Response to reviewers

Anders Munch and Thomas A. Gerds

October 30, 2025

We thank the the reviewers for their useful comments. We have incorporated their suggestions into the manuscript, and we find it much improved. Please see our point-to-point responses below.

Quotes from the old version of the manuscript is inserted in red boxes, quotes from the updated manuscript is inserted in gray boxes, and when relevant newly added text is show in blue.

Reviewer 1

Comments to the Author

This paper proposes a super learner-style estimator of conditional cumulative hazard functions in the context of right-censored time-to-event data with competing risks. The user specifies libraries of candidate estimators of the conditional cumulative hazards of the event, censoring, and competing risk times. The proposed method then selects the element of the product of these libraries that has the smallest cross-validated empirical risk, where the risk is based on a carefully constructed loss function. A key benefit of the proposed methods over alternatives is that it jointly estimates the three cumulative hazards, rather than iterating between estimation of them or requiring a pre-specified censoring survival estimator. The authors provided an oracle inequality demonstrating that the true risk of the estimator is no worse than a constant multiple of the oracle risk up to a logarithmic penalty in the size of the library. The authors also provide numerical studies and a simulation illustrating the good performance of the proposed method.

Overall, this is a nice contribution to the literature, and I am grateful to the authors for the methods and accompanying theory. The paper is well written and careful, though I think more effort could be made in parts to make the notation somewhat more accessible. In particular, an outline of the proposed algorithm in its totality would be useful for readers.

Thank you for this comment, we agree that it is a good idea with an outline of the proposed algorithm. We have added a pseudo-algorithm

to the paper.

1. I have two main comments. My first comment regards a sentence in the discussion: “A potential drawback of our approach is that we are evaluating the loss of the learners on the level of the observed data distribution, while the target of the analysis is either the event-time distribution, or the censoring distribution, or both.” I didn’t quite catch this on a first read-through of the paper, but I think it is an important point that deserves to be emphasized more throughout the paper. Specifically, this makes me wonder about the relevance of the oracle inequalities in Proposition 2 and Corollary 3: do we actually care about these risks? I usually care about inequalities and rates of convergence of the conditional survival (or cumulative hazard) function of the event time, censoring time, and competing risk. If I understand the above comment correctly, the risk considered in the paper doesn’t directly give such inequalities and rates. Is that correct? If so, it deserves to be mentioned and at least briefly discussed in the section on theoretical results. Similarly, this makes me question the relevance of the IPA metric considered in the numerical study. How do the different algorithms compare in terms of pointwise or integrated error for the conditional survival or cumulative hazard functions?

   This is an important point, which we agree should be made clear in the paper. We have added the following paragraph discussing this point at the end of Section 5.

   > *Quote from revised manuscript*
   >
   > The norm defined in equation (**??**) operates on functions F which are features of the observed data distribution. This means that Proposition **??** and Corollary **??** provide guarantees in terms of a how well the function $\hat{\varphi}_n$ predicts the observed data. Ideally, we would like performance guarantees for, e.g., the selected learners $\hat{\Lambda}_{jn}$ or the derived risk-prediction learner $\hat{Q}_n$ defined in equation (**??**). There is a one-to-one correspondence between the learner $\hat{\varphi}_n$ and the tuple of learners $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n)$ through equations (**??**)-(**??**) and (**??**), and we expect that the performance guarantees provided for $\hat{\varphi}_n$ will in many cases translate into similar performance guarantees for each element of the tuple $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n)$. We do not investigate this further theoretically, but investigate is empirically in our numerical experiments in Section **??**.

   We have also added the following reflection to the discussion in Section 8 (new text in blue.).

> *Quote from revised manuscript*
>
> A potential drawback of our approach is that we are evaluating the loss of the learners on the level of the observed data distribution, while the target of the analysis is either the event-time distribution, or the censoring distribution, or both. Our numerical experiments suggest that our approach does provide estimates of the conditional survival functions which perform well, also when the predictive performance is measured against the true survival function with no censoring present.

