Inverse-Weighted Survival Games

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep discrete-time models have become a default choice for survival analysis. Practitioners train these models with maximum likelihood, but often evaluate them with other criteria like binary losses evaluated at various time horizons, e.g. Brier score (BS) and Bernoulli log likelihood (BLL) at 5 years. Due to finite data, these models may have poor BS and BLL. However, likelihood does not directly optimize these criteria. Directly optimizing criteria like BS requires re-weighting by the censorship distribution which is generally unknown. We introduce *Inverse-Weighted Survival Games*, a non-adversarial game where failure and censoring models simultaneously optimize their respective binary losses at many time horizons. The two models re-weight each other's loss. When the loss is proper, we show that the games always has the true failure and censoring distributions as a stationary point. This means models in the game do not leave the correct distributions once reached. When the game is based on BS, the true distributions are the *only* stationary point, making BS-based games practical. We show that these games optimize the true uncensored BS on simulations and then apply these principles on real world data.

1 Introduction

Survival analysis is the modeling of time-to-event distributions and is widely used in healthcare to predict time from diagnosis to death, risk of disease recurrence, and changes in level of care. In survival data, events, known as *failures* may be *censored*, i.e. only a time range is known. Under right-censoring, we only observe a lower-bound on some failure times, for instance when a patient leaves a study before failing [Huang et al., 2004]. Though the usual goal is to model failures, discarding censored points biases estimates. In this work, we take careful consideration of the role of censoring.

Under independent censoring, likelihood training only requires evaluating a failure model on observed failure and censoring times [Kalbfleisch and Prentice, 2002]. Most practitioners make these assumptions and avoid the nuisance of modeling censoring explicitly. However, we show that, even under these assumptions, modeling censoring can help failure model learning and evaluation.

Deep survival models have dominated recent survival work [Ranganath et al., 2016, Alaa and van der Schaar, 2017, Katzman et al., 2018, Miscouridou et al., 2018, Kvamme et al., 2019, Zhong and Tibshirani, 2019, Goldstein et al., 2020]. We focus in this work on deep *discrete-time* models, which have become a default choice even when data is continuous [Lee et al., 2018, Fotso, 2018, Kvamme and Borgan, 2019b] because they can draw on deep classification architectures and can approximate continuous densities well with enough bins [Miscouridou et al., 2018].

Though training is often done through maximum likelihood, survival analysis is full of other evaluation criteria. A conceptually simple way to evaluate models that is particularly suitable for discrete models is to apply binary classification losses at various time horizons, for instance Brier score (BS) for

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

5-year survival, or similarly, Bernoulli log likelihood (BLL) [Kvamme and Borgan, 2019b, Lee et al., 2019, Haider et al., 2020].

The story of deep models and survival evaluations is not so simple. Due to finite data, models may have poor BS and BLL. However likelihood does not directly optimize these criteria. Due to missingness, directly optimizing criteria like BS requires re-weighting using inverse probability of censor-weighting (IPCW) [Van der Laan et al., 2003]. However, IPCW requires the censorship distribution, which is generally unknown. Though without weighting, using only uncensored points shifts the covariate distribution and renders estimates inconsistent.

We introduce *Inverse-Weighted Survival Games*, a non-adversarial game featuring a failure model and censoring model that each optimize a set of time-dependent binary losses. These losses are built from IPCW estimates and the two models re-weight each other's losses. Throughout training, the models become effective in weighting to account for censoring. In IPCW games, when one distribution is correct, the other directly optimizes the loss that would be computed without missing times.

When the loss is *proper* (e.g. BS, BLL) [Gneiting and Raftery, 2007], we show that IPCW games have the true failure and censoring distributions as a stationary point. This means the models in the game do not leave the correct distributions once reached. We then show that when the game is based on BS, these distributions are the *only* stationary point. We show that these games optimize the uncensored BS on simulations and apply these principles on real world data.

2 Notation and background on IPCW

Notation. Let T be a failure time with CDF F, *survival function* $\overline{F} = 1 - F$, and model F_{θ_T}. Let C be a censoring time with CDF G, survival function $\overline{G}$, and model G_{θ_C}. Under right-censoring, $U = \min(T, C)$, $\Delta = \mathbb{1}[T \leq C]$ and we observe (X_i, U_i, Δ_i). We use $\overline{G}(t^-)$ to denote $P(C \geq t)$.

