Suggestions on Using Machine Learning Models and Cautions for Analyzing Censored Time-To-Event Outcomes

Yushu Shi, PhD[1]; Miki Horiguchi, PhD[2]; and Ying Lu, PhD[3]

DOI https://doi.org/10.1200/PO.24.00220

Accepted April 12, 2024
Published May 23, 2024

JCO Precis Oncol 8:e2400220

The advancement of machine learning techniques and statistical methods have piqued the interest of many clinical investigators. As a journal that aims to advance the science and practice of precision oncology and define genomics- and other biomarker-driven clinical care of patients with cancer, *JCO PO* is highly receptive to cutting-edge data analysis methods. Nonetheless, we have noted some incorrect applications of machine learning models in previous paper review processes. In addition, inappropriate survival data analyses, particularly in defining of study groups and initial time of the analysis, have been frequently observed by reviewers. In this short commentary, we would like to shed light on common errors we have encountered and provide advice to potential authors.

Understand the Difference Between Machine Learning Models and Conventional Statistical Inference

Both conventional statistical models and machine learning models have their own strengths. Conventional statistical models are universally recognized, easier to comprehend, and provide statistical inferences such as CIs and *P* values. Conversely, machine learning models are considerably more flexible and often deliver more accurate predictions. The most important disadvantage of machine learning models is the lack of interpretability. There are tools improving the interpretability of machine learning models, including Local Interpretable Model-Agnostic Explanations,[1] Shapley values,[2] and Integrated Gradients.[3] Nevertheless, many times, machine learning models operate like a black box.

Tip: Many machine learning models, such as random forest,[4] boosting,[5] BART,[6] and deep neural network–based models, can handle feature interactions to some extent automatically. As a result, it is usually unnecessary to manually code interaction terms into these models.

Use the Most Appropriate Method, Not the Fanciest or the Most Accessible

Machine learning is one of the fastest-evolving fields in modern science and technology. We expect clinical researchers to use cutting-edge methods for data analysis, but the solidity of the biomedical science should take precedence over the novelty of the model used. Also, some R or Python packages offer easy-to-implement functions or machine learning methods. We encourage authors to use those tools, but not at the expense of losing important information.

Example of improper model application: Authors find scikit-learn has an easy-to-access function for categorical outcomes, so they dichotomize survival outcomes at a landmark time point.

The above practice can have two potential drawbacks:

1. It discards the rich information on individual survival time contained in the survival data.
2. For censored individuals, exact event times are unknown. Simple dichotomizing may introduce bias.

When dichotomizing survival outcomes at a landmark time point *t* (eg, 5-year overall survival), survival status at time *t* is unknown for those with censored observations before time *t*. Therefore, it is important to clarify how those patients were handled in the analysis. Removing patients with censored observations from the analysis or assuming they were alive at *t* may lead to biased results. We recommend using appropriate methods to handle the censored

observations. For example, for evaluating a model to predict survival status at 5 years from diagnosis, there are methods that can accommodate censored observations such as an inverse probability of censoring weighting (IPCW)-based method[7] or a time-dependent receiver operating characteristic (ROC) method.[8] These methods can be implemented using the PHREG procedure (option = IPCW) in SAS and the survivalROC package in R.

Speaking of machine learning models, one of the most trending topics in the field is deep learning, or artificial deep neural networks (DNNs). DNNs use multiple layers to progressively extract high-level features from raw input. However, for tabular data, DNNs might not be the most suitable choice, and most biomedical data sets are tabular, with each column representing a specific attribute, such as blood pressure or age. Compared with other data types such as raw text or images, tabular data are typically clearer but more costly to gather. They also often have smaller sample sizes, and certain data augmentation methods, such as upsampling, are not applicable.

Previously, Shwartz-Ziv and Armon[9] performed a comparative study of various machine learning techniques, including DNNs, in their paper, "Tabular data: Deep learning is not all you need." The study concluded that XGBoost[10] consistently outperformed deep models across different data sets, even those used in papers proposing the deep models. Apart from performance issues, Shwartz-Ziv and Armon[9] noted that deep learning may require more fine-tuning, as factors such as the number of layers, neurons in each layer, and the choice of activation function will significantly affect the results. Here, we want to point out that although machine learning techniques generally prioritize prediction over inference, some methods can provide interpretable feature importance measures. For instance, random forest provides a feature importance index that is straightforward to interpret. These merits should be considered when deciding which model to use.