In the numerical studies, we in fact do use the IPA for the risk prediction model (1 − the conditional survival function) of interest. We have now attempted to make this clearer by updating the following paragraph in Section 6:

> *Quote from old version of manuscript*
>
> Each super learner provides a learner for the cumulative hazard function for the outcome of interest. From the cumulative hazard function, we obtain a risk prediction model as described in Section 4, with the special case of $\Lambda_2 = 0$. We measure the performance of each super learner by calculating the index of prediction accuracy (IPA) [Kattan and Gerds, 2018] at a fixed time horizon (36 months) for the risk prediction model provided by the super learner. The IPA is 1 minus the ratio between the model's Brier score and the null model's Brier score, where the null model is the model that does not use any covariate information. The value of IPA is approximated using a large ($n = 20,000$) independent data set of uncensored data.

> *Quote from revised manuscript*
>
> Each super learner provides a learner for the cumulative hazard function for the outcome of interest. From the cumulative hazard function, we obtain a risk prediction model as described in Section **??**, see in particular equation (**??**) with the special case of $\Lambda_2 = 0$. We measure the performance of the risk prediction model provided by each super learner by calculating the index of prediction accuracy (IPA) (Kattan and Gerds, 2018) at a fixed time horizon (36 months) for the risk prediction model provided by the super learner. For a risk prediction model $r \colon \mathcal{X} \to [0,1]$, IPA at time τ is
>
> $$1 - \frac{\mathbb{E}_Q[(r(X) - \mathbb{1}\{T \leq \tau\})^2]}{\mathbb{E}_Q[(Q(T \leq \tau) - \mathbb{1}\{T \leq \tau\})^2]}.$$
>
> We chose IPA as a performance measure because it is proper, incorporates both discrimination and calibration, and is easy to interpret as it measures the relative performance gain compared to the null model which does not use any baseline information. The definition of IPA involves the uncensored survival time T, which is not available in practice. However, in the numerical studies, this quantity is available because we know the data-generating mechanism used to generate T. In practice, we Monte Carlo approximate the IPA by generating a large ($n = 20,000$) independent data set of uncensored survival times, and calculate the empirical version of the IPA in there.

2. Second, and more minor, I wonder if the authors could expand on why it is difficult to extend the method to an ensemble estimator in this setting.

Thank you for this comment. We do not believe that it needs to be particularly difficult, but there are at least two strategies that could be pursued, and we think that some additional thought on this is needed. We have expanded on this in the discussion by updating the following paragraph:

> *Quote from old version of manuscript*
>
> We have focused on a discrete version of the joint survival super learner, but it is of interest to extend the method to a proper ensemble learner, where learners are combined, e.g., through stacking. How an ensemble should be build for tuples of learners is an interesting topic for future research.

> *Quote from revised manuscript*
>
> We have focused on a discrete version of the joint survival super learner, but it is of interest to extend the method to a proper ensemble learner, where learners are combined, e.g., through stacking. There are at least two possible directions for constructing an ensemble version of the joints survival super leaner. One option is to construct a single convex combination of the F-learners $\varphi \in \Phi(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$. Another, perhaps more interesting option, is to construct three separate convex combinations of the learners in $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$. How such an ensemble should be build and implemented is an interesting topic for future research.

Reviewer 2

Section 3:

- P5. The point that the partial log-likelihood does not work well as an evaluation criterion is interesting and warrants further elaboration. An additional sentence explaining this point would be helpful. Additionally, I wonder if this problem can be circumvented in other ways: perhaps models that normally yield piecewise constant hazards could be included in a slightly modified form with smoothed hazards. What would be the consequences of this approach?

  We agree that this is an important point, and we have elaborated further. It is an interesting idea to attempt to smooth the problematic piecewise constant cumulative hazard functions. We believe that this introduces other issues that would have to be addressed, and we briefly reflect on that now. Please see the updated paragraph below with new text in blue.

