Flexible Group Fairness Metrics for Survival Analysis

RAPHAEL SONABEND*, Technische Universität Kaiserslautern, Germany, University of Cambridge, UK, and Imperial College London, UK
FLORIAN PFISTHER*, Ludwig-Maximilians-Universität München, Germany
ALAN MISHLER, J.P. Morgan AI Research, USA
MORITZ SCHAUER, Chalmers Technical University and University of Gothenburg, Sweden
LUKAS BURK, Ludwig-Maximilians-Universität München, Germany and Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH, Germany
SUMANTRAK MUKHERJEE, Birla Institute of Technology and Science, India
SEBASTIAN VOLLMER, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany, Technische Universität Kaiserslautern, Germany, and University of Warwick, UK

Algorithmic fairness is an increasingly important field concerned with detecting and mitigating biases in machine learning models. There has been a wealth of literature for algorithmic fairness in regression and classification however there has been little exploration of the field for survival analysis. Survival analysis is the prediction task in which one attempts to predict the probability of an event occurring over time. Survival predictions are particularly important in sensitive settings such as when utilising machine learning for diagnosis and prognosis of patients. In this paper we explore how to utilise existing survival metrics to measure bias with group fairness metrics. We explore this in an empirical experiment with 29 survival datasets and 8 measures. We find that measures of discrimination are able to capture bias well whereas there is less clarity with measures of calibration and scoring rules. We suggest further areas for research including prediction-based fairness metrics for distribution predictions.

1 INTRODUCTION

The use of machine learning (ML) models, especially in the context of clinical decision making [72] can lead to, or exacerbate, disparities in health outcomes for marginalized populations [55, 56, 79]. This can arise due to multiple reasons, such as access to different standards of care [4], historical inequity [17, 28] or under-representation in data collection [13, 45]. This can lead to models with differing (predictive) effectiveness depending on the subpopulation or models that perpetuate historical injustices if they are subsequently used to inform medical decisions [78]. The goal of algorithmic fairness is to detect and mitigate such biases [6, 50]. This has been discussed in great detail in classification and regression settings [50, 57, 61]; however very little discussion exists for survival analysis. This is problematic given the sensitive nature of survival predictions. For example, hospitals with insufficient resources may require survival models to accurately and fairly rank patient outcome risks [46, 63, 68]. It is crucial that algorithmic fairness is considered in the survival setting. Zhang and Weiss [83] have begun exploring debiasing methods for survival analysis, however only a limited number of measures are considered. In this article we examine whether existing metrics that are used in survival analysis can be adapted to detect unfairness in survival models. The code required to reproduce the results in this paper is publicly available at https://github.com/Vollmer-Lab/survival_fairness. For a comprehensive overview to fairness and survival analysis we recommend Mitchell et al. (2021) [51] and Wang et al. (2019) [80] respectively.

2 RELATED WORK

2.1 Fairness metrics

Many notions of fairness exist, including individual fairness [22], causal/counterfactual fairness [18, 42, 82], group fairness, and intersectional fairness. Individual fairness measures require defining a metric space that encodes differences between individuals [22] and it is unclear in general how such a metric should be chosen. Causal fairness measures require defining causal relationships between covariates and protected attributes and outcomes, for example in the form of a directed acyclic graph [42], which is especially challenging to construct in high dimensional settings. In this paper we focus on group fairness definitions. These metrics have the advantage that they can be measured without causal assumptions and without defining a metric over individuals.

Discrimination criteria can be understood as measuring adherence to one of the following independence statements defined based on a model’s predicted score or class, $R$, protected attribute, $S$, and target variable, $Y$: Independence: $R \perp S$, Separation: $R \perp S \mid Y$ and Sufficiency $Y \perp S \mid R$ [7]. We will refer to metrics that require
one of these three independence statements above as (statistical) group fairness metrics since independence is observed at the level of the protected groups. Group fairness metrics are usually defined for the classification setting, however many of them naturally lend themselves to regression scenarios or can be adapted simply [69]. Intersectional fairness [13] extends group fairness for a more fine-grained, and intersectional, assessment of bias. For example, group fairness measures may assess if a dataset is unbiased across race and gender, whereas an intersectional fairness measure will also assess if the interaction between race and gender are also unbiased.

