Effective Ways to Build and Evaluate Individual Survival Distributions

Humza Haider, Bret Hoehn, Sarah Davis, Russell Greiner
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8
{hshaider, bhoehn, sdavis1, rgreiner}@ualberta.ca

November 16, 2018

Abstract

An accurate model of a patient's individual survival distribution can help determine the appropriate treatment for terminal patients. Unfortunately, risk scores (*e.g.*, from Cox Proportional Hazard models) do not provide survival *probabilities*, single-time probability models (*e.g.*, 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 which provides an individual survival *distribution* 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, Brier score, and various versions of L1-loss– and 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 datasets.

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

1 Introduction

When diagnosed with a terminal disease, many patients ask about their prognosis [21]: "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 [10]) provide only *relative* survival measures, but not the calibrated probabilities desired; (2) single-time probability models (*e.g.*, the Gail model [9]) provide a probability value but *only for a single time point*; and (3) class-based survival curves (like Kaplan-Meier, KM [31]) 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. This paper considers several ISD models: the Kalbfleisch-Prentice extension of the Cox (COX-KP) [29] and the elastic net Cox (COXEN-KP) [55] model, the Accelerated Failure Time (AFT) model [29], the Random Survival Forest model with Kaplan-Meier extensions (RSF-KM), and the Multi-task Logistic Regression (MTLR) model [57]. Figure 1(middle, right) show survival curves (generated by MTLR) for two of these stage-4 stomach cancer patients, which incorporate other information about these individual patients, such as the patient's 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 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.

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 giving accurate information – *i.e.*, only if it is appropriate to tell a patient that s/he has a 50% chance of dying before the median survival time of this predicted curve, and a 25% chance of dying before the time associated

[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.

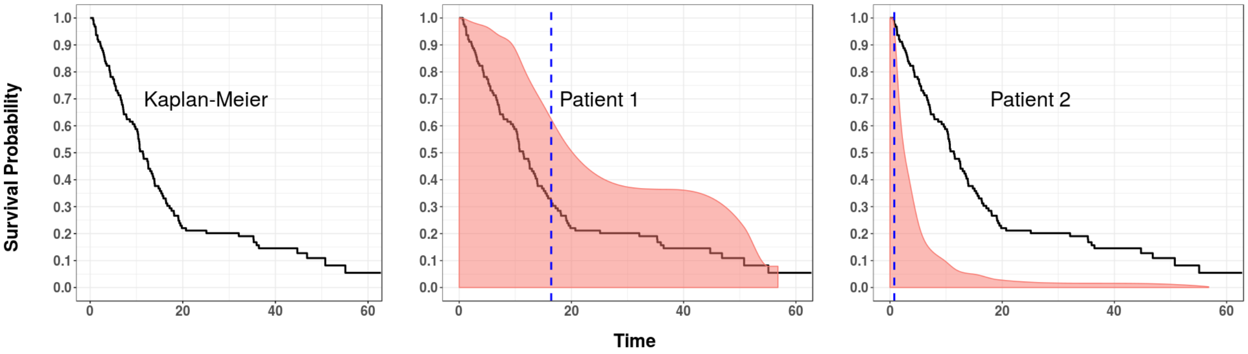


Figure 1: [left] Kaplan-Meier curve, based on 128 patients with stage-4 stomach cancer. (middle, right) Two personalized survival curves, for two patients (#1 and #2) with stage-4 stomach cancer. The blue dashed lines indicate the true time of death.

with the 25% on the curve, etc.

We focus on ways to *learn* such models from a "survival dataset" (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 she lived *at least* (say) 5 years after the starting time, but do not know whether she actually lived 5 years and a day, or 30 years. This also happens if a subject drops out of a study, after say 2.3 years, and is then lost to follow-up; 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 Table 4. 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 dataset" 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.

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 (L1-loss, Log-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 datasets, 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 and Lemeshow [26]. Wang et al. [51] 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 (with a focus on medical situations) then providing empirical results showing the strengths and weaknesses of each of the models considered.

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 dataset,

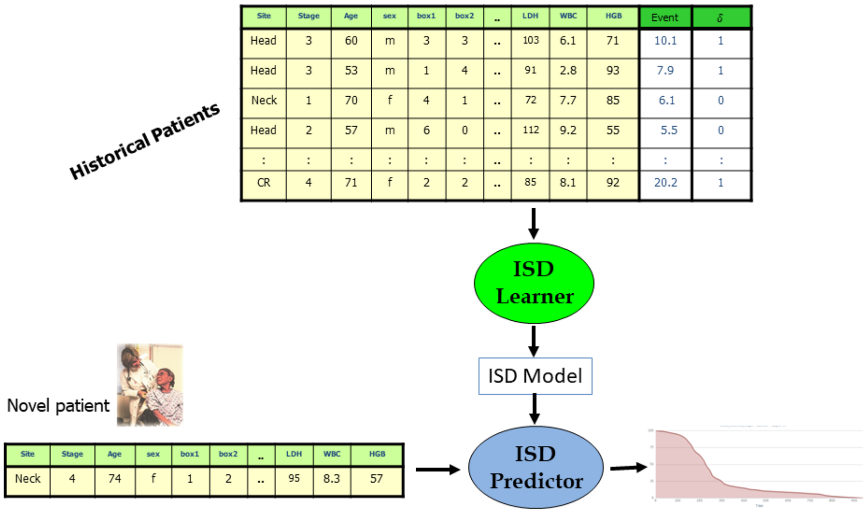
$$D \quad = \quad \{\, [\vec{x}_i,\, t_i,\, \delta_i]\,\}_i \tag{1}$$

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 2, in the context of our ISD framework.

Here, we assume $\vec{x}$ is a vector of feature values describing a patients, using information that are 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 $\delta_i = \mathcal{I}\left[\, d_i \leq c_i \,\right]$ where $\mathcal{I}\left[\,\cdot\,\right]$ 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:

[2] Throughout this work we focus on only 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".



Site	Stage	Age	sex	box1	box2	..	LDH	WBC	HGB	Event	δ
Head	3	60	m	3	3	..	103	6.1	71	10.1	1
Head	3	53	m	1	4	..	91	2.8	93	7.9	1
Neck	1	70	f	4	1	..	72	7.7	85	6.1	0
Head	2	57	m	6	0	..	112	9.2	55	5.5	0
:	:	:	:	:	:	..	:	:	:	:	:
CR	4	71	f	2	2	..	85	8.1	92	20.2	1

Site	Stage	Age	sex	box1	box2	..	LDH	WBC	HGB
Neck	4	74	f	1	2	..	95	8.3	57

Figure 2: Machine Learning paradigm for learning, then using, an ISD (Individual Survival Distribution) Model.

- *[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 $\{\ [t,\ \hat{S}(\,t\,|\,\vec{x}\,)]\ \}_{t\geq 0}$ over all future times $t \geq 0$.

- *[i vs g]* whether the result is "i" specific to a single individual patient (*i.e.*, based on a large number of features $\vec{x}$) or is "g" 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 3), giving typical uses of each, then discusses how they are interrelated.

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 $\vec{x}_a$

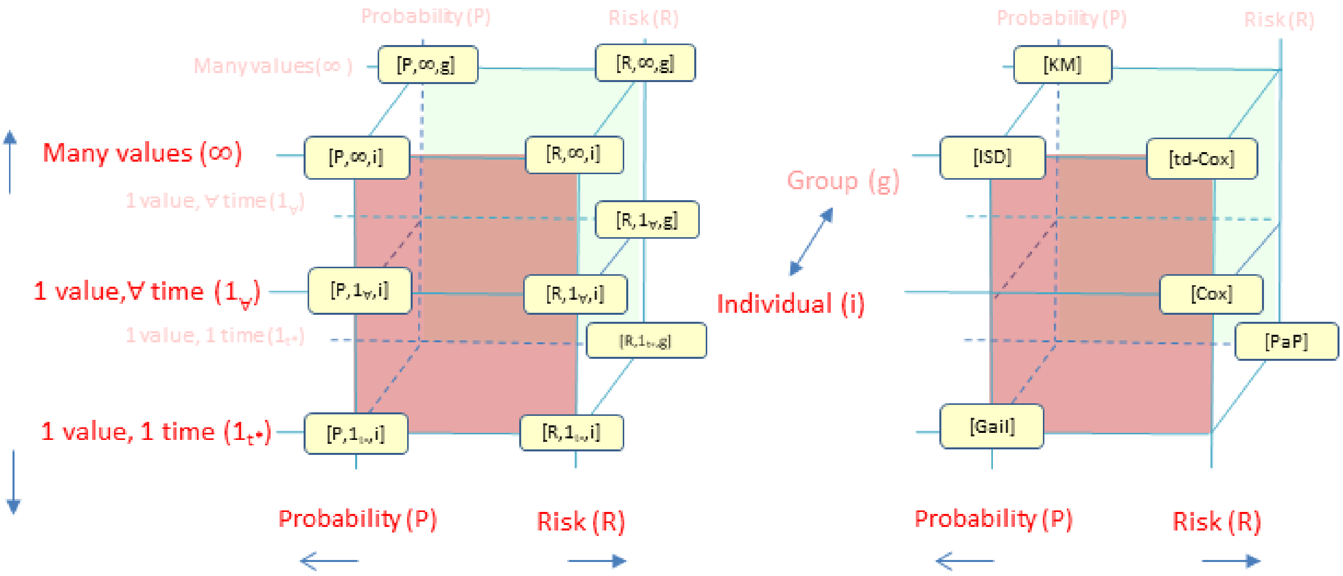


Figure 3: Dimensions for cataloging types of Survival Analysis/Prediction tools [left] – and examples of certain tools.

will die before $\vec{x}_b$. Hence, this is a *discriminative* tool for comparing pairs of patients, or perhaps for "what if" analysis of a single patient (*e.g.*, if he continues smoking, versus if 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 [10], which is used in a wide variety of applications. This models the hazard function[3] 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 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-Rosner model [8], and the myriad of others appearing on the Disease Risk Index website[4]. 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, that return a risk score associated across all time point; see Section 3.1.

[3] 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, $h(t;\vec{x}) = -\frac{d}{dt}\log S(\,t\,|\,\vec{x}\,)$.

[4] http://www.diseaseriskindex.harvard.edu/update/

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 [40]. For example, the Palliative Prognostic Index (PPI) [35] 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) [38] 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 [7, 2, 23].) Again, these tools are typically evaluated using Concordance.[5]

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] [9] [6] 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] [50] [7] 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 Apervite[8] and R-calc[9] 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

[5] 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.