Assumptions. We assume i.i.d. data and random censoring: $T \perp\!\!\!\perp C \mid X$ [Kalbfleisch and Prentice, 2002]. Derivations in this work also require the censoring positivity assumption [Gerds et al., 2013]. Let f be a failure density or mass,

$$\forall x, \forall t \in \{t \mid f(t|x) > 0\}, \quad \exists \epsilon > 0 \quad \text{s.t.} \quad \overline{G}(t^-|x) > \epsilon > 0. \tag{1}$$

This says we observe all subjects' events with positive probability. "Censoring" typically refers to missingness of failure times. However, in this work we model censoring too, a survival modeling task where censoring times are censored by failures. We then also need the positivity condition to hold with the roles of F and G reversed in order to observe censoring events properly (Appendix A).

Models. We focus on deep discrete-time models like those studied in Miscouridou et al. [2018], Lee et al. [2018], Kvamme and Borgan [2019b]. Here X is mapped through a deep model to produce a categorical distribution. When data are continuous, a discretization scheme is necessary. Kvamme and Borgan [2019b], Goldstein et al. [2020] set bins to correspond to quantiles of the observed failure times. We represent all values that fall within an interval by the lower boundary of that interval.

IPCW Estimators. IPCW is a technique for estimation under censoring [Gerds and Schumacher, 2006]. Consider the marginal mean of T. IPCW reformulates such expectations in terms of observed data. Using this method, we can show (Appendix B) that:

$$\mathop{\mathbb{E}}_{X,T}[T] = \mathop{\mathbb{E}}_{T,C,X}\left[\frac{\mathbb{1}[T \leq C]}{\overline{G}(T^-|X)} T\right] = \mathop{\mathbb{E}}_{U,\Delta,X}\left[\frac{\Delta U}{\overline{G}(U^-|X)}\right]$$

This holds because $\Delta = 1 \implies U = T$ and means we can estimate the mean, provided that we know G and that random censoring and positivity (Equation (1)) hold. In practice, we must learn the censoring distribution, a challenging task as it is also censored.

3 Time-dependent survival evaluations

One of two evaluations we focus on in this work is the Brier score (BS) [Brier and Allen, 1951]. BS is a *strictly proper* scoring rule for classification tasks, which means that it takes its minimum

value only for the correct data-generating distribution [Gneiting and Raftery, 2007]. The BS is often adapted for survival modeling [Lee et al., 2019, Kvamme et al., 2019, Haider et al., 2020]. For time t, it computes squared error between the CDF and true event status at t, turning survival analysis into a classification problem:

$$\text{BS}(t;\theta) = \mathop{\mathbb{E}}_{T,X}\left[\left(F_{\theta_T}(t \mid X) - \mathbb{1}\left[T \leq t\right]\right)^2\right] \tag{2}$$

Unfortunately one cannot compute BS on censored data since $\mathbb{1}\left[T \leq t\right]$ is unobserved for a point censored before t. IPCW BS [Graf et al., 1999, Gerds and Schumacher, 2006] can be used to estimate $\text{BS}(t)$ under censoring (Appendix D):

$$\text{F-BS-CW}(t;\theta) = \mathop{\mathbb{E}}_{T,C,X}\left[\frac{\overline{F_{\theta_T}}(t \mid X)^2 \Delta \mathbb{1}\left[U \leq t\right]}{G(U^- \mid X)} + \frac{F_{\theta_T}(t \mid X)^2 \mathbb{1}\left[U > t\right]}{G(t \mid X)}\right]. \tag{3}$$

Negated IPCW Bernoulli log likelihood (BLL) is similar, but replaces the squared error with negated log loss (Appendix C). BS and BLL assess both discrimination and calibration [Kvamme et al., 2019, Haider et al., 2020]. Since they are strictly proper for classification, the summed BS or BLL over all times of a discrete model is strictly proper.

Though such objectives all have the same true data distribution at optimum with infinite data, they may yield significantly different solutions in practice. For example, consider that likelihood incurs infinite loss when a datapoint is assigned 0 probability while BS does not. Because solutions differ, one should not rely on one loss to yield good performance under another. Therefore it is useful to consider optimizing survival models directly for BS or BLL. However, under censoring, we must use the IPCW form for these objectives. Crucially, the G in Equation (3) is the true censoring distribution rather than a model but during training, we only have access to models F_{θ_T} and G_{θ_C}. This poses a re-weighting dilemma: can these models be used successfully for re-weighting during training?