2.2 Fairness metrics in survival analysis

Applications of fairness metrics in clinical decision making have been studied by Pfohl et al. [57], however this is restricted to the classification setting. Whilst classification metrics may be directly extended to the regression setting, the same does not hold for the more complex survival setting. This is due to: 1) time-to-event datasets including censoring, i.e. patients who are not observed to experience the event of interest; and 2) the prediction of interest is a distribution and not a single value.

We could find only two strategies evaluating survival fairness in the literature. The first is a transformation of the survival objective, the second is metric based; we briefly discuss each in turn. One approach to evaluating fairness in a survival setting is to assess fairness for binary survival predictions at fixed time points [5], e.g. three years in the future. This strategy can yield viable and fair predictors if the selected horizon perfectly coincides with the time-point at which decisions are made. However, this is generally not the case and therefore evaluation of such models usually results in over-confidence in model performance due to ‘improper’ evaluation [10], thus this strategy is not fair. On the other hand, Keya et al [38] proposed predictive metrics – individual, group, and intersectional – for survival fairness that do not require any objective transformation. Whilst this is a great step forward, their metrics are only applicable to linear predictors, such as from the Cox Proportional Hazards (PH) model [19]. Whilst the Cox PH is arguably the most popular model in survival analysis, this limitation means that their metrics cannot be applied to the majority of machine learning models.

3 FAIRNESS IN SURVIVAL ANALYSIS

Survival analysis is a task in which one attempts to predict the probability of an event occurring over time. For example, predicting the risk of a patient dying of a disease after diagnosis, predicting when a customer will default on a loan (‘duration analysis’), or predicting the probability of a lightbulb failing over time (‘reliability analysis’). Survival analysis is distinct from regression as we are interested in making predictions from censored time-to-event data. A censored observation is one in which the event of interest does not occur. Survival models are fit to estimate the functional relationship between a set of covariates X and time until an event of interest takes place Y. We assume, for all observations, that there exists both a hypothetical survival time Y and censoring time C (the last recorded time for an observation). We define \( T := \min(Y, C) \) as the observed outcome time and
\( \Delta := I(Y = T) \) as the survival indicator. Survival models are fit on the survival tuple \((X,T,\Delta)\). For this paper we only consider the right-censoring survival setting. Data is assumed to consist of \( n \) observations drawn i.i.d. from a data-generating distribution \( P_{XY} \). In fairness contexts we further assume there exists one or multiple sensitive attributes, \( S \), assigning groups to each observation.

3.1 Experiment

Analysis. We are interested in understanding how well existing losses capture bias in survival datasets. We apply biasing algorithms to 29 published survival datasets (Appendix D), fit a random survival forest (RSF) [33] on these biased datasets and then evaluate fairness as \( F_L = |L_D - L_A| \) where \( F_L \) is the fairness measured by loss \( L \) and \( L_A, L_D \) are the model performance on the advantaged and disadvantaged subgroups respectively measured by loss \( L \). Analysis is performed with mlr3proba [65] in R [60]. We fit an RSF as it robustly returns multiple prediction types [64] that can be assessed with calibration, discrimination and scoring rule measures. We increased the proportion, \( \sigma \), of biased observations in the disadvantaged data from 0% to 90% and asserted that a loss, \( L \), could capture bias in our datasets if there was a significant Spearman rank correlation between \( F_L \) and \( \sigma \), and a significant t-test after regressing \( F_L \) on \( \sigma \).

Biasing algorithms. We created biasing methods that mimic identified real-world sources of bias [49]. Our first method simulates measurement bias by randomly permuting the covariates for an increasing proportion of disadvantaged observations for each dataset, whereas our second method simulates representation bias by increasingly undersampling disadvantaged groups for each dataset. These are fully described in Appendix A.