[6] http://www.cancer.gov/bcrisktool/

[7] http://predictingdepression.com/

[8] https://apervita.com/community/clevelandclinic

[9] http://www.r-calc.com/ExistingFormulas.aspx?filter=CCQHS

website[10] provides a set of Kaplan-Meier curves for various cancers. While patients can use such information to estimate 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 [22] 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

- [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* $\{\ [t,\ \hat{S}(\, t \,|\, \vec{x}_i \,)]\ \}_t$ *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 [29], 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 [29] with the Weibull distribution (AFT), Random Survival Forests with the Kaplan-Meier extension (RSF-KM, described in Appendix C.3) [28] and Multi-task Logistic Regression system (MTLR) [57]. Figure 4 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,

[10] http://seer.cancer.gov/

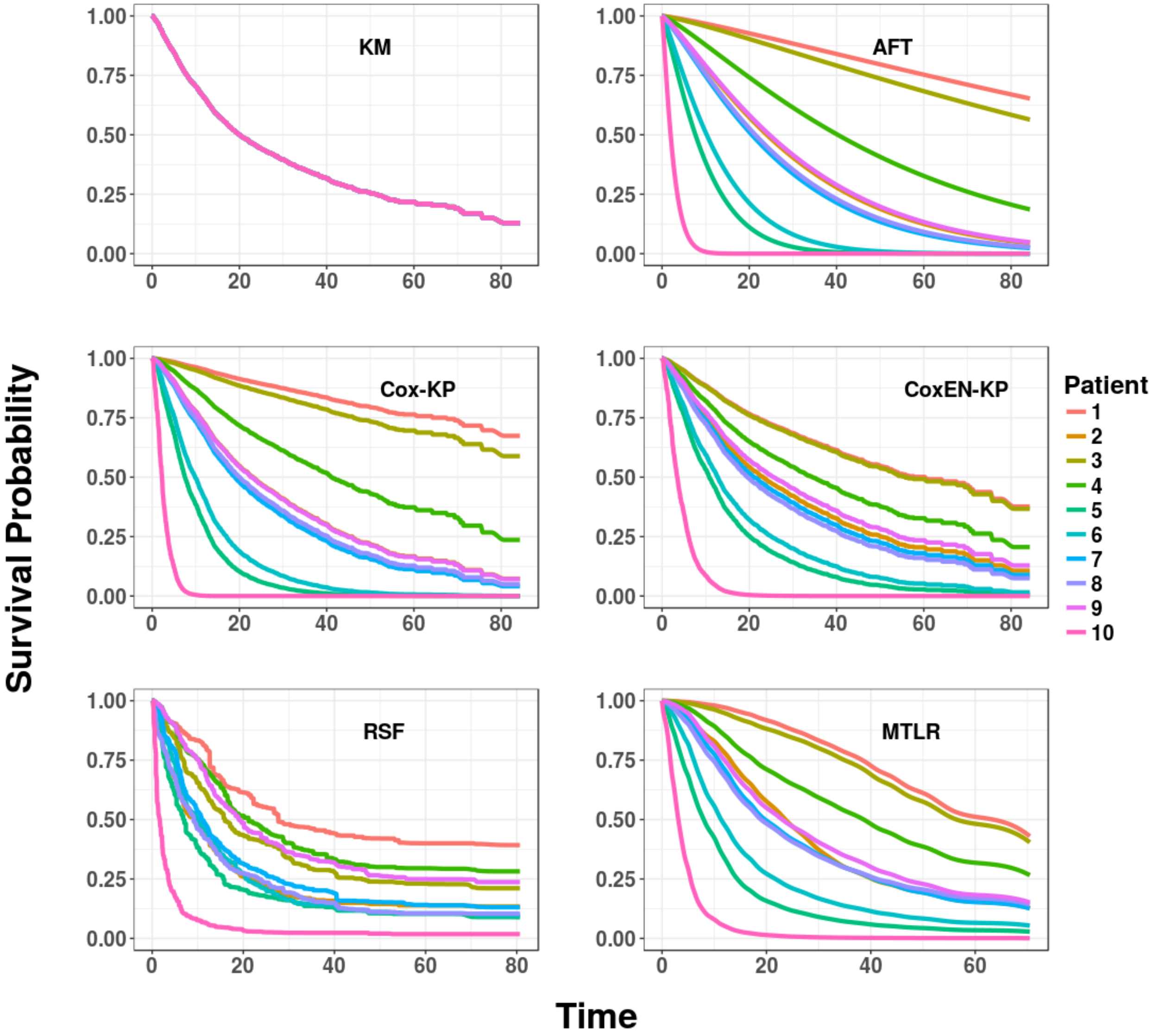


Figure 4: Survival curves (within one of 5 folds) of 10 cancer patients for the KM model and all 5 ISD models considered here, evaluated on the NACD dataset (described in Section 4.1). As the KM curve (top left) is the same for all patients by definition, we provide only 1 curve – here smoothed. Note 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.

"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 [22]. 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 [48]. (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 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.*, $\hat{S}(\, t^* \mid \vec{x}\,)$. 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%.[11] 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 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.

There are some subtle issued related to producing meaningful survival curves – *e.g.*, many curves end at a non-zero value: note the KM curve in Figure 4(top left) stops at (83, 0.12), rather than continue 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 shows 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 perform with respect to these evaluations.

For notation, we will assume models were trained on a training dataset, formed from the same triples as shown in Equation 1, that is $D = D_U \cup D_C$ where $D_U \;=\; \{\, [\vec{x}_j, d_j, \delta_j = 1] \,\}_j$

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

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".

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 dataset D is disjoint from the validation dataset, 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.*, $V = V_U = \{\, [\vec{x}_j,\, d_j,\, \delta_j = 1] \,\}_j \approx \{\, [\vec{x}_j,\, d_j] \,\}_j$ (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 $\vec{x}$, a "risk score" $r(\vec{x}) \in \Re$, 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 $\vec{x}_b$ died 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,

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

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} [33].

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 it's 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 [1] or time-dependent Cox (td-Cox) [14], which uses time-dependent features. These models can be evaluated using *time-dependent Concordance* (aka, "time-dependent ROC curve analysis") [24].

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 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, that is $\widehat{d}_i = \hat{t}_i^{(0.5)}$, leading to the following measure:

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

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 green in Figure 5), this is unrealistic.

Appendix B.2 discusses many issues with the L1-loss measure, related to censored data, and reasons to consider using the log of survival time.

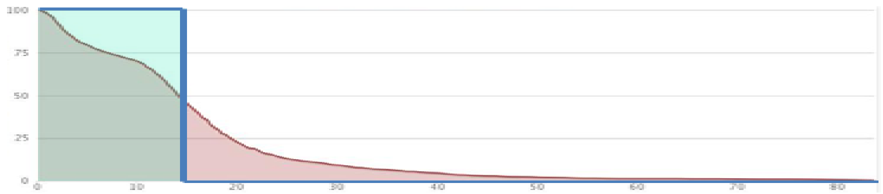


Figure 5: Example of a survival curve (in red), superimposed (in green) with a degenerate curve that puts all of its weight on a single time point (which means it assigns 100% chance of dying at exactly this time).

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 [50] 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 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 the 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 [25]. 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 6 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 dataset and t^* time). Additionally, see Appendix B.3 for an example walking through 1-Calibration.

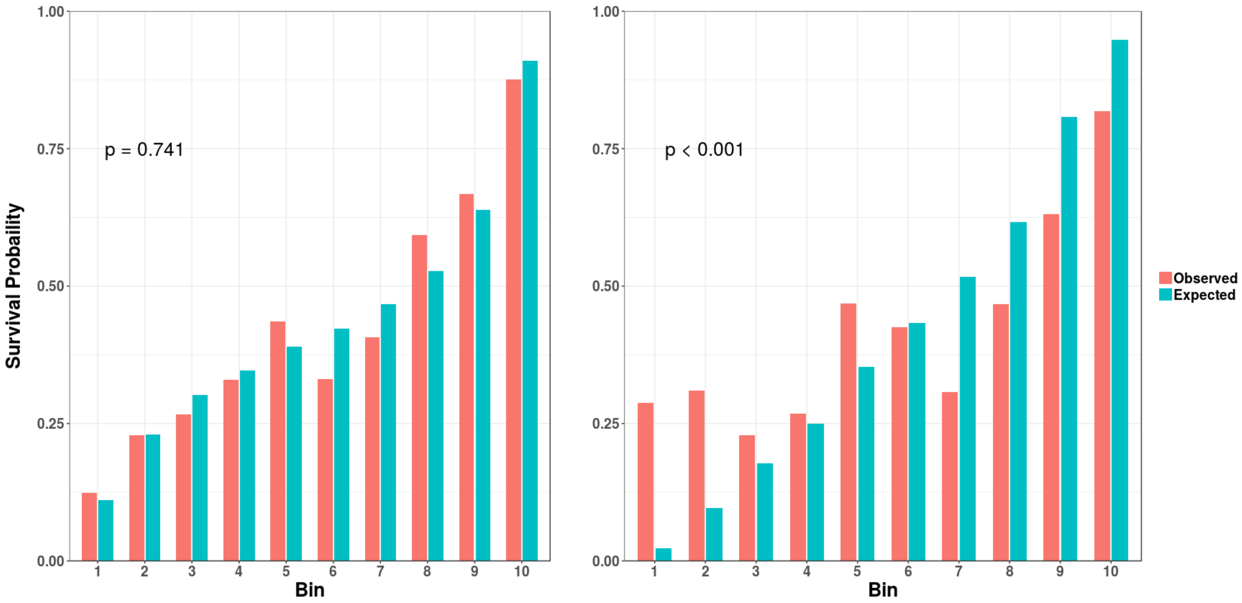


Figure 6: The bin observed and expected probabilities associated with two 1-Calibration computations, for the MTLR [left] model and the RSF-KM model applied to the GBM dataset for the 50th percentile of time (369.5 days).

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

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

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

Returning to Figure 6, 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, and $\bar{p_j}$ 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 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 ($p \geq 0.05$) and one is not ($p < 0.05$) (as p-values are not intended to be ranked). The Brier score [6] is a commonly used metric that measures both calibration and discrimination; see Appendix B.4.1. 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

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

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,

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

We will use this measure for our analysis, where τ is the maximum event time of the combined training and validation datasets – 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, and discusses the decomposition of the Brier score into calibration and discriminative components.

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 4; see Footnote 1. It would be useful to have a measure that examines the entire distribution as a distribution.[12]

A distributional calibration (D-Calibration) [3] 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

[12] 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 Appendix B.4.2 for further explanation.

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 dataset, D, and [P,∞,g]-model Θ, and any interval $[a, b] \subset [0, 1]$, let

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

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 percentile for that time, based on Θ: $\hat{S}_\Theta(\,d_i\,)$. If Θ is D-Calibrated, we expect roughly 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 bins would contain 10% of D.

