How reliable are treatment effects in clinical trials with dropout?

Yuxin Wang, Dennis Frauen, Jonas Schweisthal, Maresa Schröder, Stefan Feuerriegel
LMU Munich
Munich Center for Machine Learning (MCML)
`{Yuxin.Wang1, frauen, jonas.schweisthal, maresa.schroeder, feuerriegel}@lmu.de`

Abstract

Dropout is widespread in clinical trials and real-world oncology studies, with up to half of patients leaving before the study ends due to side effects or other reasons. When such dropout is informative (i.e., dependent on survival time), it induces censoring bias that distorts causal survival analysis and leads to biased treatment effect estimates. This challenge is particularly acute when estimating conditional average treatment effects (CATEs), which are central to personalized medicine because they reveal which patients benefit most from treatment. In this paper, we propose an assumption-lean method to assess the robustness of CATE estimates in survival analysis when facing censoring bias. Specifically, we frame the underlying task through the lens of partial identification, which allows us to obtain informative bounds on the CATE under such conditions. Importantly, this approach helps identify patient subgroups where treatment is still effective despite potential censoring. We then show that our bounds converge to the true point estimates of the CATE when the censoring bias goes to zero. We further propose a novel model-agnostic meta-learner to estimate the bounds that can be used combined with arbitrary machine-learning models and that has favorable theoretical properties such as double-robustness and quasi-oracle efficiency. We finally demonstrate the effectiveness of our meta-learner across various experiments using both simulated and real-world data.

The full version is available at: `https://arxiv.org/abs/2510.13397`

1 Introduction

Dropout is common in survival studies, particularly in oncology [Shand et al., 2024]. Patients may leave a study because of severe side effects, personal circumstances, or physician decisions about continued participation [Fizazi et al., 2017]. Such incomplete follow-up induces censoring, partially masking event times and, if unaccounted for, biasing treatment-effect estimates (we refer to this as "censoring bias" in the following), potentially making therapies appear more or less effective than they truly are for the broader patient population.

This challenge is especially acute when estimating conditional average treatment effects (CATEs), for personalized medicine, as it helps identify which patients benefit from treatment and can thereby guide personalized decision-making [Dahabreh et al., 2019, Feuerriegel et al., 2024, Wang et al., 2024]. Unlike the average treatment effect (ATE), the CATE captures the variability, which accounts for that some patients may experience substantial benefits (e.g., delayed disease progression), while others may see little or even reduced survival due to side effects. In oncology, outcomes are often measured as time-to-event variables (e.g., survival time, progression-free survival) [Falet et al., 2022,

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

Seitz et al., 2023, Buell et al., 2024]. This is referred to as survival data[1], requires tailored methods for CATE estimation from survival data [Van Der Laan and Robins, 2003, Curth et al., 2021, Xu et al., 2024, Frauen et al., 2025].

Methods have been proposed to deal with censoring bias in ATE estimation for survival data [Bai and Cui, 2025, Voinot et al., 2025], but are typically not directly applicable to CATE. Existing approaches for dealing with censoring bias in CATE estimates for survival data typically assume non-informative censoring (i.e., censoring times are fully independent or conditionally independent of survival time) [Rubin and van der Laan, 2007, Mao et al., 2018, Cai and van der Laan, 2019, Cheng et al., 2022, Schrod et al., 2022, Westling et al., 2024]. These include methods such as specific model-based estimation, such as Cox models [Gao and Hastie, 2022], tree-based [Zhang et al., 2017, Henderson et al., 2020, Tabib and Larocque, 2020, Cui et al., 2023], or neural-network-based methods [Schrod et al., 2022, Katzman et al., 2018, Curth and van der Schaar, 2021]. When the non-informative censoring assumption fails, estimates of CATE are biased. Even under it, they still have to estimate the full distribution of observational time via hazard functions, which significantly increases the complexity of the methods.

In this paper, we make three **contributions:** (for details, see the full version of this paper, including theoretical results and experiments at `https://arxiv.org/abs/2510.13397`): (1) We propose an assumption-lean framework to audit censoring bias in the CATE estimates from a censored dataset. Our method replaces the non-informative censoring assumption with sensitivity functions that use censoring strength and domain knowledge (e.g., expected survival after dropout) to form informative bounds. (2) We further introduce a model-agnostic meta-learner called **SurvB-learner** to efficiently estimate bounds. (3) We provide theoretical results for our meta-learners by showing consistency, double robustness, and quasi-oracle efficiency properties. Finally, we confirm the effectiveness of our meta-learners by performing various experiments using both synthetic and real-world data. We

