Data-Adaptive Estimation in Time-to-Event Data with Competing Risks: A Super Learner Analysis of Alzheimer's Disease Data from Clalit EHR

Yunlong Feng[1], Ran Abuhasira[2,3,4], Bella Vakulenko-Lagun[1]

[1]Department of Statistics, University of Haifa, [2]Clinical Research Center, Soroka University Medical Center, [3]Massachusetts General Hospital, [4]Harvard Medical School

אוניברסיטת חיפה
UNIVERSITY OF HAIFA
جامعة حيفا

Background

Super Learner:

An ensemble approach which provides a flexible, data-adaptive framework for optimally combining traditional statistical models and modern machine learning algorithms, with the goal of more robust estimation of complex functionals. For example:

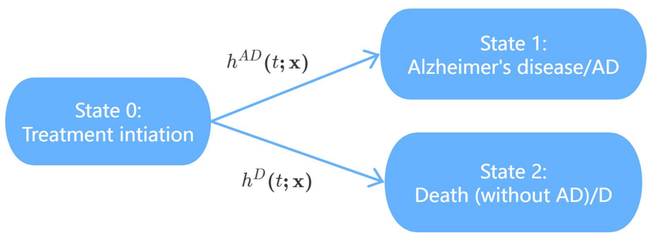
$$\hat{h}(\hat{\mathbf{w}}, t; \mathbf{x}) = \sum_{k=1}^{K} \hat{w}_k \hat{h}_k(t; \mathbf{x}), \quad \sum_{k=1}^{K} \hat{w}_k = 1, \ \hat{w}_k \geq 0.$$

*where $\hat{h}(\hat{\mathbf{w}}, t; \mathbf{x})$ is the hazard estimate of Super Learner, $\hat{h}_k(t; \mathbf{x})$ is the estimate of the k-th candidate learner, $\hat{w}_k$ is its optimal weight, t is time and **x** are the covariates.*

A Super Learner combines *K candidate learners* through an optimal convex combination via weights $\hat{\mathbf{w}}$ selected by cross-validation, to minimize a *loss* function.

Competing Risks Model:

The Super Learner was first introduced almost 20 years ago. However, there are no existing Super Learner approaches, developed specifically for right-censored time-to-event data with competing risks.



A competing risks model with two competing events: Alzheimer's disease and death, $h^{AD}(t;\mathbf{x})$ and $h^{D}(t;\mathbf{x})$ are the cause-specific hazards.

Method

We extend the Super Learner to a competing risks setting, which targets estimation of cause-specific hazards in both *continuous-time* and *discrete-time* settings.

Data Transformation:

Coutinuous-time Super Learner: we transform the data into 2 group. E,g, for a dataset of 3 observations: $(t_1, x_1, e_1 = 0), (t_2, x_2, e_2 = 1), (t_3, x_3, e_3 = 2).$

Group 1:
$(t_1, x_1, e_1 = 0), (t_2, x_2, e_2 = 1), (t_3, x_3, e_3 = 0).$

Group 2:
$(t_1, x_1, e_1 = 0), (t_2, x_2, e_2 = 0), (t_3, x_3, e_3 = 2).$

Discrete-time Super Learner: we transform data to augmented matrix. For example, an obsevation $(t_1, x_1, e_1 = 1)$: we discretize the time into intervals $\tau_1, \tau_2, \cdots, \tau_t,$ where t_1 falls in τ_t, the augmented matrix for this observation is:

$$\begin{pmatrix} \tau_1, x_1, e_1 = 0 \\ \tau_t, x_1, e_1 = 0 \\ \vdots \\ \tau_t, x_1, e_1 = 1 \end{pmatrix}$$

Framework:

Continuous-time Super Learner

Data Transformation

Discrete-time Super Learner

Data Transformation

Data Splitting: Randomly partition the data into V folds. For each fold, use $V-1$ folds as the training set, and for the v-th fold calculate predictions.

Model Training: Fit all candidate learners (eg. CoxPH, GAM in continuous-time model; KNN, SVM, XGBoost in discrete-time model) on the training set.

Loss Calculation: Compute *the loss** that compares predictions to the true observed values.

Weight Optimization: The optimal weights for the candidate learners for each event are determined by minimizing the loss, using optim() function in R.