4 Inverse-Weighted Survival Games

As a first attempt to solve the re-weighting dilemma, one might choose a failure loss weighted by the censoring model, a censoring loss weighted by the failure model, and jointly optimize the sum under the principle that both models will improve over training and yield good IPCW estimates. For concreteness, consider optimizing Equation (3) plus the same objective with the roles of the censoring and failure models reversed. Unfortunately, there exist solutions with smaller loss than the true pair of distributions, making the summed objective improper. In Figure 1, we plot this phenomenon for an IPCW $\text{BS}(t=1)$ game:

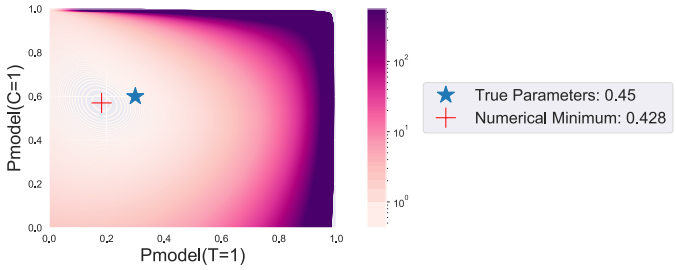


Figure 1: Sum of IPCW brier scores is improper as a joint objective for the failure and censoring models

To address this phenomenon, we introduce *Inverse-Weighted Survival Games*, a non-adversarial game that simultaneously optimizes a failure model and censoring model. We show in the experiments that these games produce failure models with good BS, BLL, and concordance relative to models trained with categorical likelihood.

Constructing the game. For simplicity, we present the games and results for the marginal categorical case. The results can be extended to conditional parameterizations with the usual

drawbacks common to maximum likelihood. Consider discrete distributions over K times. Let $\theta_T = \{\theta_{T1}, \cdots, \theta_{TK}\}$, $\theta_{Tt} = P_\theta(T = t)$, $F_{\theta_T}(t) = \sum_{k=1}^{t} \theta_{Tk}$, and likewise for C, θ_C.

Consider a loss $L_{F^\star}(F_{\theta_T})$ for model F_{θ_T} and data sampled from $F^\star$. We use L^t to denote a loss that depends on a chosen time, e.g. $\text{BS}(t)$ or $\text{BLL}(t)$. Under censoring, T is censored by $C \sim G^\star$. Define $L^{CW,t}_{F^\star,G^\star}(F_{\theta_T}; \cdot)$ (censor-weighted L^t) such that the right-hand argument is the IPCW weight:

$$L^t_{F^\star}(F_{\theta_T}) = L^{CW,t}_{F^\star,G^\star}(F_{\theta_T}; G^\star)$$

We define an *Inverse-Weighted Survival Game* at time t by IPCW estimators $L^{CW,t}$ of L^t, where the failure and censor models re-weight each other's loss:

$$V^t_F(\theta) = L^{CW,t}_{F^\star,G^\star}(F_{\theta_T}; G_{\theta_C}), \quad V^t_G(\theta) = L^{CW,t}_{G^\star,F^\star}(G_{\theta_C}; F_{\theta_T}) \tag{4}$$

We have replaced the true re-weighting distributions with models. The *failure player* minimizes V^t_F w.r.t. θ_{Tt} and the *censor player* minimizes V^t_G w.r.t. θ_{Ct}. We solve the games at all timesteps simultaneously with one failure model and one censoring model. As we will see, this formulation has advantages over the optimization in Figure 1.

Example. One instantiation of these games is the IPCW $\text{BS}(t)$ game, derived in Appendix D.

$$\begin{aligned} V^t_F(\theta) &= \mathop{\mathbb{E}}_{T,C,X}\Big[\frac{\overline{F_{\theta_T}}(t)^2 \Delta \mathbb{1}[U \le t]}{G_{\theta_C}(U^-)} + \frac{F_{\theta_T}(t)^2 \mathbb{1}[U > t]}{G_{\theta_C}(t)}\Big] \\ V^t_G(\theta) &= \mathop{\mathbb{E}}_{T,C,X}\Big[\frac{\overline{G_{\theta_C}}(t)^2 (1-\Delta) \mathbb{1}[U \le t]}{F_{\theta_T}(U)} + \frac{G_{\theta_C}(t)^2 \mathbb{1}[U > t]}{F_{\theta_T}(t)}\Big]. \end{aligned} \tag{5}$$

4.1 IPCW games have a stationary point at data distributions

At minimum, for a game to be useful, among its stationary points should be the true distributions.