2 Problem setup

Data: We consider the standard setting for estimating CATEs based on time-to-event data [Van Der Laan and Robins, 2003, Curth et al., 2021, Frauen et al., 2025, Zhang et al., 2017, Cui et al., 2023]. That is, we consider the full population $(X, A, T, C) \sim \mathbb{P}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^p$ are observed covariates, $A \in \mathcal{A} \subseteq \mathbb{N}$ is the discrete treatments, $T \in \mathcal{T} = \{0, 1, \ldots, t_{\max}\}$ is the event time of interest (e.g., the time of overall survival (OS), time of the patient or disease-free survival (DFS)), and $t_{\max}$ represents, in the general medical sense, the theoretical maximum human lifespan. $C \in \mathcal{T}$ is the censoring time (e.g., the time of a patient dropping out of the study). Because of censoring, we only observe a dataset $\mathcal{D} = \left\{\left(x_i, a_i, \tilde{t}_i, \delta_i\right)_{i=1}^n\right\}$ of size $n \in \mathbb{N}$ sampled i.i.d. from the population $Z = (X, A, \tilde{T}, \Delta)$, where $\Delta = \mathbb{1}(C \leq T)$ is a censoring indicator for the event and censoring times and $\tilde{T} = \min\{T, C\}$. The causal graph is shown in Fig. 1.

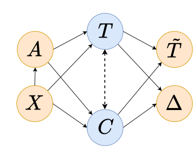

Figure 1: **Causal graph**. Variables in yellow are observed, while in blue are unobserved.

Causal estimand: We make use of the potential outcome framework [Rubin, 1974] to formalize our causal inference task. Let $T(a) \in \mathcal{T}$ denote the potential event time corresponding to a treatment intervention $A = a$. We are interested in the CATE on the survival time $\tau_{a_1,a_2}(x) = \mathbb{E}\left[T(a_1) - T(a_2) \mid X = x\right]$ with corresponding we define the conditional average potential outcomes (CAPO) of survival time via $\tau_a(x) = \mathbb{E}\left[T(a) \mid X = x\right]$.

We make standard assumptions (consistency, treatment, overlap, unconfoundedness, censoring overlap) to ensure identifiability. These assumptions (i)–(iii) are standard in causal inference for estimating CATEs [Rubin, 1974, Imbens, 2004, Shalit et al., 2017, Candès et al., 2023]. Censoring overlap is also common in survival analysis [Cai and van der Laan, 2019, Westling et al., 2024], ensuring every covariate has a positive chance of being uncensored. However, identifying CATE further requires the non-informative censoring assumption [Van Der Laan and Robins, 2003, Curth et al., 2021, Frauen et al., 2025], which assumes survival and censoring times are conditionally

[1]We deal with the problem of right censoring, which is very common in survival analysis settings. We thus assume that events have not happened before time $t = 0$.

independent given covariates and treatment. This is often violated in practice, leaving $\mathbb{E}[T \mid X = x, A = a, \Delta = 1]$ non-identified. We therefore focus on partial identification of CATE.

3 Our approach to partial identification of CATE in the presence of censoring

We now move away from point estimation to partial identification of the CATE, which allows us to obtain informative bounds in the presence of informative censoring. We define the lower and upper bounds for the CAPO, denoted by $\mu^-(x, a)$ and $\mu^+(x, a)$ respectively, which capture the range of plausible values under our partial identification framework that allows for censoring.

Lower bound: To construct a lower bound for the CAPO, we leverage the definition of $\tilde{T} = \min\{C, T\}$. We then replace T with $\tilde{T}$ to account for that our analysis is conditioned on the censored events (i.e., $\Delta = 1$), so that $T \geq \tilde{T}$ and $\nu(1, x, a) \geq \mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a]$. Therefore, we have

$$\begin{aligned}\mu(x, a) \geq \mu^-(x, a) \\ &\geq \nu(0, x, a)[1 - \xi(x, a)] + \nu(1, x, a)\xi(x, a) = \mathbb{E}[\tilde{T} \mid X = x, A = a].\end{aligned} \tag{1}$$

Upper bound: To construct an upper bound for the CAPO, we introduce the post-dropout survival time function $\gamma(x, a)$ as a sensitivity function which is naturally defined: it captures the maximum possible average survival time a patient may live after censoring. Based on $\gamma(x, a)$, the range of sensitivity function is given by

$$\mathbb{E}[T - \tilde{T} \mid \Delta = 1, X = x, A = a] \leq \gamma(x, a) \leq t_{\max} - \mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a], \tag{2}$$

for all $x \in X$ and $a \in A$. Then we can directly use it to construct informative upper bounds for the CAPO. By definition of $\gamma(x, a)$, the resulting **domain knowledge upper bound** $\mu^+(x, a)$ takes the form (we discuss a special case of the sensitivity function in the main paper as **non-informative upper bound**).

$$\mu^+(x, a) = \nu(0, x, a)[1 - \xi(x, a)] + \nu(1, x, a)\xi(x, a) + \gamma(x, a)\xi(x, a). \tag{3}$$

where the bound is expressed as a weighted combination of the uncensored survival function $\nu(0, x, a)$, the observed censored follow-up $\tilde{T}$, and the post-dropout survival captured by $\gamma(x, a)$.