Measures. We consider a range of calibration, discrimination, and scoring rule measures common in the survival literature. We avoid evaluation bias [49] by only including (strictly) proper scoring rules as other (improper) scoring rules may not accurately identify a superior model over an inferior one. We consider the right-censored log-likelihood (RCLL) [3], reweighted Brier score (RSBS) [25, 67], reweighted integrated survival logloss (RISL) [25, 67], survival negative log-likelihood (SNL) [67], van Houwelingen’s alpha (CalA) [75], D-calibration (CalD) [27], Harrell’s C (C_H) [29, 30], and Uno’s C (C_U) [74]. Scoring rules are standardised against a Kaplan-Meier baseline [26, 36].

Results. We now assess how well the metrics recover increased unfairness, controlled by \( \sigma \). Regressing \( F_L \) on \( \sigma \) we find that for the permutation biasing method, \( \sigma \) was a significant predictor of \( F_L \) for RSBS, RISL, C_H, C_U, CalA. We also find that for RCLL there is a significant correlation between \( F_L \) and \( \sigma \) however the regression slope is too small to be meaningful. There is also a significant relationship between \( F_L \) and \( \sigma \) for SNL after applying the undersampling algorithm (Table 1 and Appendix C).

3.2 Discussion

There is a significant relationship between \( \sigma \) and \( F_L \) for both measures of separation, \( C_H \) and \( C_U \). This is intuitive as the biasing methods prevent the model learning the true risk for disadvantaged
Table 1. Let \( F_2 = \alpha + \beta \beta \) be our regression model. Table shows intercept, \( \alpha \), slope, \( \beta \), and Spearman rank correlation (\( \rho \)) for each of the measures and permutation (left) and undersampling (right) biasing methods. An ‘-’ indicates a p-value less than 0.05 after Holm’s correction.

| Measure | \( \alpha \)  | \( \beta \)  | \( \rho \)  | \( \alpha \)  | \( \beta \)  | \( \rho \)  |
|---------|----------------|----------------|----------------|----------------|----------------|----------------|
| RSBS    | 0.049          | 0.078*         | 0.976*         | 0.035          | 0.066*         | 0.855*         |
| RSL     | 0.045          | 0.063*         | 0.976*         | 0.033          | 0.058*         | 0.891*         |
| SNL     | 0.018          | 0.001          | 0.248          | 0.014          | 0.083*         | 1.000*         |
| RCLL    | 0.018          | 0.009          | 0.879*         | 0.022          | 0.088*         | 1.000*         |
| CH      | 0.024          | 0.129*         | 1.000*         | 0.017          | 0.083*         | 1.000*         |
| CY      | 0.031          | 0.124*         | 1.000*         | 0.024          | 0.078*         | 0.976*         |
| CalA    | 0.027          | 0.011*         | 0.891*         | -0.015         | 0.197*         | 0.952*         |
| CalD    | 2.686          | 0.487          | 0.721*         | 2.861          | 0.646          | 0.612          |

observations and therefore cannot estimate the difference in risk between advantaged and disadvantaged observations. Secondly, CalA does detect bias in the data whereas CalD does not. CalA evaluates if a model correctly predicts the number of events in the test set. In contrast, CalD evaluates if the predicted survival functions are distributed according to \( U(0,1) \). The significant result with CalA indicates that the model cannot predict the number of events in the disadvantaged groups from either biasing method. Whereas the results with CalD may indicate a more complex relationship between distributional calibration and fairness, which has already been demonstrated in the classification setting [58]. Of the scoring rules, only RSBS and RSL could detect the bias from both methods with a regression slope of a meaningful magnitude. This is a promising finding as prior research has demonstrated the usefulness of scoring rules in evaluating fairness [24].

4 CONCLUSIONS

Algorithmic fairness is an important concept to assess how much bias is present in datasets and subsequently picked up by models. This is especially important in survival analysis, which often overlaps with areas that require strong ethical consideration. Despite this, the literature around survival fairness is in its infancy. In this paper we have performed a simple experiment to demonstrate how existing survival measures can be utilised to audit bias in algorithmic fairness. Measures of discrimination appear to be optimal for capturing bias however these do not paint a full picture as they ignore model calibration. We have found that the standardised scoring rules, RSBS and RSL, are interpretable, capture both calibration and discrimination, and can detect the bias from our algorithms. We believe future work should consider more complex biasing methods including temporal methods that introduce bias after a certain timepoint. Finally, whilst predictive metrics have been proposed for risk predictions [38], these have yet to be reviewed in the literature and there is also potential to extend this work to survival distribution predictions. Our paper should help raise awareness that current methods of measuring fairness in survival are very limited and should stimulate interest and exploration in further development.

