# Boosted Trees for Risk Prognosis

ALEXIS BELLOT* AND MIHAELA VAN DER SCHAAR*,†, † ALAN TURING INSTITUTE, * UNIVERSITY OF OXFORD

## OBJECTIVE

• **Survival Analysis:** Predict event probabilities over time with the aim of providing a *fully individualized survival function* for each patient,

$$S(t|\boldsymbol{x}_i) = \mathbb{P}(T_i > t|\boldsymbol{x}_i) \quad (1)$$

Important use as a risk score based on which clinicians design therapies for patients.

• **Example:** Survival prognosis for **chronic** or **multi-morbid** patients whose risk factors are poorly understood (e.g. cardiovascular diseases and elderly patients).

## PROBLEM: HETEROGENEITY

• **Major Challenge:** Large variability among patients in heterogeneous populations is the result of complex relationships between covariates and survival.

▷ Leads to misdiagnoses in patients with atypical disease presentation or risk factors = **major source of patient harm**.

• **Current methods** impose modelling assumptions that limit $S$ to reflect the behaviour of the *average* patient and as a result are inefficient for large portions of the population.

▷ Cox / Boosting based on Cox / Random Survival Forests

## OUR APPROACH

**Focus on complex patterns and subgroups of patients that are consistently being misdiagnosed**

• We train shallow survival trees on **re-weighted** samples of the patient population such as to iteratively improve prognosis of misdiagnosed patients.

▷ Efficient scheme for learning in high-dimensional settings
▷ No a-priori assumptions on patient behaviour
▷ Very flexible and thus provides *individualized* predictions

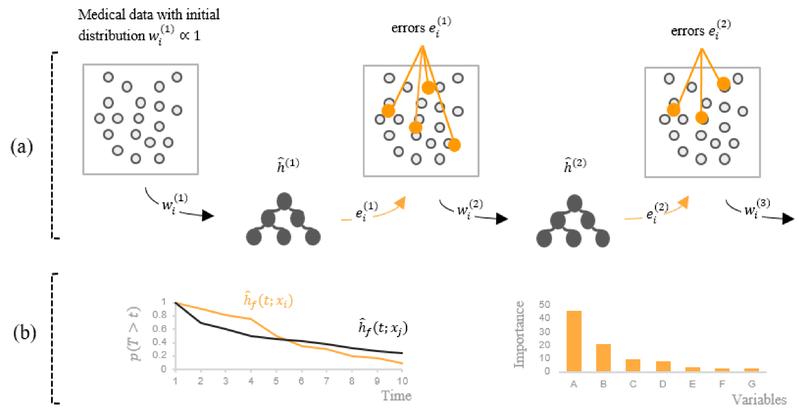
## BOOSTING SURVIVAL TREES

• **Measuring Misdiagnoses** We propose an aggregate measure that quantifies the error in survival estimation as follows,

$$\frac{1}{\tau}\int_0^\tau \mathbb{E}\left[\left(I(T_i > t) - \hat{h}(t; \boldsymbol{X}_i)\right)^2\right] dt \quad (2)$$

• **Survival Tree Construction** Trees are composed of leaves and nodes.

▷ Nodes partition the sample space into more homogeneous subgroups using the deviance as splitting criterion.
▷ Leaves estimate survival distributions with the Kaplan Meier estimator.




• **Ensemble by Boosting** In each iteration of the algorithm,

▷ Compute the error $e_i$ for each patient, given by (2).
▷ Increased weight $w_i$ is subsequently assigned to patients for which the error is largest, and lowered weight for more accurate predictions.
▷ Final survival estimates result from a weighted average of individual survival trees, weighted by a measure of overall accurateness of individual trees.

• **Covariate importance** We measure covariate importance by examining the improvement in goodness of fit given by each split involving a given variable.

## EXPERIMENTS

We investigate the discriminative ability of our methods on patients at various stages of the trajectory of cardiovascular diseases.

▷ Preventive Care: UK Biobank and MAGGIC
▷ Heart transplant wait-list management: UNOS
▷ Cancer diagnosed patients: SEER

• **Baseline Algorithms**: Cox, boosting approaches based on Cox (CBL, CBM, CindexBoost), bagging approaches (SRF, CSRF).

**Table 1:** $C$-index (higher better).

| Models | UNOS | MAGGIC | UK Biobank | SEER-I | SEER-II |
|---|---|---|---|---|---|
| Cox | $0.603 \pm 0.04$ | $0.645 \pm 0.01$ | $0.679 \pm 0.02$ | $0.772 \pm 0.03$ | $0.740 \pm 0.03$ |
| CBL | $0.605 \pm 0.04$ | $0.644 \pm 0.01$ | $0.679 \pm 0.02$ | $0.774 \pm 0.03$ | $0.738 \pm 0.04$ |
| CBM | $0.635 \pm 0.03$ | $0.625 \pm 0.01$ | $0.673 \pm 0.02$ | $0.768 \pm 0.03$ | $0.740 \pm 0.04$ |
| CindexBoost | $0.564 \pm 0.06$ | $0.592 \pm 0.01$ | $0.655 \pm 0.03$ | $0.764 \pm 0.03$ | $0.742 \pm 0.04$ |
| SRF | $0.634 \pm 0.04$ | $0.642 \pm 0.01$ | $0.627 \pm 0.01$ | $0.686 \pm 0.03$ | $0.680 \pm 0.01$ |
| CSRF | $0.635 \pm 0.05$ | $0.652 \pm 0.02$ | $0.638 \pm 0.02$ | $0.755 \pm 0.03$ | $0.717 \pm 0.04$ |
| SurvivalBoost.R | $0.636 \pm 0.03$ | $0.676 \pm 0.02$ | $0.702 \pm 0.02$ | $\mathbf{0.780 \pm 0.03}$ | $\mathbf{0.752 \pm 0.03}$ |
| SurvivalBoost.T | $\mathbf{0.647 \pm 0.04}$ | $\mathbf{0.675 \pm 0.04}$ | $\mathbf{0.725 \pm 0.03}$ | $0.775 \pm 0.04$ | $0.740 \pm 0.04$ |

## IMPACT

• **Technical Significance** First nonparametric extension of boosting architectures to survival analysis.

▷ New notion of prediction "correctness" which successfully improves predictions of mis-estimated patients.
▷ New insights into covariate importance with a novel procedure derived from our model.

• **Medical Relevance** Contribution towards **"precision medicine"**.

▷ Patient heterogeneity is one of the major reasons for the large share of misdiagnoses in chronic diseases.
▷ Individualized predictions from our model will improve long term prognosis, even for atypical patients.

**Try our App: mlhcprojects.shinyapps.io/survival_boosting_app**