> *Quote from revised manuscript*
>
> However, the partial log-likelihood loss does not work well as a general purpose measure of performance in hold-out samples when data are observed in continuous time. The reason is that the partial log-likelihood assigns an infinite value to any learner that predicts piecewise constant cumulative hazard functions, if the test set contains event times that are not observed in the training set. For instance, if no competing risks are present, a piecewise constant cumulative hazard function postulates a model for the distribution of the survival times where all probability is assigned to the finite number of time points at which the cumulative hazard function jumps. The likelihood according to such a model is zero at almost all time points, and thus the likelihood of any hold-out sample will almost surely be zero when data are observed in continuous time. This problem occurs with prominent survival learners including the Kaplan-Meier estimator, random survival forests, and semi-parametric Cox regression models, and these learners cannot be included in the library of the super learner proposed by Polley and van der Laan (2011). One might attempt to resolve this issue by smoothing an estimated cumulative hazard functions to obtain an estimate of the hazard function itself. This is a theoretically unattractive approach, as estimation of a hazard function is much harder than estimation of a cumulative hazard function. In practice, this approach would also introduces the additional problem of tuning a smoothing parameter, which may be infeasible for more complicated estimators like random survival forest, where the smoothing would have to be done conditional on baseline covariates.

Section 4:

- P5. I find the notation using “1” for the indicator function difficult to read, especially when preceded by other numerals. Why not use I, , or a bold or blackboard/double-struck 1?

  Thank you for this comment. The blackboard/double-struck was lost when converting to the journal’s template, and we have now corrected that.

- P6. The manuscript explains that it is in principle fairly easy to use the survival package to estimate libraries of models for A, A, and B. The practical value of the article would improve with a web appendix showing code implementing this for the prostate cancer study data.

  A code supplement is provided at the Github repository that is referenced at the end of the Introduction. We have now added an example that demonstrates how tools from the survival package can be used to construct learners. We cannot share the original

data, but we share an emulated data set and demonstrate how the joint survival super learner can be fitted to the data at the referenced Github repository.

Section 5:

- P8. You state that a sensible objective function must be proper. While I don't disagree, this point often generates discussion. One might argue that in certain problems, our objective need not be the correct recovery of the probability distribution. Rather, we may want to predict outcomes at specific time points while incorporating relative costs of incorrect predictions. It's not clear that using probabilities as an intermediate step is always optimal, especially when these are estimated with error. Furthermore, while a proper loss function might be optimal with large sample sizes, this may not hold for finite samples. More discussion of this point would be valuable.

  Answer notes: We now discuss utility / cost benefit. unclear what is meant by finite sample problem of proper loss but penalized likelihood is a good example.

- P8. The authors quickly move from using proper scoring rules to the Brier score, which feels less general. Are the results expected to hold for other proper scoring rules?

  Answer notes: yes, what other proper scoring rules are there? is there an integrated deviance (log - score) maybe in Putter's book?

- P9. At the top of the page, it's argued that the oracle inequality provides insights into how the number of folds, time horizon, and number of learners influence performance. Could the practical utility of this be illustrated in the context of the prostate cancer example?

  Answer notes: no, but we have now added a comment about this.. Unclear how to address this, because I do not believe there is any actual practical utility to be gained from the types of inequalities. Hmm, maybe we could say something about that in this case, our analyses indicates that fairly complex models are used. If this is true, our finite sample inequality shows that, for large enough samples, the error rate will be dominated by that flexible models and not by the cross-validation estimation step – refer to that we seem to be in case (b) of Corollary 3.

- P10. The performance will be specified using the IPA, which needs more discussion. Please briefly state how it is defined and why it was chosen. (Is the IPA itself proper? If not, does this still make sense here? Do we need other metrics as well?)

  We have now defined the IPA explicitly, please see our response to Reviewer 1's first. The IPA is just a scaled version of the Brier score so it is indeed proper.