This suggests a measure to evaluate a distributional model: see how close each of these 10 bins is to the expected 10%. We therefore use Pearson's χ^2 test: compute the χ^2-statistic with respect to the ten 10% intervals, and ask whether the bins appear uniform, at (say) the $p > 0.05$ level. Theorem B.2 (in Appendix B.5) states and proves the appropriateness of the Pearson's χ^2 goodness-of-fit test.

This addresses the question posed at the start of this subsection (Equation 8):

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

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?

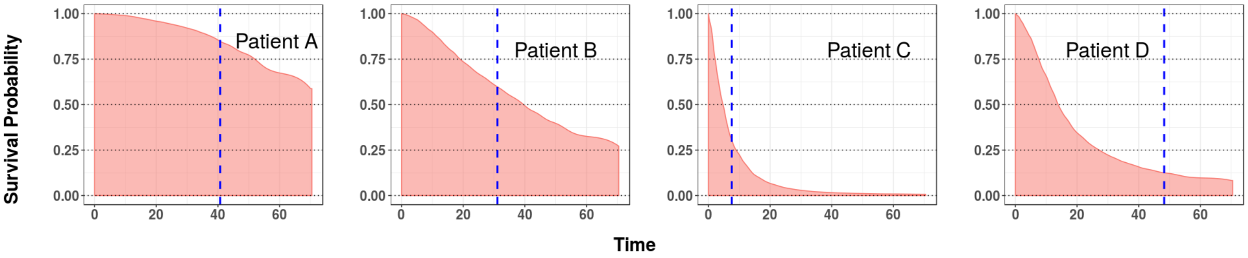


Figure 7: Four patients from the complete NACD dataset. Notice each died in a different quartile (shown with a vertical dashed line); see Table 2.

Table 2: Description of 4 patients from the NACD Dataset. (See also Figure 7)

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

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 7. 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.[13] 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.

With a slight extension to the earlier notation (Equation 9), for a dataset 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 dataset D whose time of death is assigned a probability (based on its 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%-bins", and then ask if each bin holds about 10% of the patients. The right-side of Figure 8 plots that information, for the ISD Θ

[13] Here we just consider uncensored patients; Appendix B.5 extends this to deal with censoring.

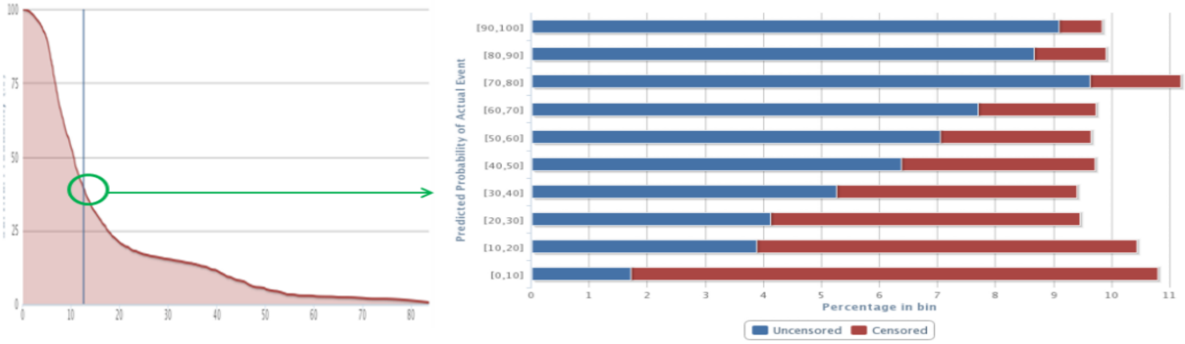


Figure 8: The right side shows the "calibration histogram" associated with the NACD dataset. The left portion shows the survival curve for a patient $\vec{x}_{27}$ – here we see that this patient's event d_{27} =12.7 months, corresponds to $\hat{S}(\, d_{27} \,|\, \vec{x}_{27}\,) \; = 39.4\%$, which 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 12.

learned by MTLR from the NACD dataset (described in Section 4.1), as a sideways histogram. We see that each of these intervals is very close to 10%. In fact, the χ^2 goodness-of-fit test yields $p = 0.433$, which suggests that this ISD is sufficiently uniform that we can believe that these survival curves are D-calibrated.

Note that Figure 8 is actually showing 5-fold cross-validation results: the survival curve 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 [53]. Also, the rust-colored intervals correspond to the censored patients; see Appendix B.5 for an explanation.

3.5.2 Relating D-Calibration to 1-Calibration

This standard notion of 1-Calibration is very 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, 4, and 7. 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 $B = 10$ bins,[14] based on the values of their associated probabilities, $\hat{S}_i(\, d_i \,|\, \vec{x}_i\,)$. Here the

[14] Note the number of bins does not have to be 10 – we chose 10 to match the typical value chosen for the 1-Calibration test.

Table 3: Summary of differences between 1-Calibration and D-Calibration.

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

goodness-of-fit test measures whether the resulting bins are approximately equal-sized, as would be expected if Θ accurately estimated the true survival curves (argued further in Appendix B.5).

Note D-Calibration tests the proportion of instances in bins 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 3, at time 20, range 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 3 summarizes the differences between D-Calibration and 1-Calibration.[15] To see that they are different, Proposition B.3, in Appendix B.5, gives a simple example of a model that is perfectly D-Calibrated but clearly not 1-Calibrated, and another example that is perfectly 1-Calibrated but clearly not D-Calibrated. In addition, we will see below several examples of this – *e.g.*, COXEN-KP is D-Calibrated for the GLI dataset, 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.

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 datasets.

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

[15] Further differences occur when considering how censored patients are handled; see Appendices B.3 and B.5.

Kaplan-Meier extension to transform RSF into an ISD is not well known, it is summarized in Appendix C.3. More recent still is the MTLR technique [57] 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) [33], some deep learning variants [39, 32, 34], and a computationally demanding Bayesian regression trees model [44], but for brevity, we focused on just the first such model, MTLR.

Note the distribution class $\mathcal{D}$ chosen for AFT certainly influences its performance – *e.g.*, it is possible that AFT[Weibull] on a dataset 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 datasets.

4.1 Datasets and Evaluation Methodology

There are many different survival datasets; here, we selected 8 publicly available medical datasets in order to cover a wide range of sample sizes, number of features, and proportions of censored patients. We excluded small datasets (with fewer than 150 instances) to reduce the variance in the evaluation metrics. Our datasets ranged from 170 to 2402 patients, from 12 to 7401 features, and percentage of censoring from 17.23% to 86.21%; see Table 4. Note that we have not included extremely high-dimensional data (with tens of thousands of features, often found in genomic datasets), as such data raises additional challenges beyond the scope of standard survival analysis; see [52] for methods to handle 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, esophagus, stomach, and other cancers. In addition to using the complete NACD dataset, we considered the subset of 950 patients with colorectal cancer (NACD-COL), with the same 53 features.

Another four datasets were retrieved from data generated by The Cancer Genome Atlas (TCGA) Research Network [15]: 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 datasets: the Dutch Breast Cancer Dataset (DBCD) [49] contains 4919 microarray gene expression levels for 295 women with breast cancer, and the Diffuse Large B-Cell Lymphoma (DLBCL) [33] dataset contains 7401 features focusing on Lymphochip DNA microarrays for 240 biopsy samples.

We applied the following pre-processing steps to each dataset: 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

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

	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
# 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

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 4 provides the dataset statistics and a full breakdown of feature numbers in each step.

Following feature selection, features were normalized (transformed to zero mean with unit variance) and passed to models for five-fold cross validation (5CV). We compute the folds by 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.

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 dataset. 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 × 8 datasets × 5 time-points); here we instead summarize the number of datasets 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, 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 [47] whereas COXEN-KP uses the *cocktail* function found in the *fastcox* package [56]. Both RSF and RSF-KM come from the *randomForestSRC* package [27]. An implementation of MTLR (and of all the code used in this analysis) is publicly available on the GitHub account[16] of the lead author.

4.2 Empirical Results

Below, we consider a dataset 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 datasets: GBM, NACD-COL, GLI, NACD – which are shown first in all of our empirical studies. We let "HIGH-CENSOR" datasets refer to READ and BRCA and "HIGH-DIMENSIONAL" datasets refer to the other two (DLBCL and DBCD).

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

Figures 9, 10 and 11 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 dataset, 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 dataset; here we find that MTLR performs best on a majority of datasets: six of eight for Concordance and L1-loss, and seven of eight for the Integrated Brier score.

NICE Datasets: Recall that the first 4 datasets are NICE. Here, we find that most models performed comparably – and 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 datasets GLI and GBM for Integrated Brier score, and for those datasets and also NACD-COL for L1-loss.)

HIGH-CENSOR Datasets – 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).

[16] https://github.com/haiderstats/ISDEvaluation

Table 5: Results from 1-Calibration evaluations. Columns represent percentiles used for each time point and rows indicate the model used. Recall there are 8 datasets – meaning no model performed perfectly for any of the percentiles.

	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**

High-Dimensional Datasets – DBCD and DLBCL: There are no entries for COX-KP for these two datasets 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 datasets – 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 5 gives the number of datasets 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 datasets 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 datasets, COXEN-KP is 1-Calibrated for two, and MTLR is 1-Calibrated for four. The 75th percentile also showed to be 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 datasets 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 datasets 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.

