arXiv:2206.03256v3 [cs.CY] 22 Jul 2022

Flexible Group Fairness Metrics for Survival Analysis

RAPHAEL SONABEND*, Technische Universität Kaiserslautern, Germany, University of Cambridge, UK, and Imperial College London, UK
FLORIAN PFISTERER*, 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.

CCS Concepts: • **Mathematics of computing** → **Survival analysis**; • **General and reference** → **Evaluation**; **Validation**; *Empirical studies*; *Metrics*; • **Computing methodologies** → Model verification and validation.

Additional Key Words and Phrases: fairness, bias, calibration, discrimination, scoring rules

ACM Reference Format:
Raphael Sonabend, Florian Pfisterer, Alan Mishler, Moritz Schauer, Lukas Burk, Sumantrak Mukherjee, and Sebastian Vollmer. 2022. Flexible Group Fairness Metrics for Survival Analysis. In . ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

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

*These authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DSHealth 2022, August 14, 2022, Washington DC
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

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 measures 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 $\Delta := \mathbb{I}(Y = T)$ as the survival indicator. Survival models are fit on the survival tuple (X, T, Δ). 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 $\mathrm{P_{x,y}}$. 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 $F_L = |L_A - L_D|$ 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, σ, 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 σ, and a significant t-test after regressing F_L on σ.

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 survival 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 σ. Regressing F_L on σ, we find that for the permutation biasing method, σ 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 σ however the regression slope is too small to be meaningful. There is also a significant relationship between F_L and σ for SNL after applying the undersampling algorithm (Table 1 and Appendix C).

3.2 Discussion

There is a significant relationship between σ 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 $F_L = \alpha + \sigma\beta$ be our regression model. Table shows intercept, α, slope, β, and Spearman rank correlation (ρ) 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	α	β	ρ	α	β	ρ
	Permutation			Undersampling		
RSBS	0.049	0.078*	0.976*	0.035	0.066*	0.855*
RISL	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*
C_H	0.024	0.129*	1.000*	0.017	0.083*	1.000*
C_U	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 RISL 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 requires 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 RISL, 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.

Disclaimer. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan 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, offer 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.

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 Krieger, Madina Agénor, 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 Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, 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.* fairmlbook.org. http://www.fairmlbook.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. pammtools: 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 t-year predicted risks. *Biostatistics* 20, 2 (apr 2019), 347–357. https://doi.org/10.1093/biostatistics/kxy006

[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. *eha: Event History Analysis.* http://ehar.se/r/eha/ R package version 2.9.0.

[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 Semi-parametric 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.).

[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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34, 2 (1972), 187–220.

[20] Angela Dispenzieri, 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

Melton 3rd, and S Vincent Rajkumar. 2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic proceedings* 87, 6 (jun 2012), 517–523. https://doi.org/10.1016/j.mayocp.2012.03.009

[21] Luc Duchateau and Paul Janssen. 2008. *The Frailty Model.* Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-72835-3

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[23] Y. Foucher and K. Trebern-Launay. 2013. *MRsurv: A multiplicative-regression model for relative survival.* https://CRAN.R-project.org/package=MRsurv R package version 0.2.

[24] Bruce Glymour and Jonathan Herington. 2019. Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 269–278. https://doi.org/10.1145/3287560.3287573

[25] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18, 17-18 (1999), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

[26] Erika Graf and Martin Schumacher. 1995. An Investigation on Measures of Explained Variation in Survival Analysis. *Journal of the Royal Statistical Society. Series D (The Statistician)* 44, 4 (jun 1995), 497–507. https://doi.org/10.2307/2348898

[27] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. 2020. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research* 21, 85 (2020), 1–63.

[28] William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American journal of public health* 105, 12 (2015), e60–e76.

