Harmonic-Mean Cox Models: A Ruler for Equal Attention to Risk

Xuejian Wang XUEJIANW@CMU.EDU
Wenbin Zhang WENBINZHANG@CMU.EDU
Aishwarya Jadhav AJADHAV2@CMU.EDU
Jeremy C. Weiss JEREMYWEISS@CMU.EDU
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213

Abstract

Survival analysis models are necessary for clinical forecasting with data censorship. Implicitly, existing works focus on the individuals with higher risks while lower risk individuals are poorly characterized. Developing survival models to represent different risk individuals equally is a challenging task but of great importance for providing accurate risk assessments across levels of risk. Here, we characterize this problem and propose an adjusted log-likelihood formulation as the new objective for survival prognostication. Several models are then proposed based on the newly designed optimization objective function which produce risks that count individuals "equally" on risk ratios thus providing representative attention to individuals of varying risk. Extensive experiments on multiple real-world datasets demonstrate the benefits of the proposed approach.

Keywords: Survival analysis, Cox model, Clinical forcasting

1. Introduction

Clinical forecasting, as a vital part of healthcare, prognosticates the time to an event of interest for the populations at risk. One critical challenge that hinders the use of data mining and machine learning approaches in clinical forecasting is the presence of censored data, *i.e.*, that followup can be incomplete, and thus the true time to event is unknown. This motivates the study of survival analysis and the development models such as Cox (1972), Aalen (1980), accelerated failure time (Lin et al., 1998) models, random survival forests (Ishwaran et al., 2008) and so on. These models encode data in the presence of censored outcome events and are a case of the more general point process, where the condition is event-free survival. These survival methods handle the unobserved ground rates but their specified objective functions do not minimize proportional error in rate, and as a result they characterize the low risk individuals poorly.

However, in healthcare the low risk individuals for morbidity and mortality are often the majority and possess risk factors different from those at high risk. Overlooking those at low risk therefore significantly limits the predictive capabilities of existing survival models (Weiss, 2019, 2021). One direct approach could be optimizing subgroups by filtering on membership to particular risk strata. However, we will see that such a two-stage approach focuses on the highest risk individuals among those subgroups yet, within subgroup, still does not produce equal representation across individuals with varying levels of risk.

When models fail to estimate risk properly, they overemphasize high risk individuals and overfit on noise rather than on signal in low risk individuals. The basic aim is therefore to build models that produce risk that count individuals "equally", *i.e*, invariant to their level of risk. Unfortunately, it is unclear how to naturally extend the existing survival models to account people "equally" on risk ratios so as to provide equal attention to low-risk individuals.

When low-risk individuals are the majority of the population, providing accurate risk estimation for them could: (i) reduce overtreatment, (ii) lower their healthcare costs, and, (iii) improve awareness of their health status at early stages of disease . As an example, according to (Fine MJ, 1997), physicians often rely on their subjective impressions of a patient's clinical appearance in making the initial decision about the site of care, tending to overestimate the risk of death in patients with pneumonia, and these overestimates are associated with the decision to hospitalize patients at low risk. Therefore, accurate predictions for low-risk patients are critical for health care field.

We approach this problem through the formulation of the objective function attempting to mitigate misattention, which results in a reweighting program that focuses on all individuals proportionally to their proportional rate misspecification. We then propose models that minimize proportional errors in the setting of point processes and proportional hazards models. The main contributions of this paper can be summarized as follows:

- A formulation of the optimization objective function that counts individuals "equally", which is generalizable and can be used in any risk estimation models.
- New equal attention time-to-event models that attend to low-risk individuals with improved predictive ability and characterization.
- Demonstration and validation with multiple experiments and case studies.

2. Related Work

Survival analysis models time-to-event information for data with potential outcome censorship. One line of research is from traditional statistics, such as the Kaplan-Meier method (Kaplan and Meier, 1958), the Cox Proportional Hazard model (Cox, 1972) and Lee and Wang (2003) enforces the hazard to follow exponential distribution when the model is fitted to the data. Among them, closer to our work is the accelerated failure time (AFT) model (Lin et al., 1998) which implicitly minimizes proportional error. Two key distinctions are: i) our work explicitly address the problem of equal risk attention while AFT models are inattentive to low-risk individuals; ii) AFT log transformation under censorship results in a challenging mathematical specification (Jin et al., 2006) requiring advanced inference methods and a good solution is thus not guaranteed, our work, on the contrary, leads to the desired solution with simplified mathematical formulation.