4.2.3 D-Calibration Results

Table 6, which gives the D-Calibration p-values for each model and dataset, shows that both KM and MTLR pass D-Calibration for every dataset, with KM receiving the highest possible p-value, p =1.000, for each. (In fact, Lemma 2 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-

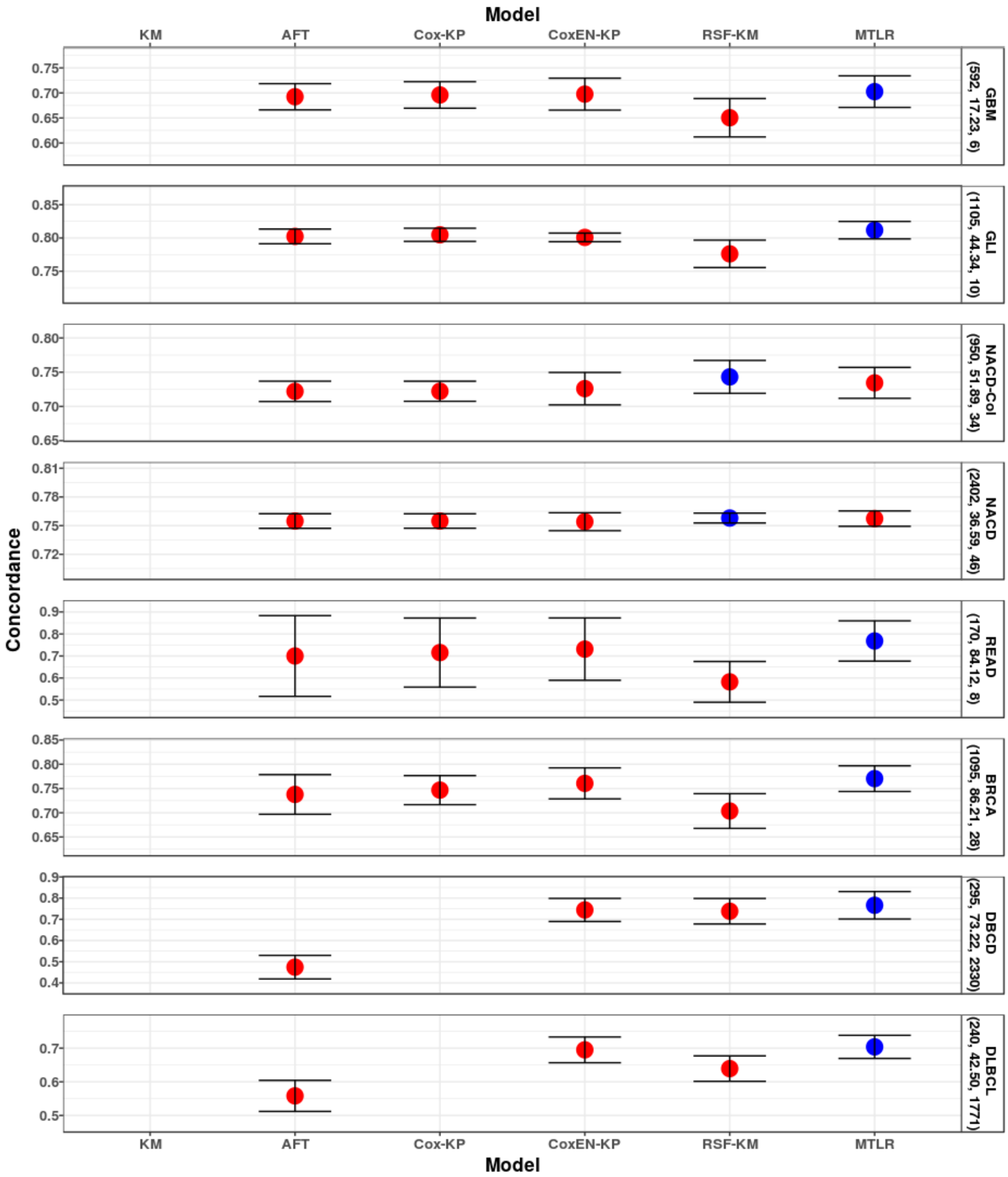


Figure 9: Concordance means and one standard deviation are given by circles and error bars, respectively. The best (highest Concordance) scoring model is given in blue; all other models are in red. We included KM so this figure would "line up" with Figures 10 and 11, but left the value blank, as the Concordance scores for KM are always 0.5. For these 3 figures: As COX-KP failed to run for datasets DBCD and DLBCL, those entries are blank. The description at the right gives the name of the dataset, and the 3 numbers "under" each dataset name are (f_{final}, % Censored, N).

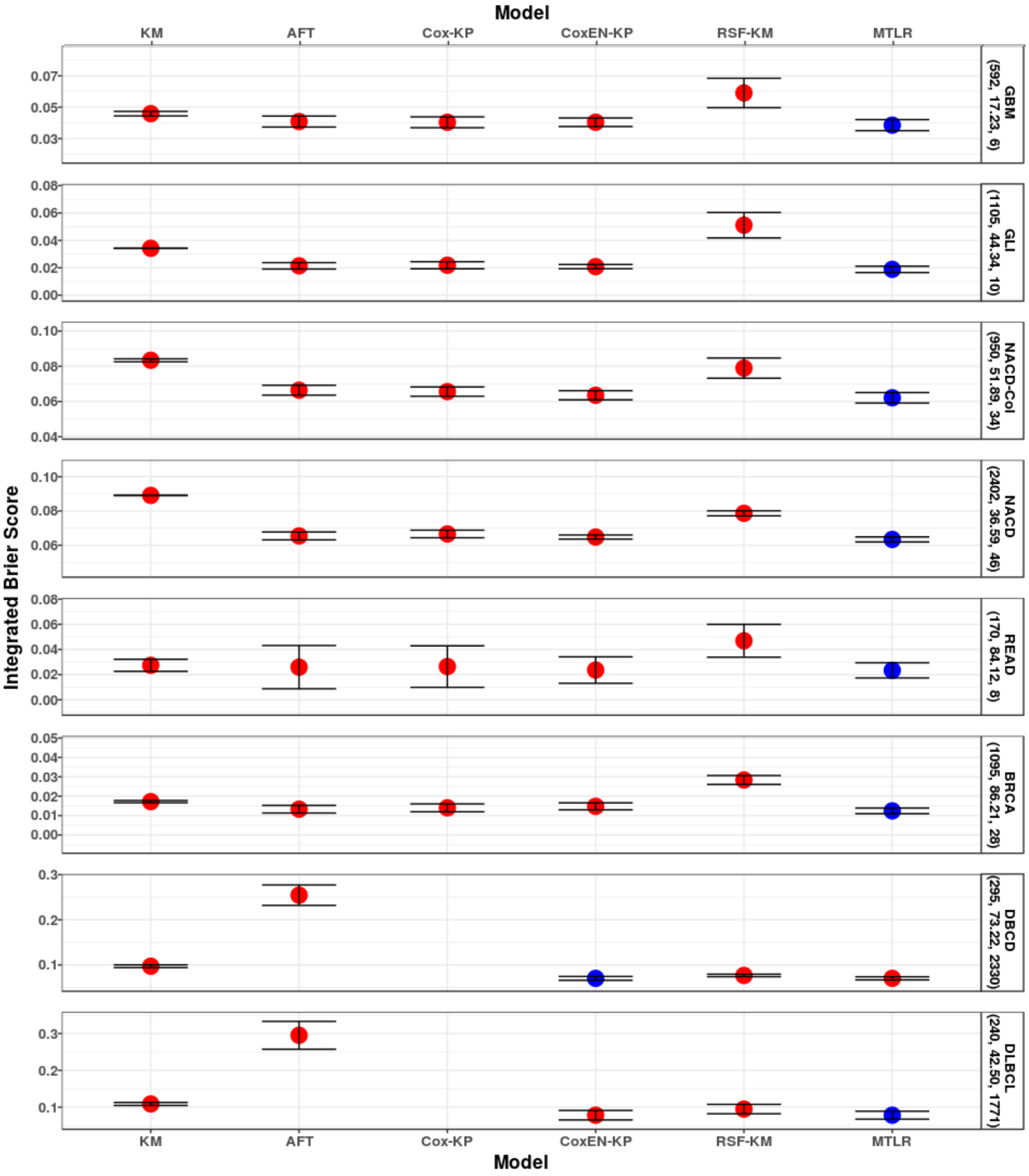


Figure 10: Integrated Brier score means and one standard deviation are given by circles and error bars, respectively. The best (lowest Integrated Brier score) scoring model is given in blue; all other models in red.

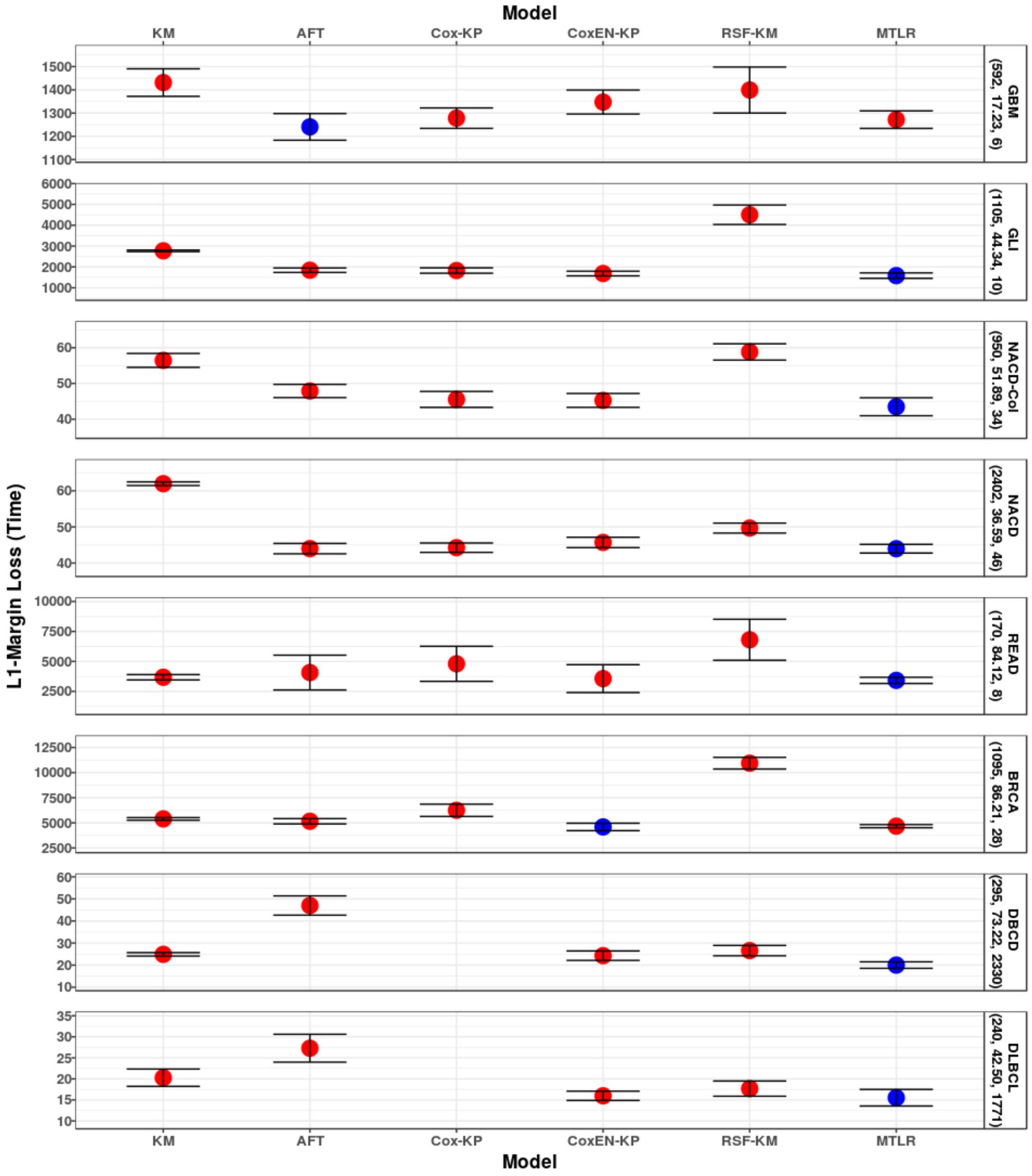


Figure 11: L1-loss means and one standard deviation are given by circles and error bars, respectively. The best (lowest L1-loss) scoring model is given in blue, all other models in red. As different datasets use different time units, we simply give the units of L1-loss as "Time" rather than days/months/years.