[29] Frank E. Harrell, Robert M. Califf, and David B. Pryor. 1982. Evaluating the yield of medical tests. *JAMA* 247, 18 (may 1982), 2543–2546. http://dx.doi.org/10.1001/jama.1982.03320430047030

[30] F E Jr Harrell, K L Lee, R M Califf, D B Pryor, and R A Rosati. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine* 3, 2 (1984), 143–152. https://doi.org/10.1002/sim.4780030207

[31] David W Hosmer, Stanley Lemeshow, and Susanne May. 2008. *Applied survival analysis regression modeling of time-to-event data* (2nd ed. ed.). Wiley-Interscience, Hoboken, N.J.

[32] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. 2011. *Applied survival analysis: regression modeling of time-to-event data.* Vol. 618. John Wiley & Sons.

[33] By Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. 2008. Random survival forests. *The Annals of Statistics* 2, 3 (2008), 841–860. https://doi.org/10.1214/08-AOAS169 arXiv:arXiv:0811.1645v1

[34] Henrik Stig Jørgensen, Hirofumi Nakayama, Jakob Reith, Hans Otto Raaschou, and Tom Skyhøj Olsen. 1996. Acute Stroke With Atrial Fibrillation. *Stroke* 27, 10 (10 1996), 1765–1769. https://doi.org/10.1161/01.STR.27.10.1765

[35] John D Kalbfleisch and Ross L Prentice. 2011. *The statistical analysis of failure time data.* Vol. 360. John Wiley & Sons.

[36] E. L. Kaplan and Paul Meier. 1958. Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.* 53, 282 (1958), 457–481. https://doi.org/10.2307/2281868

[37] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology* 18, 1 (2018), 24. https://doi.org/10.1186/s12874-018-0482-1

[38] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James Foulds. 2021. Equitable Allocation of Healthcare Resources with Fair Survival Models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 190–198.

[39] John M Kirkwood, M Hunt Strawderman, Marc S Ernstoff, Thomas J Smith, Ernest C Borden, and Ronald H Blum. 1996. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. *Journal of clinical oncology* 14, 1 (1996), 7–17.

[40] John P Klein and Melvin L Moeschberger. 2003. *Survival analysis: techniques for censored and truncated data* (2 ed.). Springer Science & Business Media.

[41] Roger Koenker. 2021. *quantreg: Quantile Regression.* https://www.r-project.org R package version 5.86.

[42] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[43] Håvard Kvamme. 2018. pycox. https://pypi.org/project/pycox/

[44] R A Kyle. 1993. "Benign" monoclonal gammopathy–after 20 to 35 years of follow-up. *Mayo Clinic proceedings* 68, 1 (jan 1993), 26–36. https://doi.org/10.1016/s0025-6196(12)60015-9

[45] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 23 (2020), 12592–12594.

[46] Wenhua Liang, Jianhua Yao, Ailan Chen, Qingquan Lv, Mark Zanin, Jun Liu, SookSan Wong, Yimin Li, Jiatao Lu, Hengrui Liang, et al. 2020. Early triage of critically ill COVID-19 patients using deep learning. *Nature communications* 11, 1 (2020), 1–7.

[47] C L Loprinzi, J A Laurie, H S Wieand, J E Krook, P J Novotny, J W Kugler, J Bartel, M Law, M Bateman, and N E Klatt. 1994. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 12, 3 (mar 1994), 601–607. https://doi.org/10.1200/JCO.1994.12.3.601

[48] M. Pohar and J. Stare. 2006. Relative survival analysis in R. *Computer methods and programs in biomedicine* 81 (2006), 272–278. Issue 3. https://doi.org/10.1016/j.cmpb.2006.01.004

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635

[50] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[51] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[52] Ulla B Mogensen, Hemant Ishwaran, and Thomas A Gerds. 2014. Evaluating Random Forests for Survival Analysis using Prediction Error Curves.