Another line of effort is from the machine learning community. In recent years, deep learning has been extended to survival analysis (Katzman et al., 2018; Ren et al., 2019; Lee et al., 2018; Kvamme et al., 2019). However, similar to the approaches from the traditional statistics community, most of these methods are inattentive to low-risk individuals and

therefore also suffer from the poor characterization of lower risk individuals leading to degraded predictive capabilities.

Closely related to our work are several recent efforts adopting alternative objective functions (Avati et al., 2020; Chapfuwa et al., 2018). Another approach is to fuse together binary classification predictions across time (Yu et al., 2011). However, unlike these methods, our method provides attention to low-risk persons in the setting of anytime prediction.

While our method is general for methods that optimize for the survival objective function, we illustrate our learning methodology based on the representative approaches from the respective research directions, and devise two models that provide attention to low-risk persons and regions: that of (i) Cox-reweighting, which extends the Cox Proportional Hazard model to make survival predictions according to the newly designed optimization objective, and (ii) DeepSurv-reweighted using harmonic mean adjusted loss function. Similarly, the approach could be used in training other recent survival frameworks (Jing and Smola, 2017; Zhang et al., 2016; Lee et al., 2018; Weiss, 2019).

3. Methodology

Background. Let Y be the event we want to model over time for N samples. Let the event times be the sequence t_{in} for $i \in \{1, \ldots, T_n\}$ for $n = \{1, \ldots, N\}$ over a period of interest $[0, \tau_n]$. We are interested in modeling the rate function:

$$\lambda(t|\cdot) = \lim_{h \to 0} \frac{P(t < T < t + h | T > t, \cdot)}{h} = \frac{f(t|\cdot)}{S(t|\cdot)},$$

where $\{\cdot\}$ varies by model and represents the information or data to use in modeling λ. The probability density function and survival function are given by f and S, with the canonical relationship: $-\frac{\partial log S}{\partial t} = -\frac{\partial S/\partial t}{S(t|\cdot)} = \frac{f(t|\cdot)}{S(t|\cdot)} = \lambda(t|\cdot)$.

Next we establish the relationship between rescaled time where a single Poisson process with rate 1 and our original time, where λ is defined. This comes from the time rescaling theorem.

> **Time rescaling theorem.** Given the rate function λ, define Λ the cumulative hazard function: $\Lambda(t|\cdot) = \int_0^t \lambda(\tau|\cdot) d\tau$. For the realization of a sequence of events from $\lambda(t|\cdot)$ with times $\{u_1, \ldots, u_k\}$ and $\Lambda(t|\cdot) < \infty$, the sequence $\{\Lambda(u_1|\cdot), \ldots, \Lambda(u_k|\cdot)\}$ is distributed according to a unit rate Poisson process (Meyer, 1971; Ogata, 1981).

Details of the proof can be found in Brown et al. (2002). The implication of the theorem is that if we could model the conditional intensity correctly, the intervals between rescaled times follow exponential distribution with rate 1.

Harmonic Mean Point Processes. From the time rescaling theorem, it is straightforward to observe that the relative contributions of each time interval to the likelihood is proportional to the rate within that interval, *i.e.*, if we care about each individual's risk in a time unit equally, then we could consider decreasing the likelihood contributions in proportion to the rate. In other words, our procedure will seek to nullify, partially or fully, the proportional factor of likelihood attention given to higher rates. We call this approach optimization of the adjusted log likelihood, which is illustrated in Figure 1.