References

[1] Alina M Allen, Terry M Therneau, Joseph J Larson, Alexandra Coward, Virend K Somers, and Patrick S Kamath. 2018. Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year-community study. Hepatology (Baltimore, Md.) 67, 5 (may 2018), 1726–1736. https://doi.org/10.1002/hep.29546
[2] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. 2012. Statistical models based on counting processes. Springer Science & Business Media.
[3] Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Ng. 2018. Countdown Regression. Sharp and Calibrated Survival Predictions. (jun 2018). arXiv:1806.08324 http://arxiv.org/abs/1806.08324
[4] Zinzi D Bailey, Nancy Krueger, Madina Agnem, Jasmine Graves, Natalia Linos, and Mary T Bassett. 2017. Structural racism and health inequities in the USA: evidence and interventions. The Lancet 389, 10077 (2017), 1453–1463.
[5] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Pinkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Soemer, Eitan Bachmat, et al. 2020. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. Nature communications 11, 1 (2020), 1–9.
[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. Fairbook.org. http://www.fairbook.org.
[7] Solon Barocas and Andrew Selbst. 2016. Big Data’s Disparate Impact. California Law Review 104, 1 (2016), 671–729. https://doi.org/10.15779/Z38BG31
[8] Andreas Bender and Fabian Scheipl. 2018. pamtools: Piece-wise exponential Additive Mixed Modeling tools. arXiv:1806.01042 [stat] (2018). http://arxiv.org/abs/1806.01042
[9] Andreas Bender, Fabian Scheipl, Wolfgang Hartl, Andrew G Day, and Helmut Küchenhoff. 2018. Penalized estimation of complex, non-linear exposure-lag-response associations. Biostatistics 20, 2 (feb 2018), 315–331. https://doi.org/10.1093/biostatistics/kxy003
[10] Paul Blanche, Michael W Kattan, and Thomas A Gerds. 2019. The c-index is not proper for the evaluation of 1-year predicted risks. Biostatistics 20, 2 (apr 2019), 327–337. https://doi.org/10.1093/biostatistics/kxy066
[11] N E Breslow and N Chatterjee. 1999. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48, 4 (jan 1999), 457–468. https://doi.org/10.1111/1467-9876.00165
[12] Göran Broström. 2021. Flexible Group Fairness Metrics for Survival Analysis DSHealth 2022, August 14, 2022, Washington DC
[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. PMLR, 77–91.
[14] Chao Cai, Yubo Zou, Yingwei Peng, and Jiajia Zhang. 2012. smcure: Fit Semiparametric Mixture Cure Models. https://CRAN.R-project.org/package=smcure R package version 2.0.
[15] Bradley P Carlin and Thomas A Louis. 2018. Supplemental Materials to Bayesian Methods for Data Analysis, 3rd Edition (3 ed.). Chapman and Hall/CRC.
[16] Daniel P. Carpenter. 2002. Groups, the Media, Agency Waiting Costs, and FDA Drug Approval. American Journal of Political Science 46, 3 (2002), 490–505. http://www.jstor.org/stable/3088394
[17] Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi. 2020. Treating health disparities with artificial intelligence. Nature medicine 26, 1 (2020), 16–17.
[18] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 7801–7808.
[19] D. R. Cox. 1972. Regression Models and Life-Tables. The Lancet 2, 6737 (1972), 389–391. https://www.jstor.org/stable/2404731
[20] Angela Dispensieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph

Disclaimer. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JP Morgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

