Abstract No 147213

Randon Survival Forest for Survival Extrapolation: feasibility and performance vs parametric modeling

JEWITI-RIGONDZA K[1], NEFF-BARO S[1], GAUTHIER A[2]

1:Amaris Consulting, Paris, France, 2:Amaris Consulting, Barcelona, Spain

INTRODUCTION

- Decision makers rely on economic evaluations, including cost-effectiveness assessments, to inform their choices of healthcare interventions.[1] These analyses rely on accurate measurements of benefits and costs of new treatments, including extrapolation of efficacy outcomes beyond clinical trial follow-up.
- The use of artificial intelligence (AI) algorithms in predictive modeling for health-related outcomes has grown steadily over the last few years. Several algorithms have been adapted to the survival framework and perform well when applied to heterogenous clinical data in oncology, including **Random Survival Forest (RSF).**[2,3,4,5,6] While these methods have been used to predict survival outcomes up to the end of study follow-up and identify prognostic factors, to our knowledge no study has evaluated the efficacy of these methods for long-term extrapolation of survival outcomes.

OBJECTIVES

The main objective consisted in assessing whether RSF algorithms can **extrapolate** patient survival beyond study follow-up and compare them to standard survival modeling techniques.

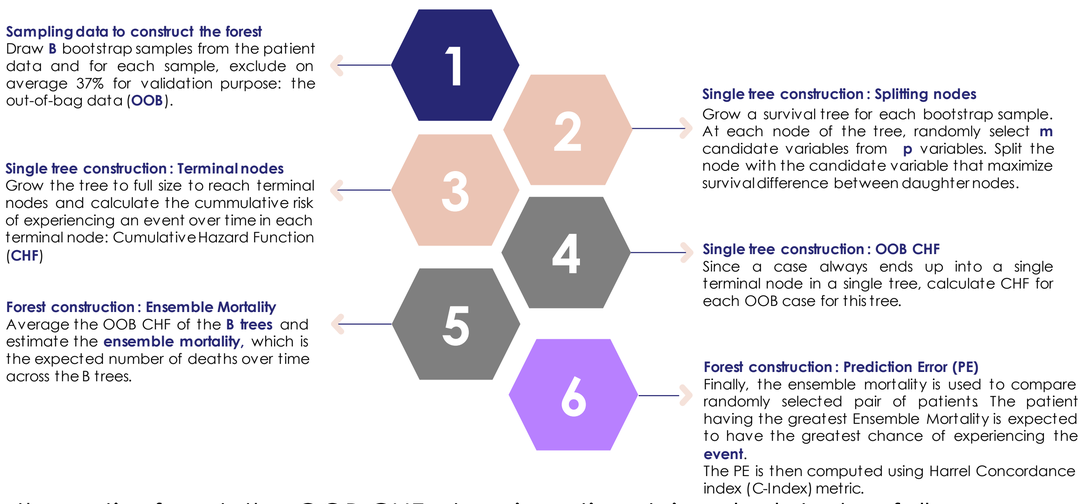
MATERIALS

- Individual patient data collected through the National Lung Screening Trial (NLST) were considered. NLST is a clinical trial assessing two ways of detecting lung cancer: Low-dose helical computed tomography (CT) and Standard chest X-ray.
- Among the 53,454 randomized participants, 2,058 were diagnosed with lung cancer.
- The RandomForestSRC package[6] developed by Hemant Ishwaran et al was used for the current analysis.

METHODS

- Random Forest Survival[6] (RSF), an extension of the Random Forest (RF) method developed by Leo Breiman[7] is a machine learning method designed for right-censored survival data analysis. **Fig.1** provides an overview of the algorithm steps :

Figure 1. RSF algorithm overview



- For the entire forest, the OOB CHF at a given time t, is calculated as follow:

$$\widehat{\mathbf{H}^{**}_{e}}(\mathbf{t}|\mathbf{x_i}) = \frac{\sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{I_{i,b}}\, \widehat{\mathbf{H}_{\mathbf{b}}}(\mathbf{t}|\mathbf{x_i})}{\sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{I_{i,b}}}$$

with $\boldsymbol{I_{i,b}}$ the boolean indicator describing if the individual ***i*** is an OOB case or not and $\widehat{H^{*}}_{b}(t)$ the CHF from a grown tree of the b[th] boostrap sample at a given time ***t.***

- The ensemble mortality is defined as follow:

$$\widehat{\boldsymbol{M}^{**}_{e,i}} = \sum_{j=1}^{n} \boldsymbol{H}^{**}_{e}\left(\boldsymbol{T_j}\middle|\boldsymbol{x_i}\right)$$

With $\boldsymbol{T_j}$ the pre-defined unique time points of OOB samples and $\boldsymbol{x_i}$ the case i set of covariate

- The C-Index, assesses the probability of having the **greatest risk of event** (based on the ensemble mortality) between two randomly selected pair of cases. The OOB PE is given by the following relationship: PE^{**}=1-C^{**} and $0 \leq PE^{**} \leq 1$. A small PE^{**} indicates a good model while a PE^{**} of 0.5 indicates that the model performs no better than random than a random model.

Data management

1. **Step 0: Select and encode relevant variables**

- Survival time (target variable) was estimated via $\boldsymbol{Y = (\delta_i, T_i)}$ composed of the censoring status and survival time (in months).
- The predictor variables $\boldsymbol{X = (X_1, .., X_p)}$ were composed of **538 selected variables** based on their clinical relevance, excluding variables with >=10% missingness.
- Categorical variables were encoded into binary variables using the one-hot encoding method.