Algorithm 1: Harmonic Mean Point Process

Result: ALL-trained model
Temporal network $F : X \mapsto \hat{\lambda}(t) \in [0, \infty)$
Proportionality coefficient γ, stability factor ϵ
while *training* **do**
  $\hat{\lambda}(t'_{j,k-1}) = F(X_j)$ piecewise on $[t'_{j,k-1}, t'_{j,k}) \ \forall k \in K$
  Copy then detach $\hat{\lambda}(t'_{j,k-1}) \ \forall j, k$
  $\mathrm{ALL}_{j,k} = \mathrm{LL}_{j,k} / \big(\hat{\lambda}(t'_{j,k-1})^{\log_{10}\gamma} + \epsilon\big)$
  ALL.sum().backward()
  optimizer.step()
end

With $\mathbf{X}$, a collection of m samples, we propose the adjusted log likelihood ALL defined as follows:

$$ALL(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \Big(\sum_{i=1}^{T_n} \frac{\log \lambda(t_{in}|\Theta)}{\lambda^*(t_{in})} - \int_0^{\tau_n} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \Big),$$

where $\lambda^*(t)$ is the ground truth intensity at time t. By assuming $\lambda(t)$ and $\lambda^*(t)$ are piecewise constant, we can view the adjusted log-likelihood as the weighted sum of log-likelihood contributions. Suppose we divide the time interval $(0, \tau_j]$ into K sub-intervals where K is a significantly large number so that $\lambda^*(t)$ is constant within any sub-interval. That is, with $0 = t'_{j,0} < t'_{j,1} < \cdots < t'_{j,K} = \tau_j$, $\lambda^*(t)$ is constant for $t \in (t'_{j,k-1}, t'_{j,k}]$ and all $k = 1, \ldots, K$. Then,

$$\begin{aligned} ALL(\mathbf{X}|\Theta) &= \sum_{j=1}^{m} \left[\int_0^{\tau_j} \frac{\log \lambda(t|\Theta)}{\lambda^*(t)} dN(t) - \int_0^{\tau_j} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \right] \\ &= \sum_{j,k=1}^{m,K} \left[\int_{t'_{j,k-1}}^{t'_{j,k}} \frac{\log \lambda(t|\Theta)}{\lambda^*(t)} dN(t) - \int_{t'_{j,k-1}}^{t'_{j,k}} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \right] \\ &= \sum_{j,k=1}^{m,K} \frac{1}{\lambda^*(t'_{j,k})} \left[\int_{t'_{j,k-1}}^{t'_{j,k}} \log \lambda(t|\Theta) dN(t) - \int_{t'_{j,k-1}}^{t'_{j,k}} \lambda(t|\Theta) dt \right]. \end{aligned}$$

is the sum of log-likelihood contributions in time intervals $(t'_{j,k-1}, t'_{j,k}]$ weighted by the reciprocal of the ground truth intensity at $t'_{j,k}$ for every j and k. The similarity is apparent when it is compared to the similar form of standard log-likelihood:

$$LL(\mathbf{X}|\Theta) = \sum_{j,k=1}^{m,K} \left[\int_{t'_{j,k-1}}^{t'_{j,k}} \log \lambda(t|\Theta) dN(t) - \int_{t'_{j,k-1}}^{t'_{j,k}} \lambda(t|\Theta) dt \right].$$

Therefore, we can weight each interval's log likelihood by the inverse of the oracle rate to get the adjusted log likelihood.

Proportional hazards models. Because of the separation of time and feature contributions to the hazard function, Cox models focus on optimizing for the parameters of

Algorithm 2: Harmonic Mean Cox Model

Result: APLL-trained model
Feature network $G : X \mapsto \hat{h} \in [0, \infty)$
Proportionality coefficient γ, stability factor ϵ
while *training* **do**
  $\hat{\lambda}(t) = \lambda_0(t)e^{\hat{h}}$
  Copy then detach $\hat{h}$
  $\text{ALL} = \text{LL}/\left((e^{\hat{h}})^{\log_{10}\gamma} + \epsilon\right)$
  ALL.sum().backward()
  optimizer.step()
end

the features in isolation, the so-called semi-parametric approach. Define $\lambda(t) = \lambda_0(t)e^{h(\cdot)}$, where $h(\cdot)$ denotes the contribution of features to the hazard function. In the Cox model, $h(\cdot) = \theta^\top x$. In DeepSurv, $h(\cdot) = g(x, \theta)$ is the output of some neural network g parameterized by θ.