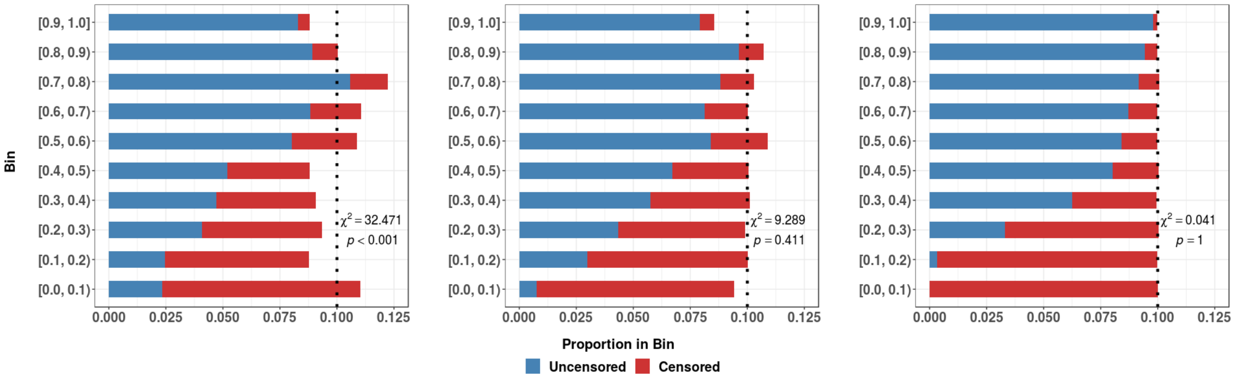


Figure 12: These figures show the (sideways) decile histogram used for the D-Calibration test. Each of these is run on the NACD dataset; from left to right: running COX-KP, MTLR and KM.

Table 6: Results for D-Calibration evaluations. Columns correspond to the dataset 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.

	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

Calibrated for all datasets, but in addition, it also provides each patients with his/her own survival curve.

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

Figure 12 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 8.

5 Discussion

Comparing different isd-models: Steyerberg *et al.* [46] 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 [30]. 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); and 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 datasets 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]" [45].) However, when we look at the Nice datasets, 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 datasets, more complex models (mtlr, rsf-km, and coxen-kp) do not offer large benefits in terms of Concordance. For the High-Dimensional datasets, 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 datasets, this suggests that rsf-km may not be optimal for these high-dimensional datasets, in terms of Concordance. For the High-Censor datasets rsf-km saw much worse performance for Concordance (among other metrics) suggesting rsf-km may not be suitable for datasets 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 7, 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 and 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, that 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 dataset. 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 datasets, 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 datasets 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 datasets. (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 datasets. As discussed in Section 4.2, AFT is a completely parametric model, which means it cannot produce different shapes (see Figure 4[top-right]), likely impacting its ability to be D-Calibrated. (Our analysis showed only that AFT[Weibull] is here not D-Calibrated; AFT[χ] for some other distribution class χ, might be D-Calibrated for more datasets.)

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 and Appendix B.4. We see MTLR performing the best for seven of the eight datasets, however, MTLR is no longer superior for DBCD, one of the high-dimensional datasets, even though it was superior for Concordance. Instead, COXEN-KP, RSF-KM, and MTLR all perform nearly identical for these HIGH-DIMENSIONAL datasets.

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 datasets – 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 dataset – MTLR and COXEN-KP were 1-Calibrated for 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 datasets.

RSF-KM was not 1-Calibrated across any percentiles for GBM, GLI, 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 datasets. 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 [43]). 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 dataset where MTLR has the best performance with an average L1-loss of 43.97 months. While this is the lowest average error, predicting the time of death up to an error of 43.97 months ($\approx$3.7 years) is likely not helpful to a patient, especially as the maximum follow-up time was 84.3 months.

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 datasets, although all ISD models performed comparably for the four NICE datasets (with the exception of RSF-KM). We see that KM is also competitive for the HIGH-CENSOR datasets, but given the construction of the L1-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 datasets.

We also compared the models in terms of "Uncensored L1-loss", which just considers the loss on the uncensored instances; see Table 11 in Appendix D.3. We see KM is no longer competitive for the HIGH-CENSOR datasets, showing how influential this effect is. Instead, at least one of the complex models {COXEN-KP, RSF-KM, MTLR} outperforms AFT and COX-KP for every dataset.

That appendix also motivates and defines the Log L1-loss, and its Table 12 shows that MTLR performs best in 4 of the datasets, and is either second or third best in the others.

Which ISD-Model to Use?: As shown above, which ISD-model works best depends on properties of the dataset, and on what we mean by "best". Table 7 summarizes our results here.

In general, for NICE datasets, MTLR was superior for calibration but for discrimination, all ISD-models were equivalent, leading us to recommend using the simplest models: (COX-

Table 7: Our recommendation for ISD models, for different types of datasets. Note we divide the High-Dimensional set into Low versus High censoring. (DBCD is 73.22% censored.)

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

KP, AFT). As we found that RSF-KM would overfit 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 datasets, 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 datasets. For High-Dimensional datasets with low censoring (less than 70% *i.e.*, 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 examine if these findings hold in general, we examined 33 other public datasets – 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 datasets were taken from the TCGA website whereas the other (High Dimensional) datasets arise from a variety of sources. The results from these 33 datasets are consistent with the findings reported here; specific results can be found on the lead author's RPubs site[17]. Given the low overall number of High-Dimensional datasets, these findings should be examined on further datasets.

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 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 4 for the MTLR model. Here, we would optimistic about this patient if we considered

[17] See http://rpubs.com/haiderstats/ISDEvaluationSupplement

the single point probability model at $t^* = 6$months, as $\hat{S}_{MTLR}(\, P5 \,|\, 6\text{months}\,) = 0.8$, but very concerned if we instead used $t^* = 12$months, as $\hat{S}_{MTLR}(\, P5 \,|\, 12\text{months}\,) = 0.3$. 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 4[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 a 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 dataset, 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 all instances in the training data are described using a fixed number of features (see the matrix in Figure 2), there is no missing values, and 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 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 handle left-censored instances (where the dataset 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 primarily ISD (individual survival distribution; [P,∞,i]) models. We discussed the advantages of having an individual survival distribution for each patient, as this can help patients and clinicians to 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 should primarily be evaluated using calibration metrics (Sections 3.3, and 3.5) as these measure the extent to which 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 datasets – over a range of #features, #instance 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.

In conclusion, this paper explains why we encourage researchers, and practioners, 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, Amii, and Borealis AI (of RBC). We also thank Adam Kashlak for his insightful discussions regarding D-Calibration.

References

[1] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view.* Springer Science & Business Media, 2008.

[2] F. Anderson, G. M. Downing, J. Hill, L. Casorso, and N. Lerch. Palliative performance scale (pps): a new tool. *Journal of palliative care*, 12(1):5–11, 1995.

[3] A. Andres, A. Montano-Loza, R. Greiner, M. Uhlich, P. Jin, B. Hoehn, D. Bigam, J. A. M. Shapiro, and N. M. Kneteman. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS one*, 13(3):e0193523, 2018.

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

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

[6] G. W. Brier and R. A. Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

[7] R.-B. Chuang, W.-Y. Hu, T.-Y. Chiu, and C.-Y. 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.

[8] G. A. Colditz and B. 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.

[9] J. P. Costantino, M. H. Gail, D. Pee, S. Anderson, C. K. Redmond, J. Benichou, and H. S. 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.

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

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

[12] R. d'Agostino and B.-H. Nam. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

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

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

[15] Genome Data Analysis Center. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.* Broad Institute of MIT and Harvard.

[16] T. A. Gerds, T. Cai, and M. Schumacher. The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4):457–479, 2008.

[17] T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

[18] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

[19] D. Guffey. *Hosmer-Lemeshow goodness-of-fit test: Translations to the Cox Proportional Hazards Model*. PhD thesis, 2013.

[20] R. C. Gupta and D. M. Bradley. On representing the mean residual life in terms of the failure rate. *arXiv preprint math/0411297*, 2004.

[21] 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.

[22] D. Harrington. Linear rank tests in survival analysis. *Encyclopedia of Biostatistics*, 2005.

[23] J. Haybittle, R. Blamey, C. Elston, J. Johnson, P. Doyle, F. Campbell, R. Nicholson, and K. Griffiths. A prognostic index in primary breast cancer. *British journal of cancer*, 45(3):361, 1982.

[24] P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

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

[26] D. W. Hosmer, S. Lemeshow, and S. May. *Applied survival analysis.* Wiley Blackwell, 2011.

[27] H. Ishwaran and U. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. R package version 2.6.1.

[28] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2:841–860, 2008.

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

[30] P. S. Kamath and W. R. Kim. The model for end-stage liver disease (meld). *Hepatology*, 45(3):797–805, 2007.

[31] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[32] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *arXiv preprint arXiv:1606.00931*, 2016.

[33] Y. Li, J. Wang, J. Ye, and C. 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.

[34] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*, 2017.

[35] T. Morita, J. Tsunoda, S. Inoue, and S. 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.

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

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

[38] M. Pirovano, M. Maltoni, O. Nanni, M. Marinari, M. Indelli, G. Zaninetta, V. Petrella, S. Barni, E. Zecca, E. 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.

[39] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

[40] M. P. Rogers, J. Orav, and P. M. 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.

[41] S. Saks. Theory of the integral. 1937.

[42] F. Sanders. On subjective probability forecasting. *Journal of Applied Meteorology*, 2(2):191–201, 1963.

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

[44] R. A. Sparapani, B. R. Logan, R. E. McCulloch, and P. W. Laud. Nonparametric survival analysis using bayesian additive regression trees (BART). *Statistics in medicine*, 35(16):2741–2753, 2016.

[45] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems*, pages 1209–1216, 2008.

[46] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[47] T. M. Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

[48] T. M. Therneau and P. M. Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

[49] H. C. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, 25(18):3201–3216, 2006.

[50] J. Wang, J. Sareen, S. Patten, J. Bolton, N. Schmitz, and A. 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.