It is also important to note that complex machine learning models, which have more parameters, often demand moderate to large sample sizes. If your labeled data set is not sufficient to train a deep neural network from scratch, transfer learning could be a viable alternative, provided you have access to a network pretrained on a large data set for a similar task. This pretrained neural network should have the same form of input. From another standpoint, we kindly ask authors to share their trained neural networks as part of their paper submission, which will serve as a springboard for future researchers interested in transfer learning.

Machine Learning Model Does Not Guarantee Superior Performance

Example of inferior machine learning model accuracy: Authors used and reported both the Cox model and the DeepSurv model for a survival data set, and Cox model has a c-index of 0.75, higher than that of the DeepSurv (0.52).

Often, when people opt for machine learning models over conventional statistical models, they exchange interpretability for prediction accuracy. In the above example, the DeepSurv[11] model performance index is worse than that of conventional method, which suggests a failure in model fitting. We provide a table of measurements of model fit as well as conventional model to compare for different types of outcomes (Table 1). Regardless of the method used (whether a machine learning method or a conventional statistical method), point estimates of model performance measures should be reported with corresponding (bootstrap) CIs.

Complex models are susceptible to overfitting, where the trained model perfectly fits the provided data set but fails to generalize to new data. Overfitting can be mitigated by using cross-validation during model training. We recommend having a separate test data set to avoid overestimating the performance of the trained models.

Define the Comparison Groups and Time Zero in Survival Analysis

When comparing survival probabilities between responder and nonresponder groups, there are two possible analysis populations—(1) all patients and (2) only patients who have survived beyond a landmark time point, *t*. Regarding the analysis population (1), a patient's response status is unknown at the start of treatment (ie, time zero of the survival curves). Therefore, a comparison of survival probabilities between two groups is subject to a so-called survival bias or guaranteed time bias in favor of the responder group and thus is potentially invalid. Although this issue has been discussed in several tutorial papers a long time ago,[15-17] we still see some manuscripts that do not appropriately handle this issue.

TABLE 1. Conventional Models and Measures of Model Fit, Organized by Outcome Type

Outcome Type	Conventional Model	Measurement Index
Continuous	Linear regression	Mean squared error, absolute prediction error
Binary	Logistic regression	c-index (ROC-AUC), accuracy, F1 score
Survival	Cox regression	c-index[12,13] Integrated Brier score[14]

Abbreviation: ROC-AUC, area under the receiver operating characteristic curve.

Example: Authors compared the survival probabilities between a group of patients who had adverse events and a group of patients who did not. However, the starting time point of the Kaplan-Meier curves in the analysis was the time of treatment initiation, at which no patient's adverse event status can be determined.

In the example above, everyone starts as the negative group (the nonresponder or nonadverse event group). We will only know the group label later in the study. Patients in the adverse event group must survive long enough for adverse events to be evaluated. However, patients with poor survival prognoses who die early (eg, death due to disease) do not have a chance to enter the adverse event group, which may contribute to poor survival for the nonadverse event group.

One of the alternatives would be a landmark analysis using the analysis population (2). Although the analysis population is restricted to a subset of the entire study population, a comparison of survival time between the two groups (ie, responders *v* nonresponders) will not be subject to survival bias. The responder and nonresponder groups will be defined on the basis of the response status information available at a landmark time point, *t*. Another alternative would be a method that considers response status (or adverse event status) as a time-varying covariate.[18]

Concluding Remarks and Guidance for Future Authors

The machine learning discussion of this brief commentary paper mainly focuses on supervised learning, where data are clearly labeled. Important problems related to unsupervised learning and semisupervised learning are untouched in this short commentary. Also, several important aspects of machine learning are not discussed, including the unintentional bias introduced in artificial intelligence[19] and patient privacy protection.[20] To summarize, we suggest the following when using and reporting machine learning models:

1. Specify the model optimization goal.
2. Apply conventional statistical methods on your data to ascertain whether the machine learning model truly offers a predictive advantage.
3. Elucidate your model structure. For instance, for DNN models, indicate the number of layers in the model and the number of neurons in each layer (can be in supplemental materials).
4. Disclose training techniques used, such as dropout, batch normalization, L1 or L2 normalization (can be in supplemental materials).

From our review experience, we propose the following key considerations for handling censored time-to-event data:

1. Opt for survival models that leverage the comprehensive data available, rather than dichotomizing the outcomes. If a survival outcome that is dichotomized at a landmark time point (eg, 5-year overall survival) is of primary interest, use an appropriate method that handles patients whose survival status at the landmark time point is unknown because of censoring.
2. Ensure fairness in comparison when the groups are defined later in the study. To appropriately compare survival between groups, make sure that the groups to be compared are defined using the information available at the starting time point of the survival curves (ie, time zero).