The partial likelihood is given by the product of the probability at time t_n that the event occurred for individual n, given the at risk set $\mathbf{n}_n, \mathbf{n}_n \subseteq \{1, \ldots, N\}$. Define the failure event indicator $Y = 1$ to indicate failure and $Y = 0$ to indicate censorship. Then with parameters Θ, the partial log likelihood is given by:

$$PLL(X|\Theta) = -\sum_{n:Y_n=1} \left(h_\theta(\cdot) - \log \sum_{\mathbf{n}_n} e^{h_\theta(\cdot)}\right).$$

From above, both the hazard λ and cumulative hazard Λ are proportional to $e^{h_\theta(\cdot)}$. Therefore, the oracle equal attention adjustment $1/\lambda^*(t)$ to the partial likelihood, under assumption that the oracle belongs to the semi-parametric model family, *i.e.*, $\lambda^*(t) = \lambda_0^*(t)e^{h^*}$, is the feature-based component of the full likelihood, which is constant over time: e^{-h^*}. Thus, we can write the adjusted partial log likelihood (APLL):

$$APLL(X|\Theta) = \quad -\sum_{n:Y_n=1} \left(h_\theta(\cdot) - h^* - \log \sum_{\mathbf{n}_n} e^{-h^*} e^{h_\theta(\cdot)}\right).$$

Oracle Approximation. Without access to λ^*, we must resort to some approximation of the reweighting. One choice for λ^* is our current estimate $\hat{\lambda}$. However, this could lead to unstable weightings because a single example could dominate the weight distribution. To address this fairness-variance tradeoff, we introduce the proportionality coefficient γ and stability factor ϵ to help stabilize the weights. Algorithm 1 illustrates the training procedure and the stabilization modification. Similarly, the APLL corresponds to reweighted samples, where each individual n is reweighted according to e^{-h^*} instead of each weighted as unity. Therefore, in practice, we can weight the examples according to the inverse of the hazard to pose an equivalent optimization program (Algorithm 2).

We call our methods harmonic mean point processes (HMPP) and harmonic mean Cox models (HMCM) because if the oracle is known and doubly stochastic, *i.e.*, frail, then

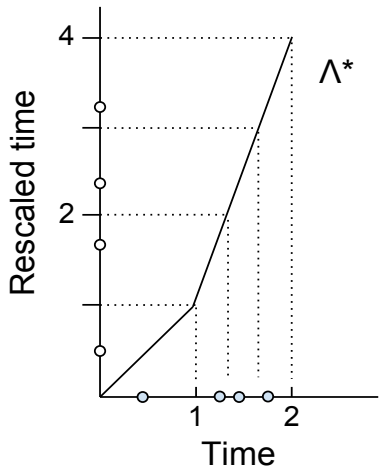


Figure 1: A unit of time with $\lambda = 1$ has three times less likelihood weight than a unit with $\lambda = 3$, evidenced by its proportion of the length on the y-axis.

the estimates we get from the training procedure are harmonic mean estimates of the rate distribution. In practice, when we use an dynamic graph for gradient updates, the denominator must be copied and detached from the computation graph so that the graph of which λ is a part is not further connected by the current model's predictions $\hat{\lambda}$ and $\hat{h}$ (Algorithms 1 and 2).

4. Experiments

In this section, we evaluated the performance of our proposed reweighting adjustment on several popular benchmark datasets and discussed the results.

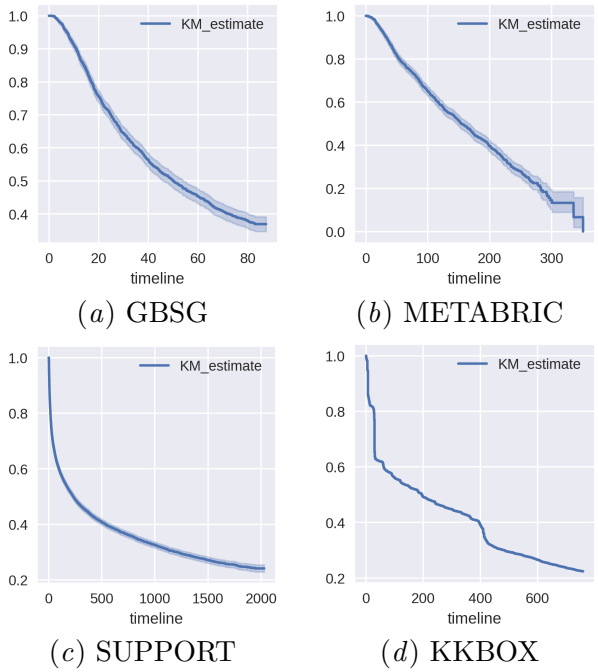
Datasets	Size	Censored Data #	Censored Rate	Feature #	MST
GBSG	2,232	965	0.432	7	50
METABRIC	1,904	801	0.420	9	154
SUPPORT	8,873	2,837	0.320	14	231
KKBOX	2,814,735	975,834	0.347	18	195

Table 1: Summary of the four datasets. # stands for number and MST stands for Median Survival Time from the Kaplan-Meier estimator.