[53] John V. Monaco, Malka Gorfine, and Li Hsu. 2018. General Semiparametric Shared Frailty Model: Estimation and Simulation with frailtySurv. *Journal of Statistical Software* 86, 4 (2018), 1–42. https://doi.org/10.18637/jss.v086.i04

[54] W N. Venables and B D. Ripley. 2002. *Modern Applied Statistics with S.* Springer. http://www.stats.ox.ac.uk/pub/MASS4

[55] Linda Nordling. 2019. A fairer way forward for AI in health care. *Nature* 573, 7775 (2019), S103–S103.

[56] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[57] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. 2021. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* 113 (Jan. 2021), 103621.

[58] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30. Curran Associates, Red Hook, NY, 5680–5689. arXiv:1709.02012 https://papers.nips.cc/paper/7151-on-fairness-and-calibration

[59] Hein Putter. 2015. dynpred: Companion Package to "Dynamic Prediction in Clinical Survival Analysis". https://cran.r-project.org/package=dynpred

[60] R Core Team. 2017. R: A Language and Environment for Statistical Computing.

[61] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169, 12 (2018), 866–872.

[62] Dimitris Rizopoulos. 2010. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software* 35, 9 (2010), 1—-33. http://www.jstatsoft.org/v35/i09/

[63] Logan Ryan, Carson Lam, Samson Mataraso, Angier Allen, Abigail Green-Saxena, Emily Pellegrini, Jana Hoffman, Christopher Barton, Andrea McCoy, and Ritankar Das. 2020. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: A retrospective study. *Annals of Medicine and Surgery* 59 (2020), 207–216.

[64] Raphael Sonabend, Andreas Bender, and Sebastian Vollmer. 2021. Avoiding C-hacking when evaluating survival distribution predictions with discrimination measures. (dec 2021). arXiv:2112.04828 http://arxiv.org/abs/2112.04828

[65] Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. 2021. mlr3proba: An R Package for Machine Learning in Survival Analysis. *Bioinformatics* (feb 2021). https://doi.org/10.1093/bioinformatics/btab039

[66] Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. 2021. mlr3proba: An R Package for Machine Learning in Survival Analysis. *Bioinformatics* (feb 2021). https://doi.org/10.1093/bioinformatics/btab039

[67] Raphael Edward Benjamin Sonabend. 2021. *A Theoretical and Methodological Framework for Machine Learning in Survival Analysis: Enabling Transparent and Accessible Predictive Modelling on Right-Censored Time-to-Event Data.* PhD. University College London (UCL). https://discovery.ucl.ac.uk/id/eprint/10129352/

