000	**0.098**	**0.157**	4

Table 17: 1-Calibration Results at t^* = 90th Percentile of Event Times

References

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Fern Anderson, G Michael Downing, Jan Hill, Lynn Casorso, and Noreen Lerch. Palliative performance scale (pps): a new tool. *Journal of palliative care*, 12(1):5–11, 1995.

Axel Andres, Aldo Montano-Loza, Russell Greiner, Max Uhlich, Ping Jin, Bret Hoehn, David Bigam, James Andrew Mark Shapiro, and Norman Mark Kneteman. A novel

learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS one*, 13(3):e0193523, 2018.

John E Angus. The probability integral transform and related results. *SIAM review*, 36(4): 652–654, 1994.

Norman Breslow and John Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, pages 437–453, 1974.

Glenn W Brier and Roger A Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

Rong-Bin Chuang, Wen-Yu Hu, Tai-Yuan Chiu, and Ching-Yu Chen. Prediction of survival in terminal cancer patients in taiwan: constructing a prognostic scale. *Journal of pain and symptom management*, 28(2):115–122, 2004.

GA Colditz, KA Atwood, K Emmons, RR Monson, WC Willett, D Trichopoulos, and DJ Hunter. Harvard report on cancer prevention volume 4: Harvard cancer risk index. *Cancer causes & control*, 11(6):477–488, 2000.

Graham A Colditz and Bernard Rosner. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the nurses' health study. *American journal of epidemiology*, 152(10):950–964, 2000.

Joseph P Costantino, Mitchell H Gail, David Pee, Stewart Anderson, Carol K Redmond, Jacques Benichou, and H Samuel Wieand. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*, 91(18):1541–1548, 1999.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246.

Sándor CsörgŐ and Lajos Horváth. The rate of strong uniform consistency for the product-limit estimator. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62(3): 411–426, 1983.

RB d'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

M.H. DeGroot and S.E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1):12–22, 1983.

Lloyd D Fisher and Danyu Y Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.

Genome Data Analysis Center. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run*. Broad Institute of MIT and Harvard, 2016. URL `https://doi.org/10.7908/C11G0KM9`.

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, Tianxi Cai, and Martin Schumacher. The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4): 457–479, 2008.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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.

Danielle Guffey. *Hosmer-Lemeshow goodness-of-fit test: Translations to the Cox Proportional Hazards Model*. PhD thesis, University of Washington, 2013.

Ramesh C Gupta and David M Bradley. On representing the mean residual life in terms of the failure rate. *arXiv preprint math/0411297*, 2004.

B Gwilliam, V Keeley, C Todd, C Roberts, M Gittins, L Kelly, S Barclay, and P Stone. Prognosticating in patients with advanced cancer – observational study comparing the accuracy of clinicians' and patients' estimates of survival. *Annals of oncology*, 24:482–488, 2012.

Humza Haider. Individual survival distributions: A more effective tool for survival prediction. Master's thesis, University of Alberta, 8 2019. `https://era.library.ualberta.ca/items/f83088e6-eda0-4596-b2da-6eec608998ee`.

David Harrington. Linear rank tests in survival analysis. *Encyclopedia of Biostatistics*, 2005.

JL Haybittle, RW Blamey, CW Elston, Jane Johnson, PJ Doyle, FC Campbell, RI Nicholson, and K Griffiths. A prognostic index in primary breast cancer. *British journal of cancer*, 45(3):361, 1982.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

Robin Henderson and Niels Keiding. Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31(12):703–706, 2005.

Hanne Hollnagel. Explaining risk factors to patients during a general practice consultation: conveying group-based epidemiological knowledge to individual patients. *Scandinavian journal of primary health care*, 17(1):3–5, 1999.

David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

David W Hosmer, Stanley Lemeshow, and Susanne May. *Applied survival analysis*. Wiley Blackwell, 2011.

H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.6.1.

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

Ping Jin. Using survival prediction techniques to learn consumer-specific reservation price distributions. Master's thesis, University of Alberta, 2015. `https://pdfs.semanticscholar.org/58c6/413b6cb68844b42ca0eeaf7b51b06bbb0b6a.pdf`.

J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data*. Wiley New York:, 2002.

Patrick S Kamath and W Ray Kim. The model for end-stage liver disease (meld). *Hepatology*, 45(3):797–805, 2007.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. ISSN 0162-1459.

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):24, 2018.

Luke Kumar and Russell Greiner. Gene expression based survival prediction for cancer patients—a topic modeling approach. *PloS one*, 14(11), 2019.

Yan Li, Jie Wang, Jieping Ye, and Chandan K Reddy. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1715–1724. ACM, 2016.

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.

T Morita, Junichi Tsunoda, Satoshi Inoue, and Satoshi Chihara. The palliative prognostic index: a scoring system for survival prediction of terminally ill cancer patients. *Supportive Care in Cancer*, 7(3):128–133, 1999.

Allan H Murphy. Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology*, 11(2):273–282, 1972.

Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.

Marco Pirovano, Marco Maltoni, Oriana Nanni, Mauro Marinari, Monica Indelli, Giovanni Zaninetta, Vincenzo Petrella, Sandro Barni, Ernesto Zecca, Emanuela Scarpi, et al. A new palliative prognostic score: a first step for the staging of terminally ill cancer patients. *Journal of pain and symptom management*, 17(4):231–239, 1999.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

Malcolm P Rogers, John Orav, and Peter McL Black. The use of a simple likert scale to measure quality of life in brain tumor patients. *Journal of neuro-oncology*, 55(2):121–131, 2001.

Stanisław Saks. *Theory of the Integral*. Instytut Matematyczny Polskiej Akademi Nauk (Warszawa-Lwów), 1937.

P.K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *ICDM 2007*, pages 655–660. IEEE, 2008.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (BART). *Statistics in medicine*, 35(16):2741–2753, 2016.

Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems*, pages 1209–1216, 2008.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL `https://CRAN.R-project.org/package=survival`. version 2.38.

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

Hans C van Houwelingen, Tako Bruinsma, Augustinus AM Hart, Laura J van't Veer, and Lodewyk FA Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, 25(18):3201–3216, 2006.

JianLi Wang, Jitender Sareen, Scott Patten, James Bolton, Norbert Schmitz, and Arden Birney. A prediction algorithm for first onset of major depression in the general population: development and validation. *Journal of epidemiology and community health*, pages jech–2013, 2014.

Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*, 2017.

Andrew Wey, John Connett, and Kyle Rudser. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3): 537–549, 2015.

Daniela M Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51, 2010.

Ian Witten, Eibe Frank, and Mark Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

Guofen Yan and Tom Greene. Investigating the effects of ties on measures of concordance. *Statistics in medicine*, 27(21):4190–4206, 2008.

Yi Yang and Hui Zou. A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and its Interface*, 6(2):167–173, 2013.

Yi Yang and Hui Zou. *fastcox: Lasso and Elastic-Net Penalized Cox's Regression in High Dimensions Models using the Cocktail Algorithm*, 2017. URL `https://CRAN.R-project.org/package=fastcox`. R package version 1.1.3.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*, 2011.