Next, we present our main result: the partial identification bounds, $\tau^-_{a_1,a_2}(x)$ and $\tau^+_{a_1,a_2}(x)$, which characterize the range of the CATE in the presence of censoring bias.

Theorem 3.1: *Under the above assumptions, the CATE is bounded via $\tau^-_{a_1,a_2}(x) \leq \tau_{a_1,a_2}(x) \leq \tau^+_{a_1,a_2}(x)$, where $\tau^+_{a_1,a_2}(x) = \mu^+(x, a_1) - \mu^-(x, a_2)$ and $\tau^-_{a_1,a_2}(x) = \mu^-(x, a_1) - \mu^+(x, a_2)$. Here, $\mu^+(x, a_1)$ and $\mu^+(x, a_2)$ are given by Eq. (1), and $\mu^+(x, a_1)$ and $\mu^+(x, a_2)$ are given by Eq. (3).*

Proof: *See our main paper*

We state the width property of our bounds in our main paper. Our partial identification bounds are especially effective under low censoring, where they remain tight enough to approximate point estimates without modeling the full hazard function, enabling reliable identification of treatment-benefiting subgroups.

4 SurvB-learner: A meta-learner for estimating the bounds

We now develop our two-stage meta-learner for estimating the bounds in Theorem 3.1. For simplicity, we derive the two-stage meta-learner for the CAPOs, while the corresponding bounds for the CATE can be obtained directly by taking the difference between the two CAPOs. Importantly, our two-stage meta-learner is flexible and can be instantiated with arbitrary machine learning methods.

Formally, we first estimate the nuisance functions with any suitable machine learning models. We rely on standard nuisance functions: the propensity score $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$, censoring strength $\xi(x, a) = \mathbb{P}(\Delta = 1 \mid X = x, A = a)$, expected survival time function $\mu(x, a) = \mathbb{E}[T \mid X = x, A = a]$ and conditional survival time function $\nu(\delta, x, a) = \mathbb{E}[T \mid \Delta = \delta, X = x, A = a]$, and the post-dropout survival function $\gamma(x, a)$ which denotes the expected maximum survival time after dropout for patients with covariates x under treatment a. Second, we combine them with observed data to construct a debiased estimator. This design ensures consistency, double-robustness, and quasi-oracle efficiency.

Theorem 4.1: *Our SurvB-learner is consistent, doubly-robust, and quasi-oracle efficient.*

Proof: *For the proof and the detailed theorem, see the full version of our paper.*

Using the pseudo-outcomes derived above, our SurvB-learner first estimates the nuisance functions and then computes the pseudo-outcomes. In future work, we plan to extend our methods to continuous treatment settings and observational data (see our full version).

5 Experiments

We now evaluate the effectiveness of the proposed bounds and SurvB-learner by performing experiments on synthetic and open-access public datasets. Synthetic data are commonly used to evaluate causal inference methods [Van Der Laan and Robins, 2003, Curth et al., 2021, Frauen et al., 2025] as they have the advantage that we have access to the ground-truth CATEs and thereby can make comparisons against oracle estimates. Further, medical data allows us to demonstrate both the applicability and relevance of our method in practice.

Data. Following Frauen et al. [2025], we simulate datasets from different functions under varying censoring strengths ($\xi = 0.2, 0.4, 0.6$). Since the ground-truth data-generating process is known, we compare SurvB-learner against the oracle CATE and oracle bounds derived from ground-truth nuisance estimators. Both domain-knowledge and non-informative bounds are evaluated against the plug-in learner.

Bound Type	Dataset	Exponential function		
	Censoring strength ξ	0.2	0.4	0.6
Domain knowledge	Plug-in learner	3.219 ± 3.528	4.063 ± 3.214	4.529 ± 2.576
	SurvB learner	$\mathbf{0.143 \pm 0.003}$	$\mathbf{0.147 \pm 0.006}$	$\mathbf{0.152 \pm 0.008}$
Non-informative	Plug-in learner	5.455 ± 6.573	6.359 ± 5.801	6.620 ± 4.581
	SurvB-learner	$\mathbf{0.138 \pm 0.003}$	$\mathbf{0.137 \pm 0.004}$	$\mathbf{0.135 \pm 0.006}$

* Smaller is better. Best value in bold.