[68] Charles L Sprung, Gavin M Joynt, Michael D Christian, Robert D Truog, Jordi Rello, and Joseph L Nates. 2020. Adult ICU triage during the coronavirus disease 2019 pandemic: who will live and who will die? Recommendations to improve survival. *Critical care medicine* 48, 8 (2020), 1196.
[69] Daniel Steinberg, Alistair Reid, and Simon O'Callaghan. 2020. Fairness measures for regression via probabilistic classification. *arXiv preprint arXiv:2001.06089* (2020).
[70] Richard J. Sylvester, Adrian P.M. Meijden, J. Alfred Witjes, Christian Bouffioux, Louis Denis, Donald W.W. Newling, and Karlheinz Kurth. 2006. Predicting Recurrence and Progression in Individual Patients with Stage Ta T1 Bladder Cancer Using EORTC Risk Tables: A Combined Analysis of 2596 Patients from Seven EORTC Trials. *European Urology* 49, 3 (2006), 466–477. https://doi.org/10.1016/j.eururo.2005.12.031
[71] The Benelux C M L Study Group. 1998. Randomized Study on Hydroxyurea Alone Versus Hydroxyurea Combined With Low-Dose Interferon-α2b for Chronic Myeloid Leukemia. *Blood* 91, 8 (apr 1998), 2713–2721. https://doi.org/10.1182/blood.V91.8.2713.2713_2713_2721
[72] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.
[73] Katy Trébern-Launay, Magali Giral, Jacques Dantal, and Yohann Foucher. 2013. Comparison of the Risk Factors Effects between Two Populations: Two Alternative Approaches Illustrated by the Analysis of First and Second Kidney Transplant Recipients. *BMC Medical Research Methodology* 13 (Aug. 2013), 102. https://doi.org/10.1186/1471-2288-13-102
[74] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L J Wei. 2011. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in Medicine* 30, 10 (2011), 1105–1117. https://doi.org/10.1002/sim.4154 arXiv:NIHMS150003
[75] Hans C. Van Houwelingen. 2000. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 19, 24 (2000), 3401–3415. https://doi.org/10.1002/1097-0258(20001230)19:24<3401::AID-SIM554>3.0.CO;2-2
[76] Hans C van Houwelingen and Hein Putter. 2008. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime data analysis* 14, 4 (dec 2008), 447–463. https://doi.org/10.1007/s10985-008-9099-8
[77] J C Van Houwelingen, W W ten Bokkel Huinink, M E Van der Burg, A T Van Oosterom, and J P Neijt. 1989. Predictability of the survival of patients with advanced ovarian cancer. *Journal of Clinical Oncology* 7, 6 (1989), 769–773.
[78] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.
[79] Sebastian Vollmer, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine SL McAllister, Puja Myles, et al. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj* 368 (2020).
[80] Ping Wang, Yan Li, and Chandan K. Reddy. 2019. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.* 51, 6, Article 110 (feb 2019), 36 pages. https://doi.org/10.1145/3214306
[81] Paula Williamson, Ruwanthi Kolamunnage-Dona, Pete Philipson, and Anthony G. Marson. 2008. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* 27 (2008), 6426–6438.
[82] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making: the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, Vol. 32. 2037–2045. https://ojs.aaai.org/index.php/AAAI/article/view/11564
[83] Wenbin Zhang and Jeremy C. Weiss. 2022. Longitudinal Fairness with Censorship. https://doi.org/10.48550/ARXIV.2203.16024

A BIASING ALGORITHM

We have a generic biasing algorithm (Algorithm 1) that is then specialised with two biasing methods. 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 σ% 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 dataset, 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 of 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: $D = (X, T, \Delta)$, survival dataset. σ, proportion of disadvantaged observations to bias. L, survival loss. M, survival model. G, biasing algorithm. Output: F_L, fairness metric.

for $i = 1, ..., 10$ **do**
  $D_B, D_U \leftarrow$ Randomly split D into equal-sized datasets
  $D_{BA}, D_{BD} \leftarrow$ Randomly split D_B w.p. σ
  Apply G to D_{BD}
  $D_B \leftarrow (D_{BA}, D_{BD})$
  $L_B \leftarrow$ 3-fold CV of D_B with model M and loss L
  $L_U \leftarrow$ 3-fold CV of D_U with model M and loss L
  $F_{L;i} \leftarrow |L_B - L_U|$
end for
$F_L \leftarrow \sum_i F_{L;i}/10$
return F_L

B METRIC DEFINITIONS

Let X, Y, C be random variables taking values in $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}, \mathcal{C} \subseteq \mathbb{R}$ respectively. In addition, define $T := \min(Y, C)$, and $\Delta := \mathbb{I}(Y = T)$. Finally let $(X_i, Y_i, C_i, T_i, \Delta_i) \overset{\text{i.i.d.}}{\sim} (X, Y, C, T, \Delta)$. We define the following measures and include a brief explanation of their interpretation and use.