This paper outlines some common problems we observed during the review process, and we hope it can serve as a guide for future authors. To summarize, we offer these general recommendations for data analysis in research:

1. Engage with statisticians or data scientists regularly.
2. For clinical studies, prioritize biomedical significance over the novelty of data analysis methods.
3. Clearly document model inputs and sample size, possibly through a table that summarizes predicting covariates, specifying which are categorical and which are continuous.
4. Select models that effectively use all pertinent data, avoiding the temptation to chase after new models.
5. Precisely define your outcome of interest, noting whether it is continuous, categorical, or time-to-event.
6. Allocate a portion of your data as test data for validation purposes to avoid overestimating the performance of the derived model.
7. Share your model publicly on platforms such as GitHub. Provide comprehensive details about input and outcome variables, the model itself, and a thorough model description.

AFFILIATIONS

[1]Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY
[2]Dana-Farber Cancer Institute, Department of Data Science, Boston, MA
[3]Stanford Cancer Institute and Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA

CORRESPONDING AUTHOR

Ying Lu, PhD; e-mail: ylu1@stanford.edu.

EQUAL CONTRIBUTION

Y.S. and M.H. contributed equally to this work.

SUPPORT

Supported by the Sandra and Edward Meyer Cancer Center at Weill Cornell Medicine, Cornell University (Y.S.) and NCI 3P30CA124435 (Y.L.).

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Yushu Shi

This author is a member of the *JCO Precision Oncology* Editorial Board. Journal policy recused the author from having any role in the peer review of this manuscript.

Ying Lu

This author is an Associate Editor for *JCO Precision Oncology*. Journal policy recused the author from having any role in the peer review of this manuscript.

Consulting or Advisory Role: Nektar, Abeona Therapeutics (Inst), Gilead Sciences, Roche/Genentech, Emergent BioSolutions, Bavarian Nordic, WCG Clinical Inc

Research Funding: UCB Biopharm Inc

No other potential conflicts of interest were reported.

REFERENCES

1. Ribeiro M, Singh S, Guestrin C: "Why should I trust you?": Explaining the predictions of any classifier, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego, CA, Association for Computational Linguistics, 2016, pp 97-101
2. Štrumbelj E, Kononenko I: Explaining prediction models and individual predictions with feature contributions. Knowledge Inf Syst 41:647-665, 2014
3. Sundararajan M, Taly A, Yan Q: Axiomatic attribution for deep networks, in Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research. PMLR, 2017, pp 3319-3328
4. Ho TK: Random decision forests, in Proceedings of 3rd international conference on document analysis and recognition, volume 1. IEEE, 1995, pp 278-282
5. Friedman JH: Greedy function approximation: A gradient boosting machine. Ann Stat 29:1189-1232, 2001
6. Chipman HA, George EI, McCulloch RE: BART: Bayesian additive regression trees. Ann Appl Stat 4:266-298, 2010
7. Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc 102:527-537, 2007
8. Heagerty PJ, Lumley T, Pepe MS: Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337-344, 2000
9. Shwartz-Ziv R, Armon A: Tabular data: Deep learning is not all you need. Inf Fusion 81:84-90, 2022
10. Chen T, Guestrin C: XGBoost: A scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. New York, NY, ACM, 2016, pp 785-794
11. Katzman JL, Shaham U, Cloninger A, et al: DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24, 2018
12. Harrell FE Jr, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. JAMA 247:2543-2546, 1982
13. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30:1105-1117, 2011
14. Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18:2529-2545, 1999
15. Anderson JR, Cain KC, Gelber RD: Analysis of survival by tumor response. J Clin Oncol 1:710-719, 1983
16. Anderson JR, Cain KC, Gelber RD, et al: Analysis and interpretation of the comparison of survival by treatment outcome variables in cancer clinical trials. Cancer Treat Rep 69:1139-1146, 1985
17. Anderson JR, Cain KC, Gelber RD: Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. J Clin Oncol 26:3913-3915, 2008
18. Simon R, Makuch RW: A non-parametric graphical representation of the relationship between survival and the occurrence of an event: Application to responder versus non-responder bias. Stat Med 3:35-44, 1984
19. Cho MK: Rising to the challenge of bias in health care AI. Nat Med 27:2079-2081, 2021
20. Murdoch B: Privacy and artificial intelligence: Challenges for protecting health information in a new era. BMC Med Ethics 22:122, 2021