Flexible Group Fairness Metrics for Survival Analysis

DSHealth 2022, August 14, 2022, Washington DC
We have a generic biasing algorithm (Algorithm 1) that is then specialised with two biasing techniques. The algorithm splits the dataset into two equal-sized smaller datasets, one to add bias to ($D_B$) and one to leave untouched ($D_U$). Then we further split $D_B$ into two more datasets, $D_{BA}, D_{BD}$, such that $\sigma$ of $D_B$ are in $D_{BD}$. The specific biasing algorithm, $G$, is then applied to $D_{BD}$. $D_{BA}$ and $D_{BD}$ are recombined and then three-fold cross-validation is utilised to evaluate model performance on $D_B$ and $D_U$ and fairness is computed. By splitting the data in this way we are able to: 1) ensure that the proportion of bias is simply controlled; 2) ensure that the model performance does not deteriorate for the unbiased datasets, i.e. if we do not split the original dataset in $D_B$ and $D_U$ then model performance will deteriorate overall due to the bias added to the disadvantaged group, whereas by splitting the dataset we are able to clearly distinguish between the ‘normal’ model performance (when no bias has been added), and the reduced performance due to the added bias.

**Biasing method 1 - Permutation.** For the first method we artifically add bias by randomly permuting the covariates of disadvantaged observations. This breaks the relationship between the covariates and outcomes and mimics the real-world problem of data being lower-quality for disadvantaged groups of people.

**Biasing method 2 - Undersampling.** For the second method we artifically add bias by undersampling disadvantaged observations. In Algorithm 1 this amounts to deleting all observations in $D_{BD}$. This mimics the real-world problem of not capturing enough data for disadvantaged groups of people.

**Algorithm 1** Input: $D = (X, Y, T, \Delta)$, survival dataset. $\sigma$, proportion of disadvantaged observations to bias. $L$, survival loss. $M$, survival model. $G$, biasing algorithm. Output: $F_L$, fairness metric.

1. \[ \text{for } i = 1, ..., 10 \text{ do} \]
2. \[ D_{BU} \leftarrow \text{Randomly split } D \text{ into equal-sized datasets} \]
3. \[ D_{BA}, D_{BD} \leftarrow \text{Randomly split } D_B \text{ w.p. } \sigma \]
4. \[ \text{Apply } G \text{ to } D_{BD} \]
5. \[ D_B \leftarrow (D_{BA}, D_{BD}) \]
6. \[ L_B \leftarrow 3\text{-fold CV of } D_B \text{ with model } M \text{ and loss } L \]
7. \[ L_U \leftarrow 3\text{-fold CV of } D_U \text{ with model } M \text{ and loss } L \]
8. \[ F_{Li,j} \leftarrow |L_B - L_U| \]
9. \[ \text{end for} \]
10. \[ F_L \leftarrow \sum_i F_{Li,j}/10 \]
11. \[ \text{return } F_L \]

## B Metric Definitions

Let $X, Y, C$ be random variables taking values in $X \subseteq \mathbb{R}^p, Y \subseteq \mathbb{R}, C \subseteq \mathbb{R}$. In particular, define $T := \min(Y, C)$, and $\Delta := I(Y > T)$. Finally let $(X_0, Y_0, C_0, T_0, \Delta_0)_{i=1}^{14d}$ $(X, Y, C, T, \Delta)$. We define the following measures and include a brief explanation of their interpretation and use.

- Reweighted survival Brier score (RSBS) \cite{25, 67}

\[
L_{RSBS}(Y, T, \Delta|\hat{G}_{KM}) = \frac{\int_T^\infty (I(T \leq \tau) - F_Y(\tau))^2 \, d\tau}{\hat{G}_{KM}(T)}
\]  
(1)

This is a strictly proper approximate survival loss that evaluates a survival distribution prediction by measuring the squared distance between the predicted survival probability and whether the event occurs. It is the CRPS with an inverse probability of censoring weighting (IPCW) adjustment.