[51] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*, 2017.

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

[53] I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[54] G. Yan and T. Greene. Investigating the effects of ties on measures of concordance. *Statistics in medicine*, 27(21):4190–4206, 2008.

[55] Y. Yang and H. 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.

[56] Y. Yang and H. Zou. *fastcox: Lasso and Elastic-Net Penalized Cox's Regression in High Dimensions Models using the Cocktail Algorithm*, 2017. R package version 1.1.3.

[57] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*, 2011.

A Extending Survival Curves to 0

In practice, survival curves often stop at a non-zero probability – see Figure 4 and Figure 13[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 datasets) 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 13[right]. Note that curves cannot cross within the extended interval, which means this extension will not change the discriminatory criteria.

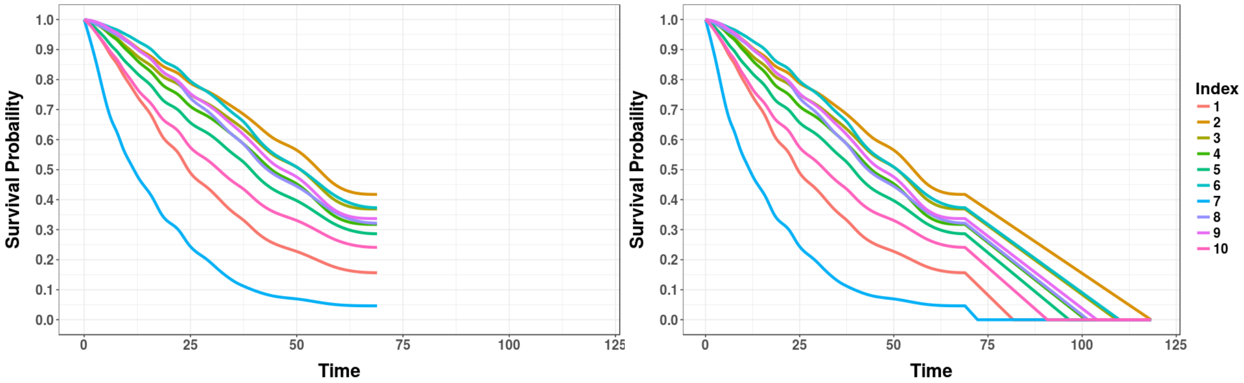


Figure 13: On left, survival curves generated from MTLR for the Nacd-Col dataset. Left shows this model's survival curves end at 68.9 months. On right, linear extensions of those survival curves go as far as 118 months.

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 datasets. 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}.

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. 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*, which is the subset of pairs of indices (here using the validation dataset (V) and recalling that $\delta = 1$ indicates a patient who died (uncensored)) containing all pair of instances when we know which patient died first:

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

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 14. 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 index 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 [54]. 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.

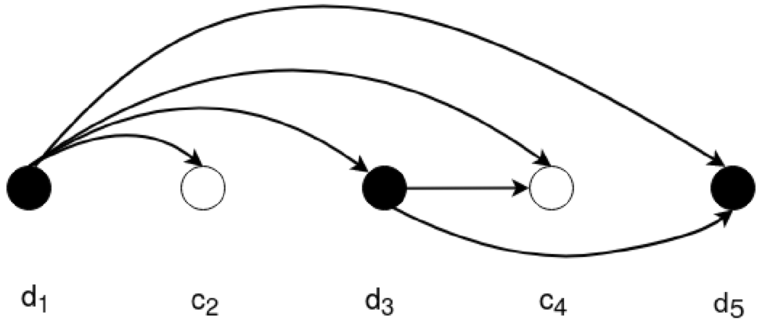


Figure 14: Depiction of Concordance comparisons, including censored patients. Black and white circles indicate uncensored and censored patients, respectively. Each d_i is the death time for an uncensored patient, and each c_j is the censoring time for a censored patient. We can only compare: uncensored 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 [51].

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 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. Table 11 in Appendix D.3 provides the results for this Uncensored L1-loss over the 8 datasets. (We see that MTLR is best for 6 of these datasets.)

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:

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

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

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

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 datasets 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 dataset (disjoint from the validation dataset) 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

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

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 11, with details in Table 10.

For completeness, we prove Equation 16. (This claim is also proven by Gupta and Bradley [20], 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

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

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

□

Step 18 is an application of Tonelli's theorem [41], 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

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

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

B.2.1 Log L1-loss

The L1-loss measure implicitly assumes that the quality of a prediction, $\hat{t}_j^{(0.5)}$, depends only on how close it is to the truth d_j– *i.e.*, on $|d_j - \hat{t}_j^{(0.5)}|$. But this does not always match how we think of the error: if we predict Patient A will live for 120 months then found that he actually lived 117 months, we would consider our prediction very accurate. By contrast, if we predict Patient B will live 1 month, but then find she lived 4 months, we would consider this to be a poor prediction. Notice, however, the L1-loss for Patient A is $|d_A - \hat{t}_A^{(0.5)}| \;=\; |120 - 117| = 3$ months, which is the same as the L1-loss for Patient B: $|d_B - \hat{t}_B^{(0.5)}| \;=\; |1 - 4| \;= 3$ months!

This motivates us to consider the *relative* error, rather than an *absolute* error: here, as our prediction for Patient A is off by only 3 / 120 = 2.5%, we consider it good, whereas our prediction for Patient B is off by 3 / 1 = 300%. The Log-L1-loss reflects this:[18]

$$\ell_{LogL1}(\, d_i,\, \hat{t}_i^{(0.5)}\,) \;=\; |\log(d_i) - \log(\hat{t}_i^{(0.5)})| \tag{19}$$

To compute the average Log-L1-loss over the dataset V_U, we can use Equation 4 but using $\log(d_j)$ rather than d_j, etc. To avoid taking $\log 0$, we replace 0 with half the minimum, positive death time (see Section B.6). Table 12 in Appendix D.3 provides the results here, over the 8 datasets. (We see that MTLR is best for 4 of these datasets.)

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.

[18] Note that the times mentioned in “Doc, do I have a day, a week, a month or a year?” are basically in a log-scale.

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 [19]; we use the D'Agostino-Nam translation [12], which uses the *within bin* Kaplan-Meier curve in place of O_j. Specifically, the test statistic is given by,

$$\widehat{HL}_{DN}\,(\,V,\ \hat{S}(\,t^*\,|\,\cdot\,)\,) \quad = \quad \sum_{j=1}^{B} \frac{(\ n_j\ (1 - KM_j(t^*))\ -\ n_j\,\bar{p}_j)^2}{n_j\,\bar{p}_j\,(1-\bar{p}_j)}, \tag{20}$$

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.

B.4 Brier Score Details

This section supplements the description of the Brier score given in Section 3.4, discussing (1) the decomposition of the Brier score into calibration and discrimination components, (2) the failure of the Integrated Brier score to incorporate the full distribution of probabilities in survival curves, and (3) how to incorporate censoring into the Brier score.

B.4.1 Brier Score Decomposition

As mentioned in Section 3.4, the Brier score can be separated into calibration and discriminatory components. The original separations were the the work of Sanders [42] and Murphy [36, 37] and later put into the context of calibration and discrimination (also known as refinement) by DeGroot and Fineberg [13].

Recall the notation and mathematical expression of the Brier score for a set of uncensored instances, V_U,

$$BS\Big(\ \hat{S}(\,t^*\,|\,\cdot\,),\ \{\vec{x}_i\}\ \Big) \quad = \quad \frac{1}{|V_U|}\sum_{i\in V_U}\Big(\mathcal{I}\,[\,d_i \le t^*\,] - \hat{S}(t^*|\vec{x}_i)\Big)^2.$$

To simplify notation, let $p_i = \hat{S}(\,t^*\,|\,\vec{x}_i\,)$. The separation of the Brier score requires that a discrete, distinct number of predictions exist; here, assume there are K distinct values for p_k for $k = 1, \ldots K$.

Further, let n_k be the total number of patients with p_k as their prediction and hence $|V_U| = \sum_{k=1}^{K} n_k$. Finally, let λ_k be the observed proportion of patients who have died by t^* and thus $(1 - \lambda_k)$ is the proportion still alive. The separation theorem of the Brier score states that $BS = C + D$, where C and D are nonnegative calibration and discriminatory scores where

$$C = \frac{1}{|V_U|} \sum_{k=1}^{K} n_k(\lambda_k - p_k)^2 \tag{21}$$

$$D = \frac{1}{|V_U|} \sum_{k=1}^{K} n_k \lambda_k (1 - \lambda_k). \tag{22}$$

Note the calibration score, C, is nearly equivalent (up to a factor of n_k) to the numerator of the Hosmer-Lemeshow test (Equation 5). However, the Hosmer-Lemeshow test subscript refers to bins whereas here the subscript refers to a distinct value of p_k. One can see that C represents a calibration score as the estimated probabilities, p_k, must be close to the true proportion of deaths, λ_k in order to have a small score (lower is better). In fact, to satisfy $C = 0$, all predictions, p_k must be equal to λ_k (Equation 21).

There are also similarities between D and the denominator of the Hosmer-Lemeshow test. However, note Equation 22 uses the the true proportion of deaths λ_k, whereas the Hosmer-Lemeshow test uses an estimated value, $\bar{p}$. Note that D has a "good" (low) score if all patients associated with a prediction probability p_k have the same status – *i.e.*, they either all die or are all still alive. To understand why this means D is a discriminatory measure, consider the extreme case where $BS(\cdot, \cdot) = 0$, which means both $D = 0$ and $C = 0$. For $D = 0$, all patients associated with each probability value must either be dead by t^* or all be alive at t^* – *i.e.*, $\lambda_k \in \{0, 1\}$ for $k = 1, 2$; note only $K = 2$ is possible here. In turn, for $C = 0$, we require $p_k = \lambda_k$ for $k = 1, 2$, that is $p_k \in \{0, 1\}$ – all predictions will be 1 or 0. Here we are discriminating perfectly between the patients who have died and the patients who are still alive, with a model that predicts only 1's or 0's. Of course, we should not require a model to estimate survival probabilities to be precisely 1 or 0, for the same reason that we do not expect the learned distribution to correspond to the Heaviside distribution shown in Figure 5.

