Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis

Xiajing (Jean) Gong[1,2], Meng Hu[1], Liang Zhao[1]

[1] Division of Quantitative Methods and Modeling, Office of Research and Standards, Office of Generic Drugs, CDER, FDA

[2] ORISE Fellow, Oak Ridge Institute for Science and Education

CONTACT INFORMATION:
Liang.Zhao@fda.hhs.gov; Xiajing.Gong@fda.hhs.gov

INTRODUCTION

- Time-to-event analysis, also referred to as survival analysis, is performed to analyze the expected time to event occurrence. This technique was originally developed for clinical studies, and now has been applied to many other areas.
- In clinical studies, the Cox proportional hazards regression model, a *de facto* standard for the survival analysis, is essentially semiparametric with underlying assumptions including proportional hazards, linearity and additivity, which may be oversimplified in practice. Sub-standard performance of the Cox model dealing with high dimensional data also limits its utilization.
- In the past decade, the development of machine learning (ML) methods has impacted a broad spectrum of research areas including survival analysis.
- Despite these applications of ML algorithms, the ML-based survival analysis has not been well recognized in the community of pharmacometrics or quantitative clinical pharmacology.
- There is currently no systematic evaluation for ML algorithms with regard to their performance advantages over the conventionally used regression based methods (e.g., Cox model).

OBJECTIVES

- We performed extensive simulations to evaluate the utilization and performance of ML-based approaches for survival analysis, as a big data pharmacometrics tool alternative to the conventional Cox regression model.

METHODS

➢ Simulation of time-to-event data

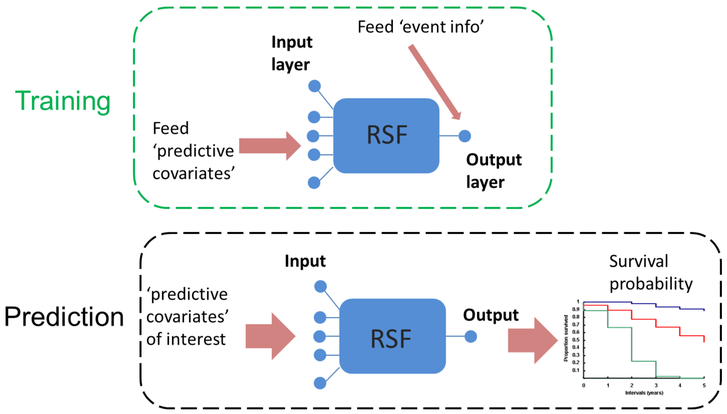
- Survival data were simulated using preset Cox models [1], yet with specific changes.
- Weibull distribution was used for survival time generation.
- By changing the relations of predictor variables in the hazard function, various complex scenarios were created, i.e., linear/independent predictors, nonlinear and/or dependent predictors, and data with a large number of predictors.
- We simulated the survival data via two approaches: **1)** by six hypothetical mathematical models, **2)** by real-world clinically relevant models.

➢ ML-based survival analysis

- We adopted the well-developed artificial neural network (ANN) [2] and random survival forest (RSF) [3] as proxies for the ML-based methods.

➢ Both simulations and analysis were performed in R.

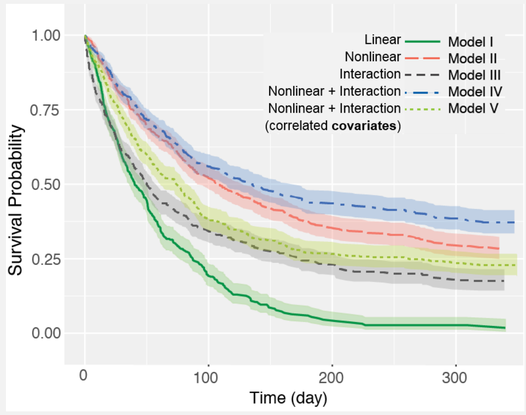
Machine Learning (ML) Methods for Time-to-Event Analysis



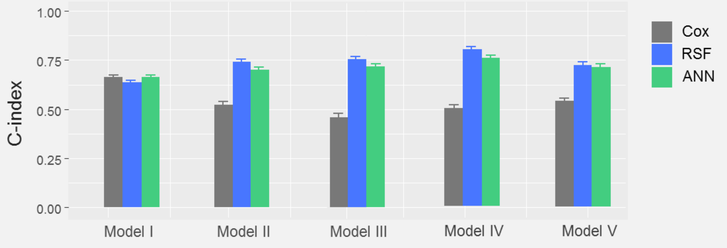
RESULTS

➢ Mathematical models

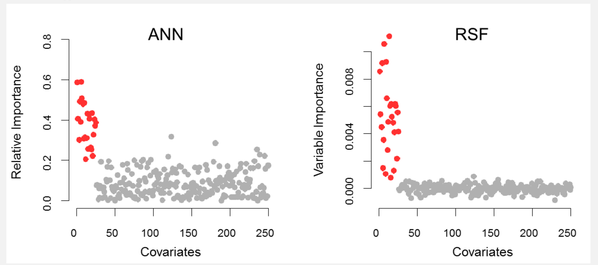
Survival data (Kaplan-Meier curves shown below) generated from Models I-V to represent increasingly complex relationships among predictor covariates. The covariates being drawn from the same distribution.



Cox model, ANN and RSF were applied on simulation data to perform survival analysis. Prediction performance are evaluated in terms of C-index (concordance index). Results are based on 500 repeated simulations for each model.



Model VI: High-dimensional simulation data



The first preset 25 important covariates (indicated in red) were successfully identified with relatively larger importance values than the non-significant covariates by both RSF and ANN.

➢ Clinically relevant models

E-R relationship for an anticancer drug

Response: survival probability

Exposure: C_{trough}

Confounding covariates: tumor size, ECOG

*ECOG: a performance score to measure a patient's daily living abilities)

Model	Description	Relationship for covariates in hazard function
A	Interaction between ECOG and C_{trough}	$\beta_1 \times ECOG + \beta_2 \times Tumor\ size + \beta_3 \times C_{trough} + \beta_{13} \times ECOG \times C_{trough}$
B	Nonlinear drug (E_{max}-type) exposure effects	$\beta_1 \times ECOG + \beta_2 \times Tumor\ size + \beta_{13} \times \frac{60 \times C_{trough}}{30 + C_{trough}}$

Data generated from non-linear, non-additive clinical relevant models. ML-based methods, both ANN and RSF can offer C-index of ~0.7 (model A) and ~0.6 (model B) vs. 0.5 by the Cox model.

CONCLUSION(S)

ML-based methods provide a powerful phamracometrics tool for time-to-event analysis, with built-in capacity for high dimensional data and better performance when the predictor variables assume nonlinear relationships in the hazard function.

REFERENCES

1. Bender, R., et. al. . Generating survival times to simulate Cox proportional hazards models. Stat. Med. 24(11), 1713-1723 (2005).
2. Biganzoli, E., et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Med. 17(10), 1169-1186 (1998).
3. Ishwaran, H., et al. Random survival forests. Appl. Stat. 841-860 (2008).

ACKNOWLEDGEMENTS

This work was supported in part by FDA Critical Path Initiative. Dr. Gong was supported by an appointment to the Research Participation Program at the Center for Drug Evaluation and Research, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. FDA. This poster presentation reflects the views of the authors and should not be construed to represent FDA's views or policies.