- Reweighted survival Brier score (RSBS) [25, 67]

$$L_{RSBS}(Y, T, \Delta|\hat{G}_{KM}) = \frac{\Delta \int_{\mathcal{T}} (\mathbb{I}(T \leq \tau) - F_Y(\tau))^2 \, d\tau}{\hat{G}_{KM}(T)} \quad (1)$$

  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) [25, 67]

$$L_{RISL}(Y, T, \Delta|\hat{G}_{KM}) = \frac{\Delta \int_{\mathcal{T}} \mathbb{I}(T \leq \tau) \log[F_Y(\tau)] + \mathbb{I}(T > \tau) \log[S_Y(\tau)] \, d\tau}{\hat{G}_{KM}(T)} \quad (2)$$

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 log-likelihood with an inverse probability of censoring weighting (IPCW) adjustment

- Survival negative log-likelihood (SNL) [67]

$$L_{SNL}(Y, T, \Delta|\hat{G}_{KM}) = -\frac{\Delta \log[f_Y(T)]}{\hat{G}_{KM}(T)} \tag{3}$$

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]

$$L_{RCLL}(Y, T, \Delta) = -\log[\delta f_Y(T) + (1-\Delta)(S_Y(T))] \tag{4}$$

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]

$$C_H(\phi, T, \Delta) = \frac{\sum_{i \neq j} \mathbb{I}(T_i < T_j, \phi_i > \phi_j)\Delta_i}{\sum_{i \neq j} \mathbb{I}(T_i < T_j)\Delta_i} \tag{5}$$

where ϕ 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]

$$C_U(\phi, T, \Delta|\tau) = \frac{\sum_{i \neq j} W(T_i)\mathbb{I}(T_i < T_j, \phi_i > \phi_j, T_i < \tau)\Delta_i}{\sum_{i \neq j} W(T_i)\mathbb{I}(T_i < T_j, T_i < \tau)\Delta_i} \tag{6}$$

where ϕ 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 τ 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]

$$\alpha(\hat{H}, T, \Delta) = \frac{\sum_i \Delta_i}{\sum_i \hat{H}_i(T_i)} \tag{7}$$

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

- D-calibration (CalD) – Algorithm in [27].
This is a calibration measures 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:

$$\tilde{L} = 1 - \frac{L_M}{L_K} = \frac{L_K - L_M}{L_K} \tag{8}$$

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. σ 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 / σ	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

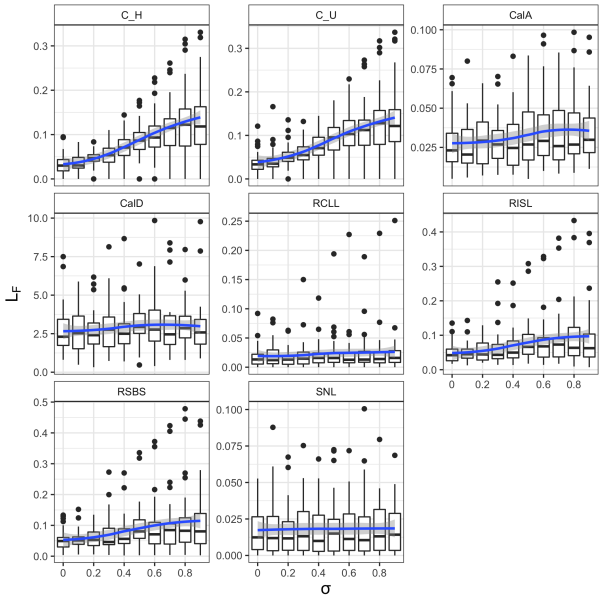

Fig. 1. Boxplots of σ against F_L over the 29 datasets with the permutation biasing method applied. Blue lines are fit with local polynomial regression.