B.4.2 Integrated Brier score does not involve the Entire Distribution

At the beginning of Section 3.5, we claimed the Integrated Brier score (IBS) does not utilize the survival curves' full distribution of probabilities over all times. For example, on a KM curve, we expect that 10% of patients will die in every 10% interval – *e.g.*, 10% of all patients will die in the [0.5, 0.6) interval. While D-Calibration will debit a model that fails to do this, this Integrated Brier score does not require this. The most obvious example is the perfect model, where each patient is given the appropriate Heaviside distribution (Figure 5) at his/her time-of-death: here the only probabilities are {0,1} – here IBS$(\cdot, \cdot) = 0$, even though no patient's $\hat{S}_{Heaviside}(\, d_i \,|\, \vec{x}_i\,)$ is ever in [0.5, 0.6). However, as we have previously noted, the inherent stochasticity of the world means that meaningful distributions should include non-zero probabilities in other places as well, rather than placing all weight on a single time point.

Since the Integrated Brier score fails to account for this, there is no guarantee that probabilities are meaningful across individual survival curves. This motivated us to introduce D-Calibration, to determine whether a proposed ISD-model produces meaningful distributions, with probabilities that reflect the number of deaths that have occurred in the population. To see that these two metrics are measuring different aspects, note that the Integrated Brier scores for the (AFT, COX-KP, COXEN-KP, and MTLR) models are all well within 1 standard error of one another for the GBM dataset, but only COXEN-KP and MTLR are D-Calibrated. (This is also true for the GLI dataset.)

B.4.3 Incorporating Censoring into the Brier score

In 1999, Graf *et al.* [18] 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.* [16, 17] formalizes and proves this intuition.

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

$$BS_{t^*}\left(V,\ \hat{S}(t^*|\cdot)\right) \;=\; \frac{1}{|V|}\sum_{i=1}^{|V|}\left[\frac{\mathcal{I}\left[\,t_i \le t^*, \delta_i = 1\,\right]\left(0 - \hat{S}(t^*|\vec{x}_i)\right)^2}{\hat{G}(t_i)} + \frac{\mathcal{I}\left[\,t_i > t^*\,\right]\left(1 - \hat{S}(t^*|\vec{x}_i)\right)^2}{\hat{G}(t^*)}\right], \tag{23}$$

where $t_i = \min\{d_i, c_i\}$. The first part of Equation 23 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 justifying why, in the case of all uncensored patients, (1) the distribution of the survival function, $\{S(t)\}_t$, should follow a uniform distribution, then (2) Following this discussion, we show how to incorporate censored patients into the D-Calibration estimate, and finally, (3) that this combination of censored and uncensored patients will produce a uniform distribution for the goodness-of-fit test to test against.

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. We also assume that censoring time is independent of death time, $c_i \perp d_i$. 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 1. *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 [4] 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, t, we have $F(t) \sim U(0,1)$. As the survival function is simply $S(t) = 1 - F(t)$, its distribution is $1 - U(0,1)$, which also follows $U(0,1)$ and hence $S(t) \sim U(0,1)$. □

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. 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 are monotonically decreasing 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) bin – *i.e.*, $P\,(\;S(d_i|\vec{x}_i) \in [0.2, 0.3)\;|\;S(d_i|\vec{x}_i) \leq 0.25)$. Using the rules of conditional probability, this is computationally straightforward[19]:

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

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

$$\begin{aligned}
P(\,S(d_i) \in [0.1, 0.2)\,|\,S(d_i) < 0.25\,) &= \frac{P(S(d_i) \in [0.1, 0.2),\; S(d_i) < 0.25)}{P(S(d_i) < 0.25)} \\
&= \frac{P(S(d_i) \in [0.1, 0.2))}{P(S(d_i) < 0.25)}
\end{aligned}$$

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

$$\begin{aligned} &= \frac{0.1}{0.25} && \text{(as } S(\cdot) \sim U(0,1)) \\ &= 0.4 \end{aligned}$$

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

This example motivates the following procedure to incorporate censored patients into the D-Calibration process: Given B bins that equally divide [0,1] into intervals of width $BW = 1/B$, suppose a patient is censored at time c with associated survival probability $S(c)$. Let b_1 be the infimum probability of the bin 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 bins:

(A) Bin $[b_1, b_2)$ (which contains $S(c)$): $\frac{S(c)-b_1}{S(c)} = 1 - \frac{b_1}{S(c)}$

(B) All following bins (*i.e.*, the bins whose survival probabilities are all less than b_1): $\frac{BW}{S(c)} = \frac{1}{B \cdot S(c)}$,

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

To further illustrate this concept of blurring a patient across bins, consider a patient who is censored at $t = 0$ with $S(c_i) = 1$. This patient is then blurred across all ($B = 10$) bins, adding a weight of 0.1 to all 10 bins. 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 bin.

This identifies a weakness of D-Calibration: if a validation set contains N_0 patients censored at time 0, then all bins are given an equal weight of N_0/B; if N_0 is large relative to the total number of patients, then the bins 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.

To perform the goodness-of-fit test, we must first calculate the observed proportion of patients within each bin. Let N_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. We can formally calculate:

$$\begin{aligned} N_k \;=\; \frac{1}{|V|} \sum_{i=1}^{|V|} \Bigg[\quad & \mathcal{I}\,[\, S(d_i) \in [p_k, p_{k+1}) \;\wedge\; d_i \leq c_i \,] && (24) \\ + \quad & \frac{S(c_i) - p_k}{S(c_i)} \cdot \mathcal{I}\,[\, S(c_i) \in [p_k, p_{k+1}) \;\wedge\; c_i < d_i \,] && (25) \\ + \quad & \frac{(p_{k+1} - p_k)}{S(c_i)} \cdot \mathcal{I}\,[\, S(c_i) \geq p_{k+1} \;\wedge\; c_i < d_i \,] \Bigg]. && (26) \end{aligned}$$

Above, (24) refers to the weight that the patients with observed events contribute to the k^{th} bin – *i.e.*, each uncensored patient whose survival probability at time of death lands in $[p_k, p_{k+1})$ contribute a value of 1. Here, we consider $d_i = c_i$ to be an uncensored event. Next, (25) gives the weight from the censored patients whose survival probability at time of censoring is within the k^{th} bin (item (A) above). Lastly, (26) gives the weights from censored patients whose survival probability was contained in a previous bin (item (B) above).

Theorem B.2 below proves that the expected value of N_k is equal for all bins – *i.e.*, $\mathbb{E}[N_k] = p_{k+1} - p_k$ – 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 equality, $d_i \leq c_i \iff S(d_i) \geq S(c_i))$. This equivalence lets us replace $d_i \leq c_i$ with $S(d_i) \geq S(c_i)$, within the indicator functions in N_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 (24) – (26) above is equal to $p_{k+1} - p_k$ – *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.2. *Given the formula for N_k (Equations (24) - (26)), if the true survival function $S(\cdot|\cdot)$ is strictly monotonically decreasing then proportions are equal across all bins –* i.e., $\mathbb{E}[N_k] = p_{k+1} - p_k$.

Proof.

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

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

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

$$= \Pr[S_d \in I_k \;\wedge\; S_d \geq S_c] \quad + \quad \Pr[S_c \in I_k \;\wedge\; S_c > S_d] \tag{I}$$

$$- p_k \, \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\left[\, S_c > S_d \,\wedge\, S_c \geq p_k \,\right]\right] \tag{II}$$

$$+ p_{k+1} \mathbb{E}\left[\frac{1}{S_c} \cdot \mathcal{I}\left[\, S_c > S_d \,\wedge\, S_c \geq p_{k+1} \,\right]\right] \tag{III}$$

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 piece 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 (1),

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

Recalling the independence assumption, $c \perp d$, we have the following equalities:

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

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

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

Now we address line (II) and analagously line (III):

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

Here f 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 $f(S_d)$ would be 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,

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

□

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, which 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 bin (specifically higher proportions within bins that contain these flat lines).

A natural corollary of Theorem B.2 is that all consistent estimators of the true survival distribution will be D-Calibrated (if the true survival distribution is strictly monotonic). Further, 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, as Kaplan-Meier is uniformly consistent [5, 11]:

Lemma 2. *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 datasets. Under all uncensored data, we would expect the typical 5% Type I error rate for claiming p <0.05 as significant, however in the presence of censored data a correct estimate of the survival distribution the proportion within bins become smoothed, boosting the p-value.

Proposition B.3. *It is possible for a* 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.*

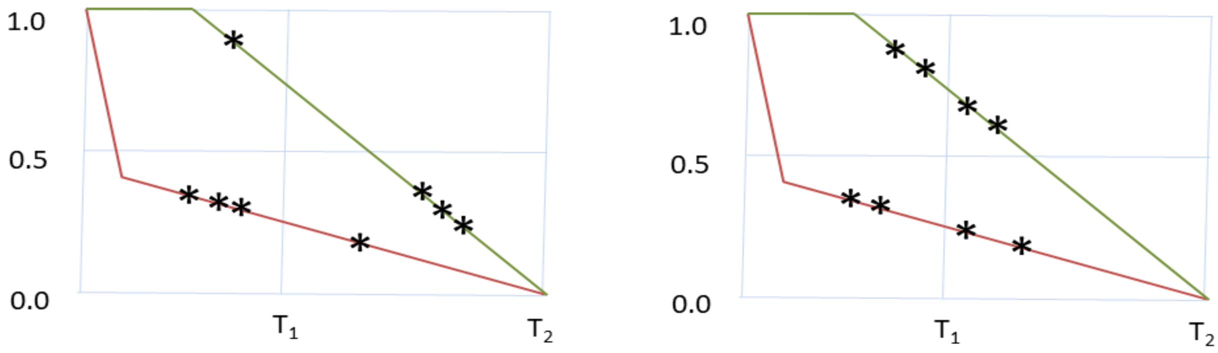


Figure 15: Simplified models to illustrate: [left] a model can have perfect 1-Calibration for a time, but not be D-Calibrated, and [right] a model can have perfect D-Calibration, but not be 1-Calibrated for a time. (See text for description.)

Proof. **"1-Calibration $\not\Rightarrow$ D-Calibration":** Consider the model shown in Figure 15[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 $\hat{S}(\,T_1\,|\,\vec{x}\,)$ values, then partition them into k sets. Here, we consider $k = 2$, rather than the deciles earlier. The first set contains the 4 patients with $\hat{S}(\,T_1\,|\,\vec{x}\,) = 0.75$ (*i.e.*, the green patients); and the second, the 4 patients with $\hat{S}(\,T_1\,|\,\vec{x}\,) = 0.25$. 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 partition the time intervals into 2 sets (not 10), as shown in Figure 15. Here, $\hat{S}(\,d_{\vec{x}}\,|\,\vec{x}\,) \in [0.5, 1]$ holds for only 1 patient, and $\hat{S}(\,d_{\vec{x}}\,|\,\vec{x}\,) \in [0, 0.5]$ holds for 7; if the model was D-Calibrated, each of these sets should contain 4 patients.