Table 1: Mean and standard deviation of the RMSE over 5 random runs for synthetic datasets.

Results. Table 1 reports RMSEs relative to oracle bounds. SurvB-learner achieves the lowest average error and variability, with RMSE up to 7.4 fold smaller than the plug-in learner, consistent with Künzel et al. [2019], Nie and Wager [2020].

In the full version of our paper, we further show that SurvB-learner reliably recovers both domain-knowledge and non-informative bounds, and the width of non-informative bounds shrinks as censoring decreases. And we will demonstrate our framework using the **ADJUVANT trial** [Zhong et al., 2018, Liu et al., 2021] of adjuvant gefitinib in resected EGFR-mutant NSCLC in the full version of our paper.

Acknowledgement

Our research was supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Jenny Shand, Elizabeth Stovold, Lucy Goulding, and Kate Cheema. Cancer care treatment attrition in adults: Measurement approaches and inequities in patient dropout rates: a rapid review. *BMC Cancer*, 24(1):1345, 2024.

Karim Fizazi, NamPhuong Tran, Luis Fein, Nobuaki Matsubara, Alfredo Rodriguez-Antolin, Boris Y. Alekseev, Mustafa Özgüroğlu, Dingwei Ye, Susan Feyerabend, Andrew Protheroe, Peter De Porre, Thian Kheoh, Youn C. Park, Mary B. Todd, Kim N. Chi, and LATITUDE Investigators. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. *The New England Journal of Medicine*, 377(4):352–360, 2017.

Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Guanbo Wang, Patrick J. Heagerty, and Issa J. Dahabreh. Using effect scores to characterize heterogeneity of treatment effects. *JAMA*, 331(14):1225–1226, 2024.

Jean-Pierre R. Falet, Joshua Durso-Finley, Brennan Nichyporuk, Julien Schroeter, Francesca Bovis, Maria-Pia Sormani, Doina Precup, Tal Arbel, and Douglas Lorne Arnold. Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nature Communications*, 13(1):5645, 2022.

Kevin P. Seitz, Alexandra B. Spicer, Jonathan D. Casey, Kevin G. Buell, Edward T. Qian, Emma J. Graham Linck, Brian E. Driver, Wesley H. Self, Adit A. Ginde, Stacy A. Trent, Sheetal Gandotra, Lane M. Smith, David B. Page, Derek J. Vonderhaar, Jason R. West, Aaron M. Joffe, Kevin C. Doerschug, Christopher G. Hughes, Micah R. Whitson, Matthew E. Prekker, Todd W. Rice, Pratik Sinha, Matthew W. Semler, and Matthew M. Churpek. Individualized treatment effects of bougie versus stylet for tracheal intubation in critical illness. *American Journal of Respiratory and Critical Care Medicine*, 207(12):1602–1611, 2023.

Kevin G. Buell, Alexandra B. Spicer, Jonathan D. Casey, Kevin P. Seitz, Edward T. Qian, Emma J. Graham Linck, Wesley H. Self, Todd W. Rice, Pratik Sinha, Paul J. Young, Matthew W. Semler, and Matthew M. Churpek. Individualized treatment effects of oxygen targets in mechanically ventilated critically ill adults. *JAMA*, 331(14):1195–1204, 2024.

Mark J. Van Der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York, NY, 2003. ISBN 978-0-387-21700-0.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. In *NeurIPS*, 2021.

Shenbo Xu, Raluca Cobzaru, Stan N. Finkelstein, Roy E. Welsch, Kenney Ng, and Zach Shahn. Estimating heterogeneous treatment effects on survival outcomes using counterfactual censoring unbiased transformations. *arXiv preprint*, arXiv:2401.11263, 2024.

Dennis Frauen, Maresa Schröder, Konstantin Hess, and Stefan Feuerriegel. Orthogonal survival learners for estimating heterogeneous treatment effects from time-to-event data. In *NeurIPS*, 2025.

Yang Bai and Yifan Cui. Partial causal identification for right censored data with noncompliance. *Journal of Nonparametric Statistics*, 2025.

Charlotte Voinot, Clément Berenfeld, Imke Mayer, Bernard Sebastien, and Julie Josse. Causal survival analysis, estimation of the average treatment effect (ATE): Practical recommendations. 2025.

Daniel Rubin and Mark J. van der Laan. A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3(1):Article 4, 2007.