Proposition 1. Assume $\exists \theta^\star_T \in \Theta_T, \exists \theta^\star_C \in \Theta_C$ such that $F^\star = F_{\theta^\star_T}$ and $G^\star = G_{\theta^\star_C}$. Assume $L^t_{P^\star}(P)$ is proper, i.e. for model P and sampling distribution $P^\star$, L^t satisfies $P \neq P^\star \implies L^t_{P^\star}(P^\star) \le L^t_{P^\star}(P)$. Then $(\theta^\star_T, \theta^\star_C)$ is a stationary point of the game Equation (4) for any t.

Proof. See Appendix H. □

$\text{BS}(t)$, $\text{BLL}(t)$, and $\text{AUC}(t)$ are examples of proper scoring rules, while concordance(t) is not [Blanche et al., 2019]. When IPCW games are built from such objectives, the set of solutions includes $(\theta^\star_T, \theta^\star_C)$, meaning the models do not leave the correct failure and censoring distributions once reached.

4.2 Finding Stationary Points

A solution is a *stationary point*, a set of states where no player will change their state. Such points (θ_T, θ_C) are characterized by the variational inequality [Harker and Pang, 1990, Gidel et al., 2018], which says that at (θ_T, θ_C), all directional derivatives are non-negative. To solve all games simultaneously using unconstrained optimization, the parameters must normalize appropriately. Let h be an invertible function mapping from unconstrained parameters $\eta_{T1:K-1}$ to categorical parameters $\theta_{T1:K-1}$. Assembling the failure player derivatives in vector form, we have:

$$dV_F(\theta) = \Big[\ \frac{dV^1_F}{d\theta_{T1}}\Big|_\theta, \quad \ldots, \quad \frac{dV^t_F}{d\theta_{Tt}}\Big|_\theta, \quad \ldots, \quad \frac{dV^{K-1}_F}{d\theta_{T,K-1}}\Big|_\theta\ \Big]$$

To solve the games, we want $dV_F(\theta) = 0$ and $dV_G(\theta) = 0$. Letting J_h be the Jacobian of h, define:

$$g_T(\eta_T) \triangleq \Big(\ J_h|_{\eta_T}\ \Big)^\top dV_F(h(\eta_T)) \tag{6}$$

We can show that setting $g_T(\eta_T) = 0$ sets $dV_F(h(\eta_T)) = 0$. When h is invertible, J_h^{-1} exists when $J_{h^{-1}}$ does. Then, provided we choose an h with differentiable inverse,

$$\Big(\ J_{h^{-1}}|_{h(\eta_T)}\ \Big)^\top g_T(\eta_T) = dV_F(h(\eta_T))$$

Algorithm 1 Following Gradients in Inverse-Weighted Games

Input: Choice of value functions V_F, V_G, normalization function h, learning rate γ.
Initialize $\eta_{T,1:K-1}$ and $\eta_{C,1:K-1}$ randomly.
repeat
  // compute both gradients simultaneously
  $\eta_T \leftarrow \eta_T - \gamma g_T$ and $\eta_C \leftarrow \eta_C - \gamma g_C$
until convergence
Output: η_T, η_C

An invertible matrix only has the 0-vector in its kernel, which means $g_T(\eta_T) = 0 \iff \forall t, \frac{dV_F^t}{d\theta_t} = 0$. Define $g_C(\eta_C)$ analogously. To set $g_T(\eta_T)$ and $g_C(\eta_C)$ to 0 one can directly minimize their norms. In in our experiments, following g_T and g_C converges (Algorithm 1 and Appendix E).

5 Discrete Brier Games

We now provide a stronger result for the IPCW BS game in Equation (5): its only stationary point is at the true failure and censoring distributions. This means that IPCW BS games do not introduce extra local optima beyond those expected when introducing deep conditional parameterizations.

Proposition 2. Assuming that $\theta^\star_{Tt} > 0$ and $\theta^\star_{Ct} > 0$, the solution $(\theta^\star_T, \theta^\star_C)$ is the only feasible stationary point common to the set games at all times.

Proof. See Appendix I. □

For the same experiment as in Figure 1, Figure 2 shows that the IPCW BS game moves to the correct solution at its unique stationary point.