Final Model Fitting and Hazards Calculation: Fit all candidate learners on the entire dataset, and use the optimal weights obtained to get the estimates of cause-specific hazards.

*Continuous-time Super Learner loss is

$$-\sum_{i \in O_{(v)}} \left\{ \mathbb{I}(E_i = 1) \ln\left(\hat{h}^{AD}_{(v)}(\mathbf{w}^{AD}, t_i; \mathbf{x}_i)\right) - \int_0^{t_i} \hat{h}^{AD}_{(v)}(\mathbf{w}^{AD}, s; \mathbf{x}_i)\, \mathrm{d}s \right\}$$

$$-\sum_{i \in O_{(v)}} \left\{ \mathbb{I}(E_i = 2) \ln\left(\hat{h}^{D}_{(v)}(\mathbf{w}^{D}, t_i; \mathbf{x}_i)\right) - \int_0^{t_i} \hat{h}^{D}_{(v)}(\mathbf{w}^{D}, s; \mathbf{x}_i)\, \mathrm{d}s \right\}.$$

$O_{(v)}$ is the v-th fold test data, $\hat{h}^{j}_{(v)}(\mathbf{w}^{j}, t_i; \mathbf{x}_i)$ is the estimated cause-specific hazard for Observation i, Event j, using the training data of Fold v.

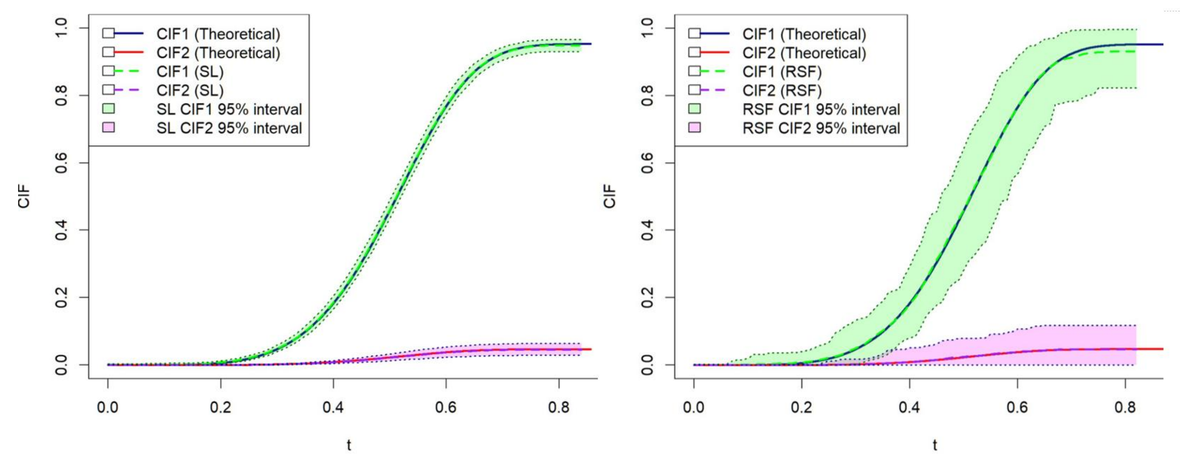
Discrete-time Super Learner loss is

$$-\sum_{i=1}^{N^*} \sum_{j=0}^{2} \mathbb{I}(e_i = j) \ln(\hat{e}_{ij}),$$

N is the total raw of the augmented matrix. $\hat{e}_{ij}$ is the one-hot code for the estimated cause-specific hazard of Observation i, Event j.*