4.1. Datasets

We evaluated the proposed models on four real-world datasets whose characteristics are summarized in Table 1 with corresonding Kaplan-Meier curve depicted in Fig. 2.

GBSG is a dataset by the Rotterdam & German Breast Cancer Study Group (Katzman et al., 2018). It contains 2232 patients and nearly 10% censored rate in training dataset and 56% censored rate in test dataset.



(a) GBSG (b) METABRIC

(c) SUPPORT (d) KKBOX

Figure 2: The Kaplan-Meier curves of four different datasets.

METABRIC is by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Katzman et al., 2018), which contains 1904 patients.

SUPPORT stands for Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) (Katzman et al., 2018), which contains 8873 patients.

KKBOX is a survival dataset created from the WSDM - KKBox's Churn Prediction Challenge 2017 with administrative censoring (Kvamme et al., 2019). KKBOX offers streaming music subscription and users can cancel their memberships at any time. To "churn" means a user stops his/her membership. It has around 2.8 million samples and is the largest dataset we experimented with.

4.2. Evaluation Metrics

4.2.1. Concordance index

Concordance index (C-index) is a widely used evaluation metric in survival analysis (Harrell et al., 1982), which equals the area under ROC curve (AUC) in the absence of censorship. It is an assessment of model predictions in terms of correct pairwise ordering among the failure times. For example, if for two samples $d_1 = (\mathbf{x}_1, y_1), d_2 = (\mathbf{x}_2, y_2)$, where $\mathbf{x}$ stands for features and y stands for the corresponding time-to-event, our prediction indicates $\hat{y}_1 < \hat{y}_2$ is true then the model get 1 point for this pair, 0 if the predicted ordering is incorrect. Concordance index calculates such scores for every possible pair and outputs the average score. Formally, we define the concordance index as

$$\text{C-index} = Pr\{\hat{S}(T_i|\mathbf{x}_i) < \hat{S}(T_i|\mathbf{x}_j)|T_i < T_j, E_i = 1\},$$

where $\hat{S}$ is the predicted survival probability, T is the death time, E stands for censored or not.

4.2.2. Brier score

Brier score (BS) (Brier and Allen, 1951) is another metric which both attends to discrimination and calibration of survival analysis models. It is roughly the mean squared error of the probability estimations and the real binary labels, e.g. death or not in survival analysis. Note that unlike concordance index, the lower the Brier score is, the better. For dataset with censorship, the Brier score is calculated as

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{S}(t|\mathbf{x}_i)^2\mathbb{1}\{T_i \le t, D_i = 1\}}{\hat{G}(T_i)} + \frac{(1-\hat{S}(t|\mathbf{x}_i))^2\mathbb{1}\{T_i > t\}}{\hat{G}(T_i)}\right],$$

where N is the number of samples and $\hat{G}(t)$ is the Kaplan-Meier estimate of the censoring survival function.

4.3. Compared Method

We aim to compare models with proportional hazard assumption and adjust them with our proposed loss function to achieve more accurate risk estimation for low risk people.