- Reweighted integrated survival logloss (RISL) \cite{25, 67}

\[
L_{RISL}(Y, T, \Delta|\hat{G}_{KM}) = \frac{\int_T^\infty \int_0^T I(T \leq \tau) \log[f_Y(\tau)] + I(T > \tau) \log[1 - S_Y(\tau)] \, d\tau}{\hat{G}_{KM}(T)}
\]  
(2)
This is a strictly proper approximate survival loss that evaluates a survival distribution prediction by measuring the logarithm of the predicted survival (or 1-survival) probability and whether the event occurs. It is the integrated loglikelihood with an inverse probability of censoring weighting (IPCW) adjustment

- Survival negative log-likelihood (SNL) \[ 67 \]
  \[
  L_{SNL}(Y, T, \Delta | \hat{G}_{KM}) = - \Delta \log \left[ \frac{f_Y(T)}{\hat{G}_{KM}(T)} \right] 
  \]
  (3)

This is a strictly proper approximate survival loss that evaluates a survival distribution prediction by taking the logarithm of the predicted probability density function. It is essentially the ‘usual’ negative log-likelihood with an IPCW weighting.

- Right-censored log-likelihood (RCLL) \[ 3 \]
  \[
  L_{RCLL}(Y, T, \Delta) = - \log \left[ \delta f_Y(T) + (1 - \Delta) S_Y(T) \right] 
  \]
  (4)

This is a strictly proper survival loss that evaluates a survival distribution prediction by taking the logarithm of the predicted probability density function for ‘dead’ observations and the logarithm of the predicted survival function for censored observations.

- Harrell’s C (\( C_H \)) [29, 30]
  \[
  C_H(\phi, T, \Delta) = \frac{\sum_{i \neq j} I(T_i < T_j, \phi_i > \phi_j) \Delta_i}{\sum_{i \neq j} I(T_i < T_j) \Delta_i} 
  \]
  (5)

where \( \phi \) are predicted risks. This is a concordance measure that evaluates the discrimination of a ranking prediction by asserting if predicted risks are concordant with observed survival times, i.e., if observation \( i \) is predicted to be of higher risk of dying than observation \( j \), then this prediction is concordant if \( i \) dies before \( j \).

- Uno’s C (\( C_U \)) [74]
  \[
  C_U(\phi, T, \Delta | \tau) = \frac{\sum_{i \neq j} W(T_i) \Delta_i I(T_i < T_j, \phi_i > \phi_j, T_i < \tau)}{\sum_{i \neq j} W(T_i) \Delta_i I(T_i < T_j, T_i < \tau)} 
  \]
  (6)

where \( \phi \) are predicted risks, \( W(t_i) = [\hat{G}_{KM}(t_i)]^{-2}, \hat{G}_{KM} \) is the Kaplan-Meier estimator fit on \((T, 1 - \Delta)\), and \( \tau \) is an upper-cutoff timepoint. This has the same interpretation as Harrell’s C but includes an IPCW adjustment to account for censoring.

- van Houwelingen’s alpha (CalA) [75]
  \[
  \alpha(\hat{H}, T, \Delta) = \frac{\sum_i \Delta_i}{\sum_i \hat{H}_i(T_i)} 
  \]
  (7)

where \( \hat{H}_i \) are individual predicted cumulative hazard functions. This is a calibration measure that evaluates if a predicted survival distribution is well-calibrated by asserting if the predicted expected number of events is equal (or close to equaling) the true number of observed events.

- D-calibration (CalD) – Algorithm in [27].
  This is a calibration measure that evaluates if a predicted survival distribution is well-calibrated by asserting if the predicted survival distributions are distributed Uniformly as expected.

Let \( L_M \) be a scoring rule evaluated on a model \( M \) and let \( L_K \) be the same scoring rule evaluated on a prediction from the Kaplan-Meier baseline, then the explained residual variation (ERV) \[ 26 \] of \( L \) is defined as the percentage decrease of \( L_M \) from \( L_K \):

\[
\tilde{L} = 1 - \frac{L_M}{L_K} = \frac{L_K - L_M}{L_K} 
\]
(8)

This allows any scoring rule to be meaningfully interpreted as a percentage increase in performance over a baseline model. We standardise all the scoring rules above (RSVS, RISL, SNL, RCLL) with this method.
C FULL RESULTS

The full results of running our experiment are provided below in tabular and graphical forms.