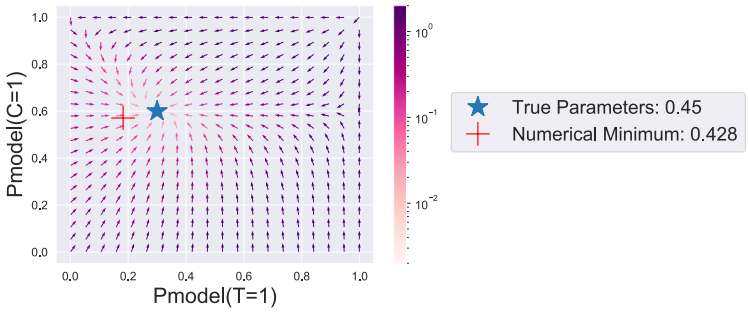


Figure 2: Gradient field of game at timestep one. The unique stationary point of game is at true distribution, contrasting the improper objective in Figure 1.

6 Experiments

We run experiments on a simulation with conditionally Gamma failure and censoring times, a semi-simulated survival dataset based on MNIST, several sources of cancer data, and data on critically ill hospital patients.

We optimize the same categorical model in three different ways: negative log likelihood (NLL), the IPCW BS game, and the IPCW negated BLL game. For these models we report BS (uncensored for simulations and Kaplan-Meier (KM)-weighted for real data), BLL (likewise uncensored or weighted), concordance which measures the proportion of datapoint pairs whose predicted risks are ordered correctly [Harrell Jr et al., 1996], and NLL.

Model Description. In all experiments except for MNIST, we use a 3-hidden-layer ReLU network. The hidden sizes are [128, 64, 64] for the Gamma simulation and [128,256,64] for the real data. We

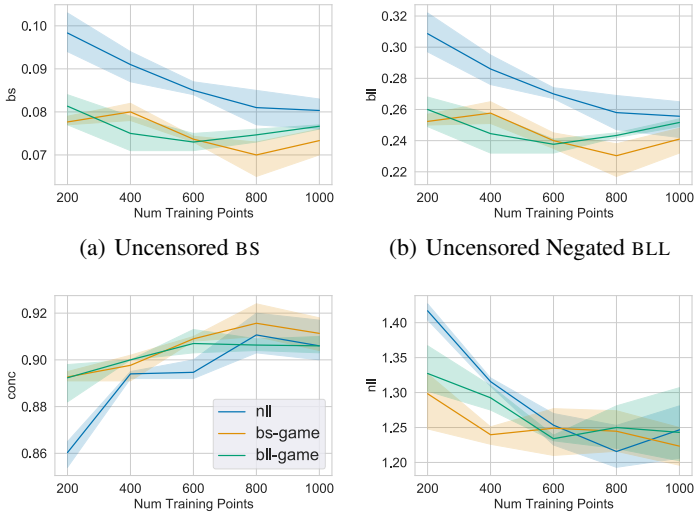
output 20 categorical bins. For MNIST we first use a small convolutional network and follow with the same fully-connected network as the Gamma experiment. We include more details in Appendix G.

For the Gamma and Survival-MNIST simulations, we found that stick-breaking, an alternative to softmax normalization (e.g. [Khan et al., 2012]), worked better for all objectives, and that softmax worked better for all objectives on the real data. Stick-breaking is described in Appendix F.

6.1 Simulation Studies

Data. We draw x from a 32D multivariate normal $\mathcal{N}(0, 10I)$. We simulate conditionally gamma failure times with mean μ_t a log-linear function of x. The censoring times are also conditionally gamma with mean $0.9 * \mu_t$. Both distributions have constant variance 0.05. T and C are conditionally independent given X. Each random seed draws a new dataset.

Results. The results in Figure 3 demonstrate the basic idea and show that it is possible to optimize the true uncensored BS with the IPCW games, and, though more slowly (with respect to training size), with log-likelihood too. We see that the IPCW games have better test-set performance on all metrics for small training size, and that all three methods converge on similar performance when there is enough data (though *enough* is highly dependent on dimensionality and model class).