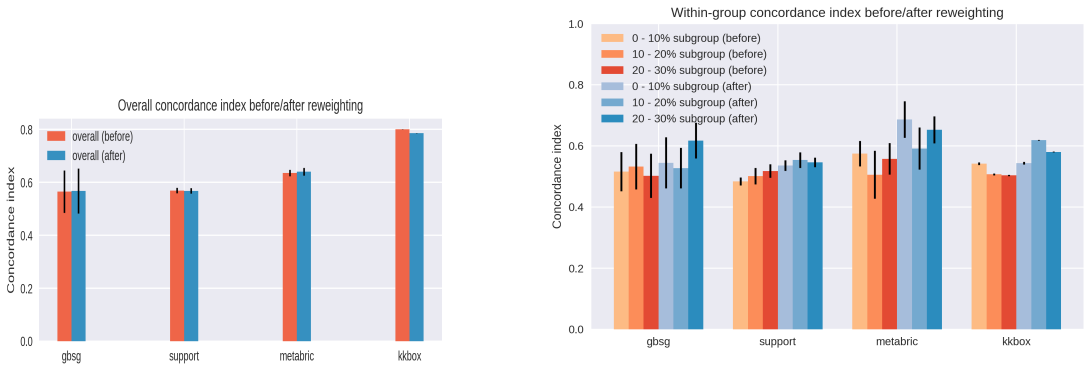
- **Cox** (1972) is a semi-parametric model which is commonly used in survival analysis.
- **Cox-reweighted** is our adjusted cox model with reweighting. We started with a weight vector of ones and retrain the model with weight assigned by the reciprocals of last risk score predictions. We train using Lifelines CoxPHFitter for 10 epoches or until convergence.
- **DeepSurv** (Katzman et al., 2018) is a Cox proportional hazard model equipped with deep neural network for modeling feature interactions. The loss function is similar to Cox with L2 regularization:

  $$l(\theta) = -\frac{1}{N_{E=1}}\sum_{i:E_i=1}\left(\hat{h}_\theta(\mathbf{x}_i) - \log\sum_{j\in\mathcal{R}(T_i)} e^{\hat{h}_\theta(\mathbf{x}_j)}\right) + \lambda\cdot||\theta||_2^2$$

  where $N_{E=1}$ is the number of sample with events, $\hat{h}_\theta(\mathbf{x}_i)$ is the risk output of the neural network and $\mathcal{R}(T_i)$ stands for all the remaining patients at time T_i.
- **DeepSurv-reweighted** is our adjusted DeepSurv model with reweighting. We modified the loss function of DeepSurv with the approach introduced previously.

4.4. Reporting of Results

We present our results in subgroups divided by risk to investigate how our reweighting techniques work in low risk subgroups which we care more about. We divide the whole dataset into 10 subgroups, where the 0 - 10% subgroup is the subgroup with lowest risk and 90 - 100% subgroup is the subgroup with highest risk. Furthermore, we performed



(a) (b)

Figure 3: Cox: (a) the concordance index before/after reweighting for four datasets. (b) the within-group concordance index of the lowest 0 - 30% risk subgroups before/after reweighting for four datasets.

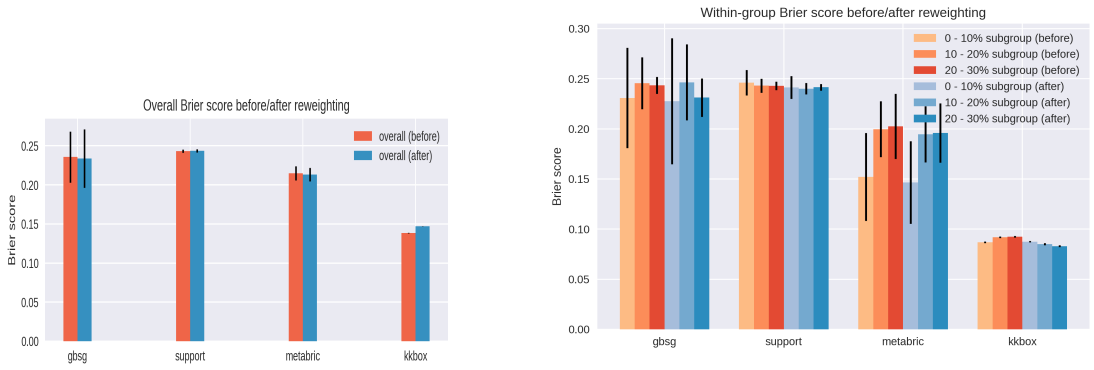
a 5-fold cross validation and report the mean and standard deviation of the concordance index for the lowest 0 - 30 percentile subgroups which we care most about. Using other numbers of subgroups gives similar results. The error bar indicates the standard deviation of the concordance index in the 5-fold cross validation and the height stands for the mean concordance index.