"D-Calibration $\not\Rightarrow$ 1-Calibration": See Figure 15[right], where again, each line represent 4 different 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 $\hat{S}(\,d_{\vec{x}}\,|\,\vec{x}\,) \in [0.5, 1]$ (the green patients), and 4 with $\hat{S}(\,d_{\vec{x}}\,|\,\vec{x}\,) \in [0, 0.5]$ (for 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; and of the 4 patients with $\hat{S}(\,T_1\,|\,\vec{x}\,) = 0.25$, 2 are alive at T_1. To be 1-Calibrated, there should be 3 living patients in the first set, and 1 in the second; hence this model is not 1-Calibrated at T_1. □

B.6 Other Subtle Points

All of these tools for producing survival curves are able to deal with “right censored” events: where the censored event time is a *lower bound* of the time of death. (This corresponds to, perhaps, the termination of a study, or when a participant left the study early.) There are other types of censoring, including “left censoring”, which provides an upper-bound on the time of death (*e.g.*, when a survey finds that the patient is currently dead, but does not know when previously this happened), and “interval censoring”, when we can constrain the time of death to some interval. While there are extensions of each of these tools that can accommodate these alternate types of censoring, here we considered the most common case of having right-censored instances, and included only datasets that had only such instances.

As a second subtle issue: some of the methods involve taking the log of a predicted value, or of a true value; see Appendix B.2.1. This is clearly problematic if that value is 0 – *e.g.*, if a patient died during a transplantation surgery. To avoid these errors, we replace any such 0 with the η for a database, which is defined as 1/2 of the minimum observed positive time of any event, in that dataset. That is, we ignore all time=0 events, and then consider the smallest remaining value. If that value is, say, 1.0 day, then we set η =0.5 days. Note that all other times are left unchanged.

C Comments about Various ISD’s

C.1 Comments about COX-KP

Notice Equation 2 embodies two strong assumptions: (1) that the individual features are independent of one another (*e.g.*, the outcome does not depend on a non-linear combination of the features), and (2) that these covariates are independent of time – which means that a blood test is as important just after an operation, as it is a year later, or a decade later. These assumptions mean the survival curves for different patients will have the same basic shape, and will not cross; see Figure 4 (middle-left). These simplifications allow the Cox model to suggest important information about individual features by examining the single coefficient β_j associated with the j^{th} feature, *e.g.*, *does “being male” increase the risk of dying from this specific cancer, or does it protect against this outcome (or neither).* This “neither” case suggests that a given feature is not relevant to the prediction; for this reason, we used univariate Cox is a feature selection technique for our results.

By contrast, MTLR and RSF-KM do not make these extreme assumptions, which means that a given feature can have different levels of importance at different times. Moreover, the curves for different patients can cross; see Figure 4. More relevant, however, we found that MTLR is more often D-Calibrated, and hence more meaningful for individual patients, than this “predictive Cox” system; see Table 6. While this Cox analysis of survival may not be directly relevant for individual patients, there are still extreme benefits in being able to identify important features. By observing how different features impact survival, clinicians can be made aware of treatments or lifestyle changes that best help patients survival.

C.2 Overview of MTLR

Consider[20] 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 $\vec{x}$,

$$S_{\vec{\theta}_i}(\, T \geq t_i \,|\, \vec{x}\,) \quad = \quad \Big(1 + \exp(\vec{\theta}_i \cdot \vec{x})\Big)^{-1}, \qquad 1 \leq i \leq m, \tag{27}$$

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[21] $[0, 0, \ldots, 1, 1, \ldots, 1]$, including the sequence of all '0's and the sequence of all '1's. Our MTLR model computes 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 dataset 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

$$\mathbf{\Theta}^* \;=\; \arg\max_{\mathbf{\Theta}} \sum_{r=1}^{n} \left[\sum_{i=1}^{m} y_j(d_r)(\vec{\theta}_i{\cdot}\vec{x}_r) - \log \sum_{k=0}^{m} \exp f_{\mathbf{\Theta}}(\vec{x}_r, k) \right] - \frac{C}{2} \sum_{j=1}^{m} \|\vec{\theta}_j\|^2 \tag{28}$$

where the C (for the regularizer) is found by an internal cross-validation process.

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.* [57] 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 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 4 (bottom-right).

[20] This paragraph is paraphrased from [57]; reprinted with permission of publisher/author.

[21] Notice there are no '0's after a '1'. This is the 'no zombie' rule: once someone dies, that person stays dead.

C.3 Extension to Random Survival Forests (RSF-KM)

Given a labeled dataset, a random survival forest learner will produce a set of T decision trees from a bootstrapped sample 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 the 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 logrank test statistic (or some other chosen splitting rule). This becomes the rule of that node; and 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, 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.[22]

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 16.[23]

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 datasets presented in the paper. We found that the Concordance scores were similar to that of RSF-KM; MTLR still outperformed RSF on the datasets where MTLR outperformed RSF-KM (data not shown).

[22] 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.

[23] 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.

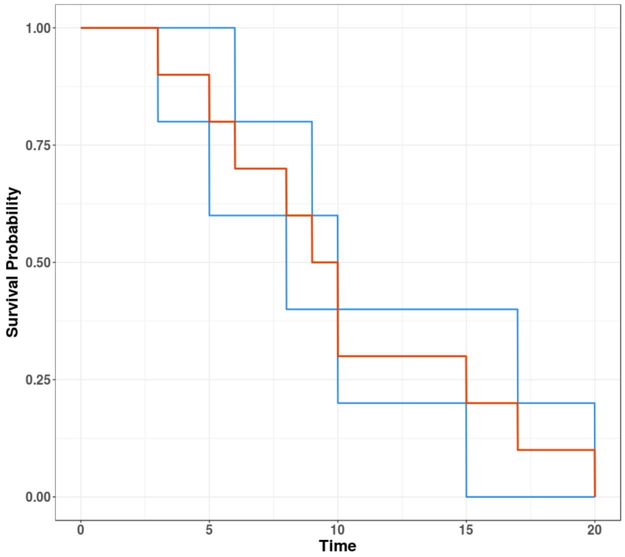


Figure 16: This figure illustrates how to combine two different survival curves, to produce a new one. (RSF-KM uses this idea to "merge" the curves obtained from the various leaf nodes reached by a novel instance.) Here, two survival curves, given in blue, are averaged to produce the survival curve shown 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 – that is a drop in either (blue) Kaplan-Meier curve.

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 8 for the results corresponding to Figure 9.

Table 8: Concordance results corresponding to Figure 9. **Bold** values indicate the best (highest Concordance) performing model, for each dataset.

	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)**

D.2 Brier Score

See Table 9 for the results corresponding to Figure 10.

Table 9: Integrated Brier score results corresponding to Figure 10. **Bold** values indicate the best performing model for each dataset – 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, based on additional digits, not shown.

	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)**

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

Here we give the the results for the Margin-L1-loss (Table 10) as given in Figure 11. Additionally, we give results for the Uncensored L1-loss (Table 11) and the Log-Margin-L1-loss (Table 12).

Table 10: Margin-L1-loss results corresponding to Figure 11. **Bold** values indicate the best performing model for each dataset – with the lowest Margin-L1-loss.

	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 11: Uncensored L1-loss (not the L1-Margin loss given in Figure 11). **Bold** values indicate the best performing model for each dataset – with the lowest L1-loss.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL
KM	318.53 (10.49)	520.32 (9.56)	19.40 (0.15)	12.17 (0.08)	1829.99 (1261.87)	2418.79 (24.55)	18.69 (0.64)	**2.90 (0.20)**
AFT	291.28 (28.99)	524.78 (64.89)	19.58 (1.38)	12.54 (1.07)	1738.27 (1191.64)	2681.47 (453.66)	23.77 (3.95)	13.48 (3.01)
COX-KP	281.61 (26.53)	542.35 (76.2)	18.91 (1.35)	12.63 (1.12)	2248.56 (1263.39)	3141.20 (521.90)	-	-
COXEN-KP	284.86 (14.93)	482.70 (44.76)	16.14 (1.12)	**10.64 (0.66)**	1936.34 (1186.59)	2647.76 (399.62)	11.52 (1.51)	3.19 (1.38)
RSF-KM	373.14 (72.48)	1204.95 (276.78)	28.33 (2.15)	15.61 (1.07)	4099.14 (1482.32)	5671.22 (511.99)	13.07 (1.51)	3.91 (0.93)
MTLR	**272.20 (27.15)**	**436.99 (62.40)**	**15.87 (1.03)**	10.71 (0.57)	**1411.46 (306.97)**	**2167.09 (208.32)**	**9.30 (1.30)**	3.70 (0.92)

Table 12: Log-Margin-L1-loss (not the L1-Margin loss given in Figure 11). **Bold** values indicate the best performing model for each dataset – with the lowest Log-Marign-L1-loss.

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL
KM	1.98 (0.01)	2.35 (0.02)	1.76 (0.04)	2.20 (0.01)	2.05 (0.19)	1.78 (0.03)	2.02 (0.04)	3.37 (0.11)
AFT	**1.64 (0.08)**	1.36 (0.06)	1.44 (0.01)	**1.48 (0.04)**	2.00 (0.60)	1.65 (0.14)	6.21 (0.68)	10.54 (2.62)
COX-KP	1.68 (0.07)	1.35 (0.06)	1.42 (0.02)	1.48 (0.04)	2.02 (0.56)	1.73 (0.11)	-	-
COXEN-KP	1.80 (0.01)	1.35 (0.05)	1.43 (0.03)	1.55 (0.05)	**1.88 (0.56)**	**1.56 (0.06)**	1.77 (0.08)	2.70 (0.11)
RSF-KM	1.85 (0.20)	1.98 (0.10)	1.51 (0.05)	1.54 (0.04)	2.62 (0.47)	2.21 (0.12)	1.90 (0.05)	3.04 (0.14)
MTLR	1.67 (0.07)	**1.25 (0.08)**	**1.37 (0.05)**	1.50 (0.04)	1.95 (0.17)	1.57 (0.06)	**1.67 (0.07)**	**2.60 (0.18)**

D.4 1-Calibration

Each table corresponds to a different percentile of event times for each dataset. Moving down the 10th, 25th, 50th, 75th, and 90th percentiles are given. **Bolded** values indicate models which passed 1-Calibration ($p > 0.05$). The "Total" column of each table gives the total number of datasets passed by each model – that is, the values in that columns correspond to Table 5.

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

	GBM	GLI	Nacd-Col	NACD	READ	BRCA	DBCD	DLBCL	Total
AFT	0.001	**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 14: 1-Calibration Results at $t^* = 25$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 15: 1-Calibration Results at $t^* = 50$th 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 16: 1-Calibration Results at $t^* =$ 75th 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 17: 1-Calibration Results at $t^* =$ 90th 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