(a) Uncensored BS (b) Uncensored Negated BLL

(c) Concordance (d) Categorical NLL

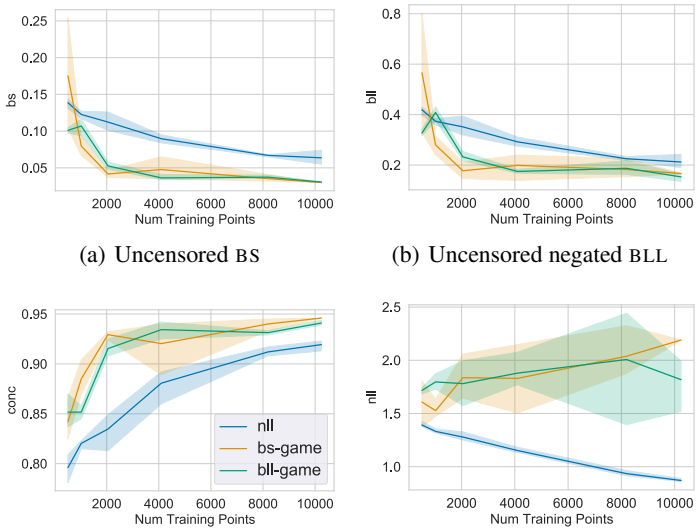
Figure 3: Test set evaluation metrics on Gamma simulation versus number of training points for three different methods. Lower is better for all the metrics except for concordance.

6.2 Semi-simulated studies

Data. Survival-MNIST [Gensheimer, 2019, Pölsterl, 2019] draws times conditionally on MNIST label Y. This means digits define risk groups and $T \perp\!\!\!\perp X \mid Y$. Times within a digit are i.i.d. The model only sees the image pixels X as covariates so it must learn to classify digits (risk groups) to model times. The PyCox package [Kvamme et al., 2019] uses Exponential times. We follow Goldstein et al. [2020] and use Gamma times. T's mean is $10 * (Y + 1)$ so that lower labels Y mean sooner event times. We set the variance constant to 0.05. C is drawn similarly but with $9.9 * (Y + 1)$. Each random seed draws a new dataset.

Results. We demonstrate in this experiment that better NLL does not correspond to better performance on BS, BLL, and concordance. Similarly to the previous experiment, we see in Figure 4 that the game methods attain better uncensored test-set BS and BLL on survival-MNIST. The game methods likewise have higher concordance. NLL training performs better at the metric it directly optimizes.

This experiment also establishes that it is possible to optimize through deep convolutional models with batch norm and pooling using the game training methods.



(a) Uncensored BS (b) Uncensored negated BLL

(c) Concordance (d) Categorical NLL

Figure 4: Test set evaluation metrics on the survival-MNIST simulation versus number of training points for three methods. Lower is better for all the metrics except for concordance.

6.3 Real Datasets

Data. We use several survival datasets available in the R survival package [Therneau, 2021] and recently used e.g. in Katzman et al. [2018] and their DeepSurv package, Kvamme et al. [2019], Kvamme and Borgan [2019b] and their PyCox package, and Chen [2020]. The datasets are

- Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) [Knaus et al., 1995] which includes severely ill hospital patients
- Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [Curtis et al., 2012]
- Rotterdam Tumor Bank (ROTT) [Foekens et al., 2000] and German Breast Cancer Study Group (GBSG) [Schumacher et al., 1994] combined into one dataset (ROTT. & GBSG)

For more description see Therneau [2021], Katzman et al. [2018], Chen [2020]. For real datasets, there is no known ground truth for the censoring distribution, which means evaluation requires assumptions. Following the experiments in Kvamme et al. [2019], we assume that censoring is marginal and apply the KM estimator to evaluate the learned models. This is also the route taken in the R packages Survival [Therneau, 2021], PEC [Mogensen et al., 2012], and riskRegression [Gerds et al., 2020].

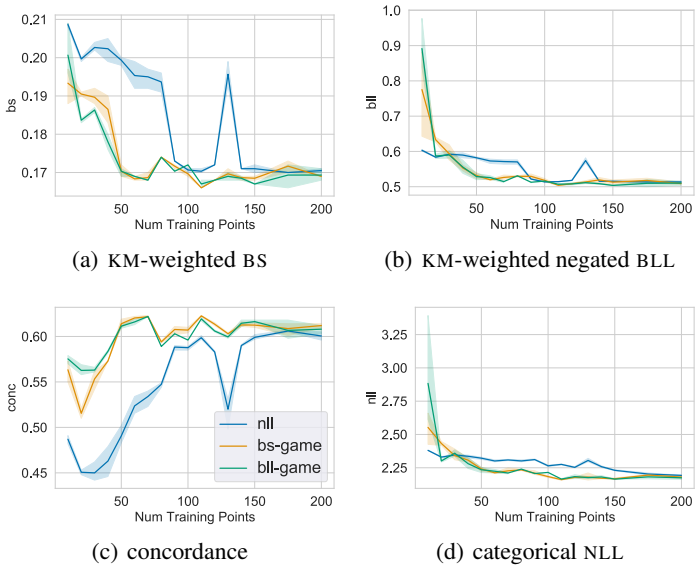
Results. On METABRIC, we see that the game training attains lower (better) KM-weighted BS and BLL than NLL training when the number of datapoints is small. We also see that game training has better concordance, and NLL though they do not directly optimize it. As the number of datapoints grows, the three methods converge the same values for all four metrics.