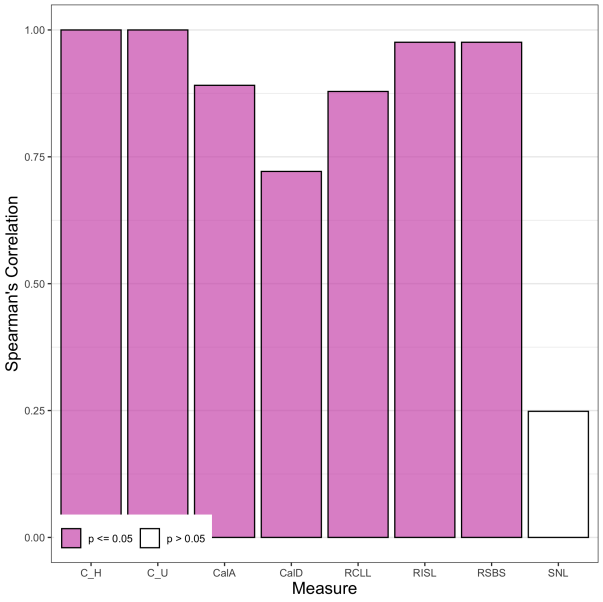

Fig. 2. Spearman rank correlation of σ against F_L over the 29 datasets with the permutation biasing method applied. Pink bars indicate correlations with $p < 0.05$ after correction by Holm's method.

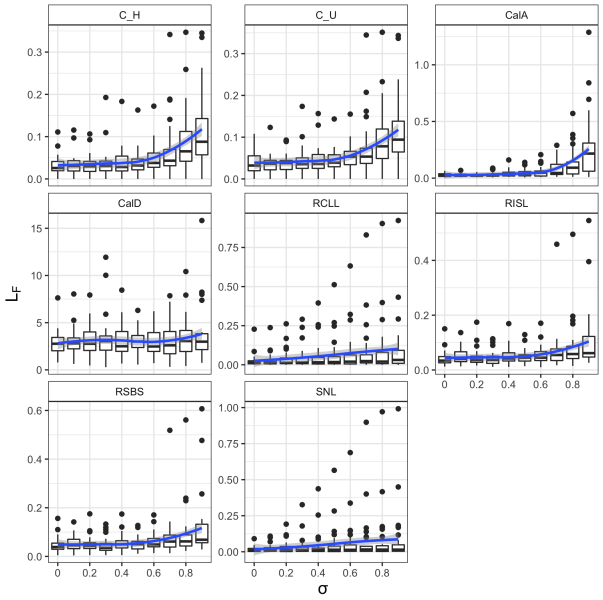

Fig. 3. Boxplots of σ against F_L over the 29 datasets with the undersampling biasing method applied. Blue lines are fit with local polynomial regression.

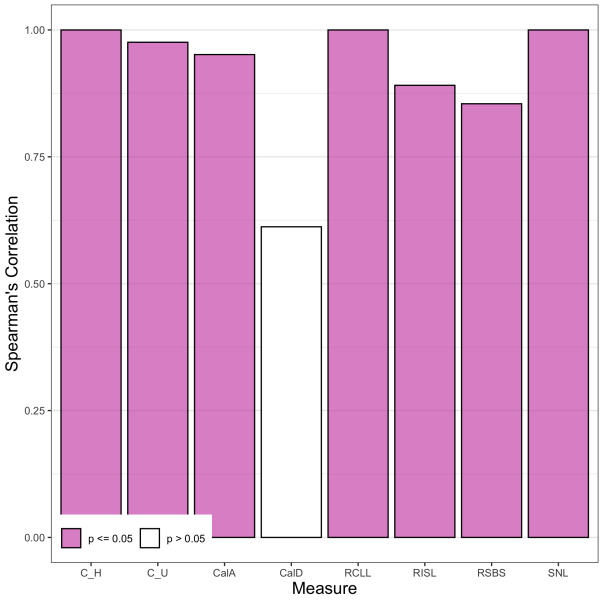

Fig. 4. Spearman rank correlation of σ against F_L over the 29 datasets with the undersampling biasing method applied. Pink bars indicate correlations with $p < 0.05$ after correction by Holm's method.

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[1]	**Cens %**[2]	n_C^3	n_D^4	**n**[5]	**p**[6]	n_E^7	**Package**[8]
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]
flchain [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]
hdfail [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**
mgus [44]	6	6	1	176	7	165	**survival**
nafld1 [1]	92	4	1	12588	5	322	**survival**
nwtco [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	**pammtools** [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	**pammtools**
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**