4.5. Results

Cox and Cox-reweighted Our Cox-reweighted model achieves higher concordance index within each subgroup (Figure 3(b)). With the adjusted loss function and reweighting technique, it outperformed the vanilla Cox model by 8.4% in the subgroup with lowest risk, and 6.0%/6.8% in 10 - 20%/20 - 30% low risk subgroup, respectively, averaging across the four datasets. Specifically, KKBOX is the largest dataset with over two million samples where our Cox-reweighted model stably outperforms the Cox model in all three low risk subgroups by a large margin, which further demonstrates our approach. At the same time, the overall concordance index on the whole datasets remains nearly the same before and after weighting (Figure 3(a)), proving that the reweighting won't degrading overall performance.

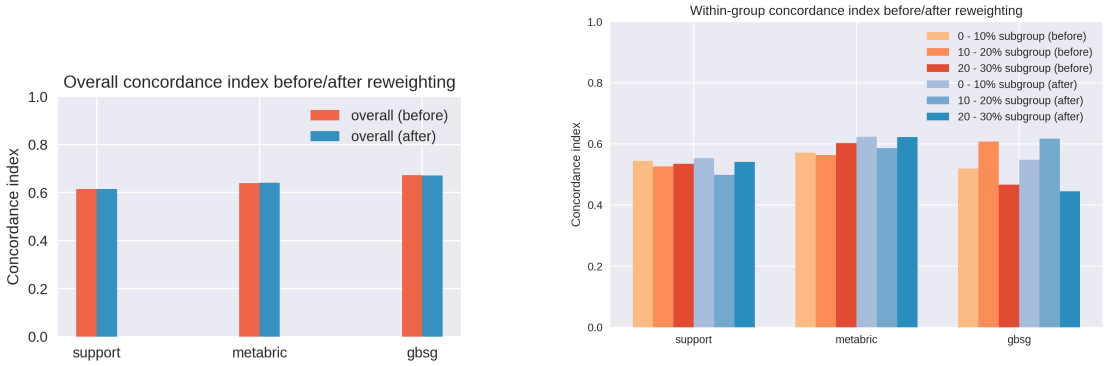
In terms of Brier score, the Cox-weighted achieved lower scores in 0 - 30% subgroups across the four dataset we compared (Figure 4(b)). Again, KKBOX has the most stable Brier score output due to its large size. The Cox-reweighted achieved 1.5%/2.6%/4.8% improvement over Cox model for 0-10%/10-20%/20-30% risk subgroups respectively. Similarly, the overall Brier scores on the four datasets stayed nearly the same, which shows our adjustment improves the prediction for low risk people while maintaining the overall performance (Figure 4(a)).

DeepSurv and DeepSurv-reweighted The within-group concordance index of the lowest 0 - 30% risk subgroups demonstrate improvements in each comparable low-risk decile



(a) (b)

Figure 4: Cox: (a) the overall Brier score before/after reweighting for four datasets. (b) the within-group Brier score of the lowest 0 - 30% risk subgroups before/after reweighting for four datasets (the lower, the better).



(a) (b)

Figure 5: DeepSurv: (a) the concordance index before/after reweighting for three datasets. (b) the within-group concordance index of the lowest 0 - 30% risk subgroups before/after reweighting for three datasets.

(Figure 5(b)), in fact by 5% in the lowest-risk decile. DeepSurv-reweighted did not degrade the performance on overall dataset (Figure 5(a)).

In addition, theoretically, the formulation of equally attentive optimization objective function of our approaches is expected to have features associated with the outcome differently to the association of inattention to low-risk individuals methods. The hazard ratio forest plot relating the features with the outcome (Figure 6) experimentally verifies this. Note that due to space constraint, only the forest plot on kkbox dataset is shown.