On Rotterdam & GBSG (ROTT. & GBSG), the differences in BS are smaller, but the BS game performs better on BS than the other two methods when the number of datapoints is small. Both game methods have better concordance than NLL for most training set sizes. Again, all three methods converge to about the same value in all four metrics when the number of datapoints is large enough.

All methods perform similarly on SUPPORT.

No method is stable for all metrics, for all training sizes, on all seeds. We see that NLL has a drop in performance on METABRIC around 130 training points and likewise the BLL game on ROTT. & GBSG around 40 training points.

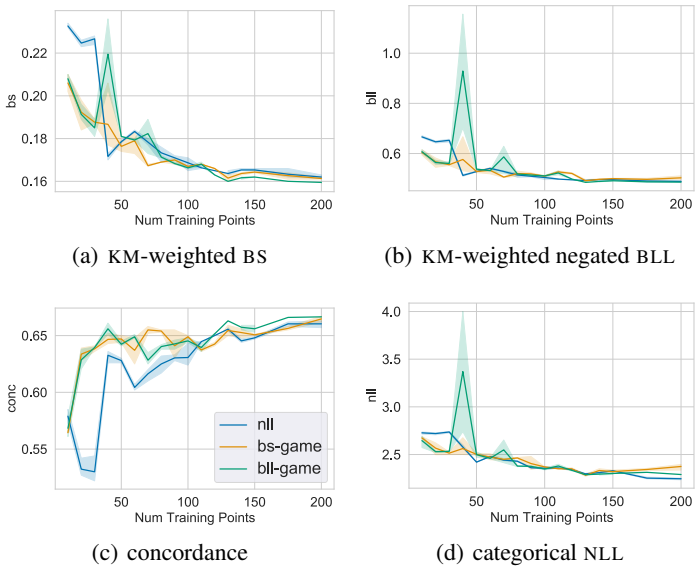
Finally, we mention two caveats. First, though popular, these survival datasets are low-dimensional, so any of the objectives can perform well on the metrics with just several hundred points. We see that this is distinct from the MNIST simulation, where thousands of points were required to improve performance. Second, though possible, it may not be true that censoring is marginal on these real datasets, which would mean that the BS and BLL results only have their interpretation conditional on a particular set of covariates.



(a) KM-weighted BS (b) KM-weighted negated BLL

(c) concordance (d) categorical NLL

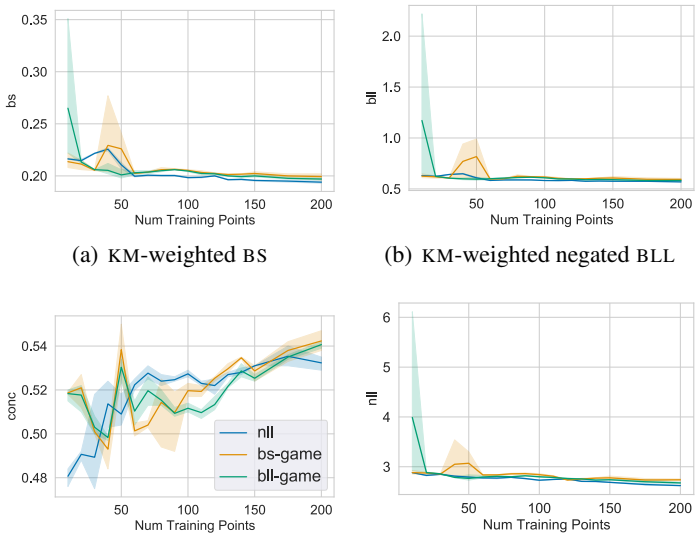
Figure 5: Test set evaluation metrics on METABRIC versus number of training points for three methods. Lower is better for all the metrics except for concordance.



(a) KM-weighted BS (b) KM-weighted negated BLL

(c) concordance (d) categorical NLL

Figure 6: Test set evaluation metrics on ROTT. & GBSG versus number of training points for three methods. Lower is better for all the metrics except for concordance.