Table 2. Measures computed after running the permutation biasing algorithm a random survival forest, arithmetic mean taken over all datasets. \( \sigma \) values are represented in the columns such that \( \sigma = 0 \) means there is no bias in the disadvantaged dataset whereas \( \sigma = 0.9 \) means 90% of observations are biased in the disadvantaged dataset. Measures with 'ERV' after them are standardised against a Kaplan-Meier baseline (Appendix B).

| Measure / \( \sigma \) | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|                        |     |     |     |     |     |     |     |     |     |     |
| Permutation Biasing Method |     |     |     |     |     |     |     |     |     |     |
| \( C_H \)               | 0.034 | 0.036 | 0.046 | 0.054 | 0.073 | 0.089 | 0.102 | 0.118 | 0.131 | 0.138 |
| \( C_U \)               | 0.038 | 0.043 | 0.054 | 0.06 | 0.076 | 0.096 | 0.109 | 0.12 | 0.132 | 0.141 |
| CalA                   | 0.028 | 0.027 | 0.03 | 0.029 | 0.03 | 0.034 | 0.036 | 0.036 | 0.036 | 0.036 |
| CalD                   | 2.679 | 2.676 | 2.647 | 2.961 | 2.832 | 3.062 | 3.154 | 3.014 | 3.027 | 3.005 |
| RCLL                   | 0.019 | 0.02 | 0.017 | 0.022 | 0.022 | 0.025 | 0.024 | 0.025 | 0.025 | 0.027 |
| RISL                   | 0.049 | 0.049 | 0.054 | 0.064 | 0.065 | 0.085 | 0.085 | 0.091 | 0.098 | 0.096 |
| RSBS                   | 0.054 | 0.054 | 0.061 | 0.072 | 0.075 | 0.098 | 0.096 | 0.105 | 0.116 | 0.113 |
| SNL                    | 0.017 | 0.018 | 0.017 | 0.02 | 0.017 | 0.019 | 0.017 | 0.019 | 0.018 | 0.019 |

| Undersampling Biasing Method |     |     |     |     |     |     |     |     |     |     |
|-----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| \( C_H \)                  | 0.032 | 0.034 | 0.034 | 0.037 | 0.039 | 0.04 | 0.047 | 0.07 | 0.09 | 0.116 |
| \( C_U \)                  | 0.04 | 0.037 | 0.04 | 0.044 | 0.045 | 0.044 | 0.055 | 0.073 | 0.094 | 0.116 |
| CalA                       | 0.028 | 0.028 | 0.027 | 0.033 | 0.041 | 0.043 | 0.055 | 0.079 | 0.133 | 0.27 |
| CalD                       | 2.853 | 2.957 | 3.049 | 3.363 | 3.066 | 2.928 | 3.041 | 3.042 | 3.409 | 3.811 |
| RCLL                       | 0.026 | 0.03 | 0.039 | 0.045 | 0.055 | 0.063 | 0.075 | 0.086 | 0.093 | 0.101 |
| RISL                       | 0.042 | 0.049 | 0.046 | 0.043 | 0.051 | 0.048 | 0.055 | 0.07 | 0.085 | 0.104 |
| RSBS                       | 0.047 | 0.053 | 0.051 | 0.046 | 0.055 | 0.052 | 0.061 | 0.079 | 0.091 | 0.117 |
| SNL                        | 0.017 | 0.022 | 0.03 | 0.037 | 0.045 | 0.055 | 0.066 | 0.075 | 0.081 | 0.086 |
Fig. 1. Boxplots of $\sigma$ against $F_L$ over the 29 datasets with the permutation biasing method applied. Blue lines are fit with local polynomial regression.

Fig. 2. Spearman rank correlation of $\sigma$ against $F_L$ over the 29 datasets with the permutation biasing method applied. Pink bars indicate correlations with $p < 0.05$ after correction by Holm’s method.

Fig. 3. Boxplots of $\sigma$ against $F_L$ over the 29 datasets with the undersampling biasing method applied. Blue lines are fit with local polynomial regression.