5. Discussion

We summarize our contributions as the following three points:

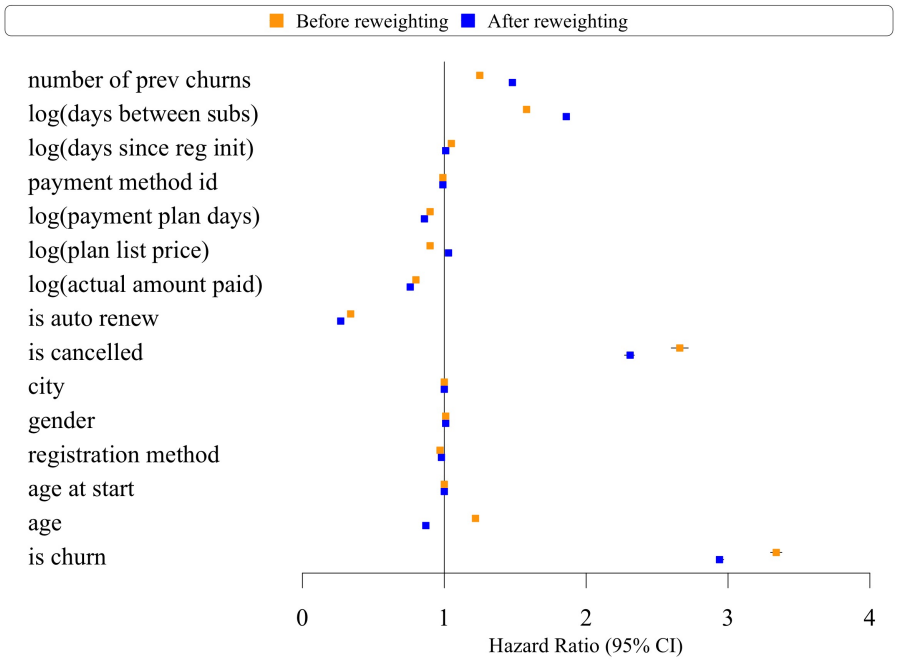


Figure 6: Forest plot of hazard ratios and 95% confidence intervals before and after reweighting on the KKBOX dataset.

- We proposed using harmonic mean to adjust the optimization objective function for models obeying proportional hazard assumption such as Cox and DeepSurv, to count individuals "equally" and achieve more accurate risk prediction for low risk people.
- We provided the analysis and pseudocode for our harmonic mean approach.
- We performed the reweighting techniques on Cox and DeepSurv models and demonstrated our proposed adjustment with experiments on four real-world datasets. The results has achieved 8.4% improvement on Cox and 5% on DeepSurv in terms of concordance index for the lowest 10% risk people.

Future work includes (1) the application of our reweighting techniques to additional semi-parametric models; (2) the integration of our approach with other reweighting methods; and, (3) the application to large-scale healthcare datasets such as MIMIC-III and eICU to perform case studies where low-risk individuals are the majority, but where the consequences of even rare outcomes mean that risk assessments should not overlook them.

References

Odd Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, 1980.

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Y Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pages 145–155. PMLR, 2020.

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

Emery N Brown, Riccardo Barbieri, Valérie Ventura, Robert E Kass, and Loren M Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin Duke, and Ricardo Henao. Adversarial time-to-event modeling. In *International Conference on Machine Learning*, pages 735–744, 2018.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Medsger AR Li YH Ricci EM Singer DE Marrie TJ Coley CM Walsh MB Karpf M Lahive KC Kapoor WN Fine MJ, Hough LJ. The hospital admission decision for patients with community-acquired pneumonia. results from the pneumonia patient outcomes research team cohort study. *Arch Intern Med.*, 1997.

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

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Zhezhen Jin, DY Lin, and Zhiliang Ying. On least-squares regression with censored data. *Biometrika*, 93(1):147–161, 2006.

How Jing and Alexander J Smola. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 515–524, 2017.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

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

Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

DY Lin, LJ Wei, and Zhiliang Ying. Accelerated failure time models for counting processes. *Biometrika*, 85(3):605–618, 1998.

Paul-André Meyer. Demonstration simplifiee d'un theoreme de knight. In *Séminaire de probabilités v université de strasbourg*, pages 191–195. Springer, 1971.

Yosihiko Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. 2019.

Jeremy C Weiss. On microvascular complications of diabetes risk: development of a machine learning and electronic health records risk score. 2019.

Jeremy C Weiss. Wavelet reconstruction networks for marked point processes. In *AAAI Symposium on Survival Prediction*, 2021.

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

Wenbin Zhang, Jian Tang, and Nuo Wang. Using the machine learning approach to predict patient survival from high-dimensional survival data. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1234–1238. IEEE, 2016.