(a) KM-weighted BS (b) KM-weighted negated BLL

(c) concordance (d) categorical NLL

Figure 7: Test set evaluation metrics on SUPPORT versus number of training points for three methods. Lower is better for all the metrics except for concordance.

7 Discussion

In this work, we propose a new training method for survival models under censored data. We argue that on finite data, it is important to close the gap between training methodology and the desired evaluation criteria. We showed in the experiments that better NLL does not correspond to better performance on BS, BLL, and concordance, all evaluations of interest in survival analysis.

The main trend in our experimental results was that data size matters: smaller meant the game methods performed better than NLL and enough data meant that they perform similarly, which is expected since all objectives are proper. However *enough* data is hard to define: it depends on dimensionality and on the data generating distribution and model class. It is a great direction of future work to build more precise understanding on how objectives behave differently even when they have the same optimum on infinite data.

Limitations. As mentioned in the experiments, evaluation on real data under censoring requires assumptions. It is important to further consider how to better assess test-set performance on metrics such as BS, BLL, and concordance. Because concordance is not proper, we do not build objectives from it in this work, but it too is not invariant to censoring.

Regarding games, we showed properties about their stationary points. More analysis is necessary to describe important properties of optimizing these games and when this converges to the stationary points.

Social Impact. Survival models are deployed in hospital settings and have high impact on public health. In this work, we saw benefits of a new training approach for these models, but no training method is a panacea. Practitioners of survival analysis must take great care to consider various training and validation approaches.

References

A. M. Alaa and M. van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334. Curran Associates Inc., 2017.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

P. Blanche, J.-F. Dartigues, and H. Jacqmin-Gadda. Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5): 687–704, 2013.

P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

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

G. H. Chen. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, pages 537–565. PMLR, 2020.

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

M. R. Elliott. Model averaging methods for weight trimming. *Journal of official statistics*, 24(4):517, 2008.

J. A. Foekens, H. A. Peters, M. P. Look, H. Portengen, M. Schmitt, M. D. Kramer, N. Brünner, F. Jänicke, M. E. Meijer-van Gelder, S. C. Henzen-Logmans, et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60(3): 636–643, 2000.

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

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

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.

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

T. A. Gerds, P. Blanche, T. H. Scheike, R. Mortensen, M. Wright, N. Tollenaar, J. Muschelli, U. B. Mogensen, and B. Ozenne. Package 'riskregression', 2020.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

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

M. Goldstein, X. Han, A. M. Puli, A. Perotte, and R. Ranganath. X-cal: Explicit calibration for survival analysis. *Advances in Neural Information Processing Systems*, 33, 2020.

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.

H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3): 161–220, 1990.

F. E. Harrell Jr, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

X. Huang, R. A. Wolfe, and C. Hu. A test for informative censoring in clustered survival data. *Statistics in medicine*, 23(13):2089–2107, 2004.

H. Hung and C.-T. Chiang. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010a.

H. Hung and C.-t. Chiang. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679, 2010b.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2 edition, 2002.

J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

M. Khan, S. Mohamed, B. Marlin, and K. Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Artificial Intelligence and Statistics*, pages 610–618. PMLR, 2012.

W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

H. Kvamme and Ø. Borgan. The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581*, 2019a.

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

H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

B. K. Lee, J. Lessler, and E. A. Stuart. Weight trimming and propensity score weighting. *PloS one*, 6 (3):e18174, 2011.

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

C. Lee, W. Zame, A. Alaa, and M. Schaar. Temporal quilting for survival analysis. In *The 22nd international conference on artificial intelligence and statistics*, pages 596–605. PMLR, 2019.

X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256, 2018.

U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23, 2012. URL https://www.jstatsoft.org/v50/i11.

S. Pölsterl. Sebastian pölsterl, Jul 2019. URL https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/.

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

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448): 1096–1120, 1999.

M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. Neumann, and H. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

T. M. Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-11.

H. Uno, T. Cai, L. Tian, and L.-J. Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

M. J. Van der Laan, M. Laan, and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

M. Wolbers, P. Blanche, M. T. Koller, J. C. Witteman, and T. A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 2014.

S. Yadlowsky, S. Basu, and L. Tian. A calibration metric for risk scores with survival data. In *Machine Learning for Healthcare Conference*, pages 424–450, 2019.

C. Zhong and R. Tibshirani. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]
(b) Did you mention the license of the assets? [N/A] The assets we use are open source.
(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our code
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]