2. **Patient selection:** 1,747 patients diagnosed with lung cancer, having no other cancer diagnoses, and no missing data for the included variables were included in the study.
3. **Extrapolation dataset:** An extrapolation set composed of 30% of patients from step 1 were randomly selected, leading to 523 cases. For each of these patients: **if survival time >=24,** an artificial cut-off of **24 months** and censoring status = 1 were attributed .

METHODS (continued)

4. **Training dataset:** A training data set was built from the complete cases dataset, excluding patients from the extrapolation set.

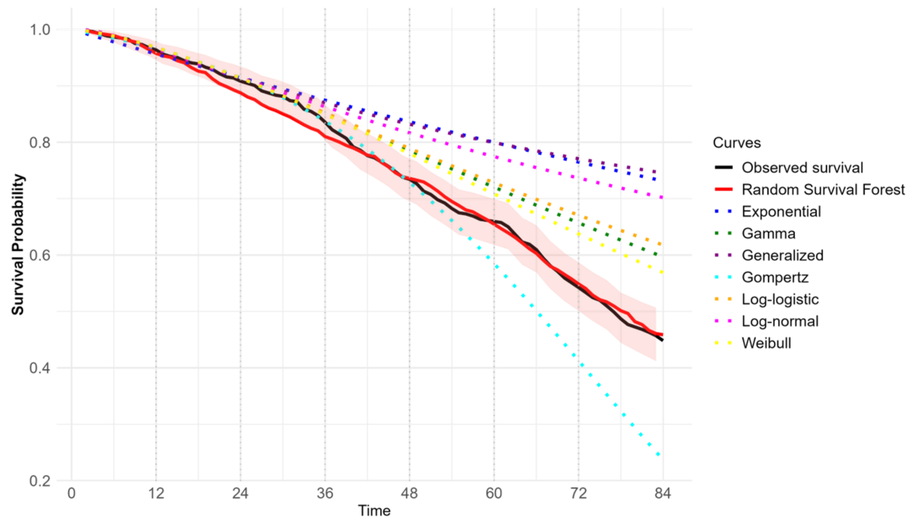
Models tuning and evaluation

- A grid search combined with K-fold (K=5) cross-validation was considered to sequentially train and evaluate the RSF model with relevant hyperparameters
- Parametric models were fitted on extrapolation data set with artificial cut-off, then overall survival was extrapolated beyond the artificial cut-off at 5 years horizon.
- On one hand, PE was not appropriate to evaluate parametric models. On the other hand, RSF predict survival probabilities at patient level hence, we averaged survival probabilities across all patient in the extrapolation dataset at each timepoint.
- Finaly, to compare performance between each model, Mean Absolute Error (MAE) was calculated accross timepoints between observed and predicted probabilities

RESULTS

- Survival curves of parametric and RSF models are represented along the true observed survival of patients in the extrapolation dataset (see **Fig.2**).
- The tuned RSF model averaged an OOB of **0.22** and seems to outperforms parametric models.
- Except for the Gompertz distribution, parametric models tended to **overestimate observed survival** overtime while RSF remained close to the observed survival curve within the confidence interval of the observed survival probability curve.

Figure 2. Survival probabilities by model vs observed data



- RSF demonstrated the lowest MAE at 0.01, with errors in other models ranging from 4 to 10 times greater.

CONCLUSIONS

- This study investigated the ability of RSF to extrapolate survival probabilities beyond study follow-up. This analysis was motivated by the ability of RSF to identify **clusters of patient** sharing the same terminal node with similar characteristics, explaining their survival probabilities
- This core concept introduces an innovative method for extrapolating survival time in clinical trials. By using RSF, we can estimate the survival of patients with limited follow-up based on the profiles of similar patients with extended follow-up.
- Our results suggests that predictive modeling using RSF has the potential to provide more reliable survival estimates than traditional parametric modelling while including a multitude of variables to reflect complex prognostic factors.
- One limitation of this work is that the standard parametric models were not adjusted on prognostic factors, while RSF included many predictor variables.
- This analysis is a standalone study using data from clinical trial. Future studies could consider replicating this work using Real World Evidence (RWE) data to assess robustness of the approach.
- While this study focused on the RSF method, additional AI models have been adapted to the survival framework as well. Future studies including these methods in the comparison of extrapolation performance is warranted.

REFERENCES and ABBREVIATIONS

1. Latimer NR and Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. BMJ Med. 2022;10(1)e000094. doi: 10.1136/bmjmed-2021-000094.
2. Xiao J *et al*. The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study. JMIR Med Inform. 2022;10(2):e33440. doi: 10.2196/33440.
3. Wang D *et al*. Development, and validation of machine learning models for predicting prognosis and guiding individualized postoperative chemotherapy: A real-world study of distal cholangiocarcinoma. Front Oncol. 2023;13:1106029. doi: 10.3389/fonc.2023.1106029.
4. Spooner A *et al*. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. 2020;10(1):20410. doi: 10.1038/s41598-020-77220-w.
5. Cygu S *et al*. Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. Sci Rep. 2023;13(1):1370. doi: 10.1038/s41598-023-28393-7.

6.Hemant I et al (2008). Random Survival Forests.The annals of Applied Statistics, 841-860
7.Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324.

Abbreviation: MAE: Mean Absolute Error; OOB: Out-Of-Bag; PE: Prediction Error; RSF: Random Survival Forest

DISCLOSURES

Authors have no conflict of interest to declare.

Presented at: ISPOR Europe 2024; 17-20 November 2024; Barcelona, Spain