Huzhang Mao, Liang Li, Wei Yang, and Yu Shen. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine*, 37(26):3745–3763, 2018.

Weixin Cai and Mark J. van der Laan. One-step targeted maximum likelihood for time-to-event outcomes. *arXiv preprint*, arXiv:1802.09479, 2019.

Chao Cheng, Fan Li, Laine E Thomas, and Fan (Frank) Li. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American Journal of Epidemiology*, 191(6):1140–1151, 2022.

Stefan Schrod, Andreas Schäfer, Stefan Solbrig, Robert Lohmayer, Wolfram Gronwald, Peter J. Oefner, Tim Beißbarth, Rainer Spang, Helena U. Zacharias, and Michael Altenbuchinger. BITES: balanced individual treatment effect for survival data. *Bioinformatics*, 38(Supplement_1):i60–i67, 2022.

Ted Westling, Alex Luedtke, Peter B. Gilbert, and Marco Carone. Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, 119(546): 1541–1553, 2024.

Zijun Gao and Trevor Hastie. Estimating heterogeneous treatment effects for general responses. *arXiv preprint*, arXiv:2103.04277, 2022.

Weijia Zhang, Thuc Duy Le, Lin Liu, Zhi-Hua Zhou, and Jiuyong Li. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378, 2017.

Nicholas C. Henderson, Thomas A. Louis, Gary L. Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68, 2020.

Sami Tabib and Denis Larocque. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics (Oxford, England)*, 36(2):629–636, 2020.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

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.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *arXiv preprint*, arXiv:2101.10943, 2021.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 2004.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint*, arXiv:1712.04912, 2020.

Wen-Zhao Zhong, Qun Wang, Wei-Min Mao, Song-Tao Xu, Lin Wu, Yi Shen, Yong-Yu Liu, Chun Chen, Ying Cheng, Lin Xu, Jun Wang, Ke Fei, Xiao-Fei Li, Jian Li, Cheng Huang, Zhi-Dong Liu, Shun Xu, Ke-Neng Chen, Shi-Dong Xu, Lun-Xu Liu, Ping Yu, Bu-Hai Wang, Hai-Tao Ma, Hong-Hong Yan, Xue-Ning Yang, Qing Zhou, Yi-Long Wu, Qun Wang, Wei-Min Mao, Lin Wu, Yi Shen, Yong-Yu Liu, Chun Chen, Ying Cheng, Lin Xu, Jun Wang, Ke Fei, Xiao-Fei Li, Jian Li, Cheng Huang, Zhi-Dong Liu, Shun Xu, Ke-Neng Chen, Shi-Dong Xu, Lun-Xu Liu, Ping Yu, Bu-Hai Wang, Hai-Tao Ma, Si-Yu Wang, Jian Hu, Wei Liu, Wei Li, and Jian-Hua Shi. Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II–IIIA (N1–N2) EGFR-mutant NSCLC (ADJUVANT/ctong1104): a randomised, open-label, phase 3 study. *The Lancet Oncology*, 19(1):139–148, 2018.

Si-Yang Liu, Hua Bao, Qun Wang, Wei-Min Mao, Yedan Chen, Xiaoling Tong, Song-Tao Xu, Lin Wu, Yu-Cheng Wei, Yong-Yu Liu, Chun Chen, Ying Cheng, Rong Yin, Fan Yang, Sheng-Xiang Ren, Xiao-Fei Li, Jian Li, Cheng Huang, Zhi-Dong Liu, Shun Xu, Ke-Neng Chen, Shi-Dong Xu, Lun-Xu Liu, Ping Yu, Bu-Hai Wang, Hai-Tao Ma, Hong-Hong Yan, Song Dong, Xu-Chao Zhang, Jian Su, Jin-Ji Yang, Xue-Ning Yang, Qing Zhou, Xue Wu, Yang Shao, Wen-Zhao Zhong, and Yi-Long Wu. Genomic signatures define three subtypes of EGFR-mutant stage II–III non-small-cell lung cancer with distinct adjuvant therapy outcomes. *Nature Communications*, 12(1):6450, 2021.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we state clearly in the last paragraph of introduction of contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We state clearly in the last sentence of SurvB-learner.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have clear assumption and proof in main paper and appendix.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the code of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have the data generation file in our code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes it is in the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we use RMSE and report the mean and standard deviation over five runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we record them in our code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: I reviewed it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We state it in our introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We don't make use of the large language or generatino model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: I correctly cited the origin paper of adjuvant dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper uses publicly available datasets and does not involve new experiments with human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study uses publicly available datasets and does not involve new experiments with human participants; hence no IRB approval is required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: I use LLM to help me correct the grammar of writing and some typos.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.