- P10. The SurvSL model should be described in more detail.

  Answer notes: Do this in the simulation setup?

- P10. For the 'second aim,' it's suggested that an advantage of the joint survival learner may be that it is a discrete super learner. This warrants more explanation—why is this an advantage? Is there more risk of overfitting in small samples?

  Draft answer: This is an interesting point. To our knowledge, the finite sample performance of a discrete versus continuous (ensemble) super learner is not well understood. In the special case where the data-generating model is included as a learner in the library, we think that a discrete super learner could have an advantage, because it is forced to pick one model from the library, and will with good change pick the correct one, while an ensemble learner will always be a mixture of the correct model and a mis-specified one. We have expanded on this and tried to clarify that this only a possible explanation and not a strict fact.

  Todo: add something to the manuscript.

Section 7:

- P11. The prostate cancer study seems like an afterthought. It would be beneficial to reference it more when introducing the problem to demonstrate practical relevance.

  Todo TAG.

- P11. The splitting into training and test data seems wasteful. Could nested cross-validation be used instead?

  Todo: Try it out and then decide.

- P11. More discussion of the results is needed. What exactly do we learn from this practical use case?

  Answer notes: Todo: ask chatgpt: prompt could you help us answer the reviewers request: here is the paragraph of our paper copy-paste

- P11. The usefulness of this manuscript for practical researchers would be greatly enhanced with a link to a code supplement.

  A code supplement is provided at the Github repository that is referenced at the end of the Introduction, please also see our answer to you comment to Section 4, page 6.

Section 8:

- P12. The stated drawback—that the authors evaluate the loss of learners at the level of the observed data distribution while the target is either the event-time distribution, censoring distribution, or both—needs more detailed explanation. Are there alternatives to this approach, and what would be their drawbacks?

We have addressed this partly in our answer to Reviewer 1's first comment, please see above. We briefly mention alternative approaches in Section 3, and we now reiterate and expand on these points in the Discussion, please see the paragraph below which we have added to the Discussion.

> *Quote from revised manuscript*
>
> Alternatives to using a performance measure defined with respect the observed data are the use of IPCW loss functions, censoring unbiased transformations, or pseudo-values. As mentioned in Section 3, the drawback of these approaches is that they all need a prespecified estimator of the censoring distribution, and hence these methods are not immediately applicable if we do not in advance know how to model the censoring distribution. We note that for the special case where the partial log-likelihood loss can be used, this loss function, like our suggested approach, also measures performance with respect to a feature defined by the observed data distribution (e.g., Hjort, 1992; Whitney et al., 2019). [e.g., Hjort, 1992, Whitney et al., 2019]. We do not know of any method that would allow us to evaluate performance of a risk-prediction model in censored data without either modeling additional nuisance parameters (such as the censoring distribution) or measuring performance directly with respect to the observed data.

- P12. The authors claim that targeted learning is also known as debiased machine learning. However, these are often presented as distinct approaches. This should be clarified.

This is a fair point. We have updated the text accordingly:

> *Quote from old version of manuscript*
>
> A relevant application of the joint survival super learner is within the framework of targeted learning [van der Laan and Rose, 2011], also known as debiased machine learning [Chernozhukov et al., 2018], – a general methodology that combines flexible, data-adaptive estimation of nuisance parameters with asymptotically valid inference for low-dimensional target parameters.

> *Quote from revised manuscript*
>
> A relevant application of the joint survival super learner is within the framework of targeted learning [van der Laan and Rose, 2011] or debiased machine learning [Chernozhukov et al., 2018], which are general methodologies for combining flexible, data-adaptive estimation of nuisance parameters with asymptotically valid inference for low-dimensional target parameters.

- P12. The discussion lacks reflection on the results from sections 6 and 7.

  TODO.

References

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

N. L. Hjort. On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, pages 355–387, 1992.

M. W. Kattan and T. A. Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2018.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.