Fig. 4. Spearman rank correlation of $\sigma$ against $F_L$ over the 29 datasets with the undersampling biasing method applied. Pink bars indicate correlations with $p < 0.05$ after correction by Holm’s method.
### D Datasets

Table 3. Datasets used in experiments. 1. Dataset ID and citation. 2. Proportion of censoring in the dataset, rounded to nearest percentage point. 3-4. Number of continuous and discrete features respectively. 5-6. Total number of observations and features respectively. 7. Number of observed events in dataset. 8. R package in which the dataset is included.

| Dataset                  | Cens % | $n_C$ | $n_D$ | $n$  | $p$ | $n_F$ | Package  |
|--------------------------|--------|-------|-------|------|-----|-------|----------|
| aids.id [15]             | 60     | 1     | 4     | 467  | 5   | 188   | JM [62]  |
| Aids2 [54]               | 38     | 1     | 3     | 2814 | 4   | 1733  | MASS [54]|
| ALL [76]                 | 63     | 0     | 4     | 2279 | 4   | 838   | dynpred  [59]|
| bladder0 [70]            | 48     | 0     | 3     | 397  | 3   | 206   | frailtyHL [59]|
| CarpenterFdaData [16]    | 36     | 15    | 11    | 408  | 26  | 262   | simPH    |
| channing [40]            | 62     | 1     | 1     | 458  | 2   | 176   | KMsurv   |
| child [12]               | 79     | 1     | 3     | 26574| 4   | 5616  | eha [12] |
| cost [34]                | 22     | 3     | 10    | 518  | 13  | 404   | pec [52] |
| e1684 [39]               | 31     | 1     | 2     | 284  | 3   | 196   | smcure  [14]|
| fchain [20]              | 72     | 4     | 3     | 7871 | 7   | 1082  | survival |
| FTR.data [73]            | 86     | 0     | 2     | 2206 | 2   | 300   | MRsurv  [23]|
| gbsg [37]                | 43     | 3     | 4     | 2232 | 7   | 1267  | pycox  [43]|
| grace [32]               | 68     | 4     | 2     | 1000 | 6   | 324   | mlr3proba [66]|
| hdfs [53]                | 94     | 1     | 4     | 52422| 5   | 2885  | frailtySurv [53]|
| kidtran [40]             | 84     | 1     | 3     | 863  | 4   | 140   | KMsurv   |
| liver [2]                | 40     | 1     | 1     | 488  | 2   | 292   | joineR  [81]|
| lung [47]                | 28     | 5     | 3     | 167  | 8   | 120   | survival |
| metabric [37]            | 42     | 5     | 4     | 1903 | 9   | 1103  | pycox    |
| mgs [44]                 | 6      | 6     | 1     | 176  | 7   | 165   | survival |
| naflid1 [1]              | 92     | 4     | 1     | 12588| 5   | 322   | survival |
| nwtc [11]                | 86     | 1     | 2     | 4028 | 3   | 571   | survival |
| ova [77]                 | 26     | 1     | 4     | 358  | 5   | 266   | dynpred  |
| patient [9]              | 79     | 2     | 5     | 1985 | 7   | 416   | pammttools [8]|
| rdata [48]               | 47     | 1     | 3     | 1040 | 4   | 547   | relsurv [48]|
| reconstitution [21]      | 19     | 0     | 2     | 200  | 2   | 162   | parfm    |
| std [40]                 | 60     | 3     | 18    | 877  | 21  | 347   | KMsurv   |
| STR.data [73]            | 82     | 0     | 4     | 546  | 4   | 101   | MRsurv  |
| support [37]             | 32     | 10    | 4     | 8873 | 14  | 2705  | pycox    |
| tumor [8]                | 52     | 1     | 6     | 776  | 7   | 375   | pammttools|
| uis [31]                 | 19     | 7     | 5     | 575  | 12  | 464   | quantreg [41]|
| veteran [35]             | 7      | 3     | 3     | 137  | 6   | 128   | survival |
| wbc1 [71]                | 43     | 2     | 0     | 190  | 4   | 109   | dynpred  |
| whas [32]                | 48     | 3     | 6     | 481  | 9   | 249   | mlr3proba |

Flexible Group Fairness Metrics for Survival Analysis

DSHealth 2022, August 14, 2022, Washington DC