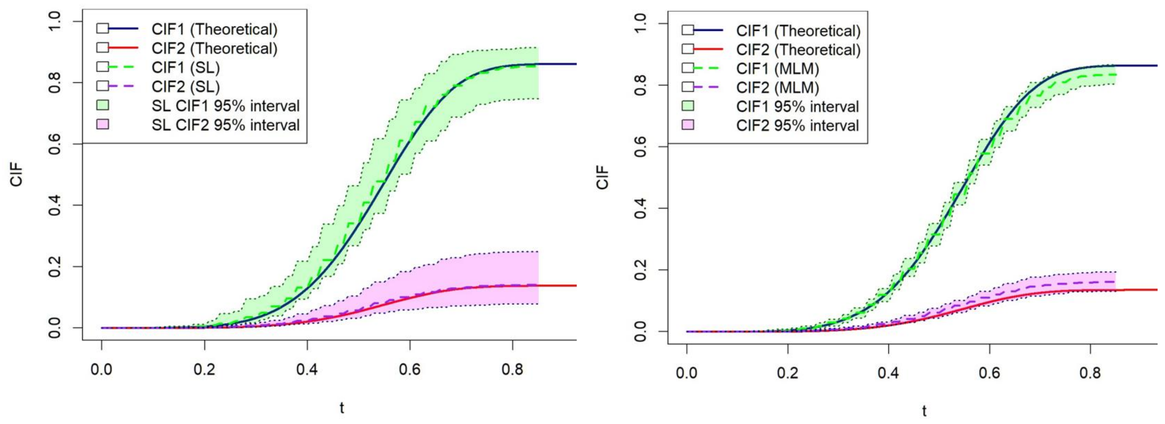
Simulations

Continous-time Super Learner:



Comparison of conditional cause-specific hazard estimates under simulated data without a common unmeasured risk factor, for a representative covariate setting. The Super Learner model (left) performs better than Random Survival Forest (RSF, right), a flexible non-parametric method that tends to exhibit high variance.

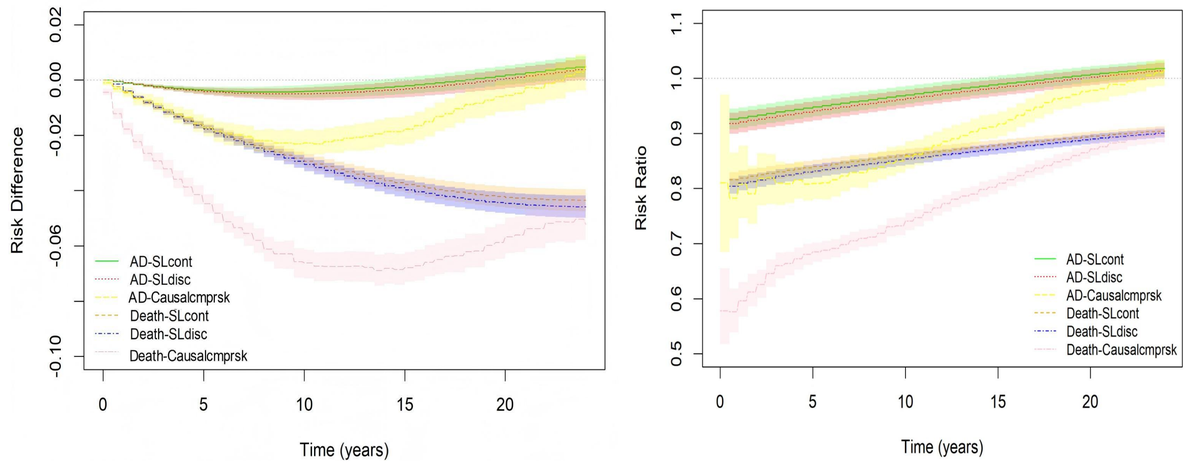
Discrete-time Super Learner:



Comparison of conditional (on x) cause-specific hazard estimates under simulated data with a common unmeasured risk factor, for a representative covariate setting. The Discrete Super Learner model (left) exhibits smaller bias than the Multinomial Logit Model (MLM, right). By combining multiple machine learning methods (e.g., SVM, KNN, XGBoost), the Discrete Super Learner provides greater flexibility, although this may lead to increased variance.

Case study

We applied our Super Learner to the data from Clalit EHR (up to 25 years of follow-up), as part of the drug-repurposing study of anti-diabetic treatments for Alzheimer's disease, and compared the effects of *metformin* vs *sulfonylurea*, on the time to Alzheimer's disease onset and the time to (competing) death (without prior AD).



Both the continuous-time and discrete-time Super Learners yield similar results and the same conclusion: treatment with metformin is associated with lower cumulative probability of both Alzheimer's disease and competing death compared with sulfonylurea.

This research has been supported by the National Institute of Health (grant number R01AG058063).