Title: Simplifying Inputs of Machine Learning-Based Risk Calculators to Improve Clinical Usability: a Report from the Childhood Cancer Survivor Study and the St Jude Lifetime Cohort

Authors: Zhe Lu1, Kenneth Liao2, Zhiyu Kang2, Linwan Xu1, Gregory T. Armstrong3, Lucie M Turcotte2, Wendy Leisenring4, Mellissa M Hudson3, Kirsten K Ness3, Cindy Im2* and Yan Yuan1*

*Equal contribution

Affiliations:
1. University of Alberta
2. University of Minnesota
3. St. Jude Children's Research Hospital
4. Fred Hutchinson Cancer Center

Abstract:
BACKGROUND: Automatic variable selection is inherent in machine learning algorithms and supports the development of medical risk calculators with better risk prediction performance. However, from the user perspective, automatic variable selection results in models with high administrative burden because data for all considered inputs is needed. For example, the published risk calculator for basal cell carcinoma (BCC) for childhood cancer survivors currently available on the Childhood Cancer Survivor Study (CCSS) website needs 19 different clinical inputs. To improve the usability of these calculators, we aim to leverage the benefits of machine learning algorithms and simplify the number of model inputs with little if any sacrifice in prediction accuracy.

METHODS: We developed a manual variable selection approach for the machine learning algorithm Extreme Gradient Boosting (XGBoost), which was the best performing algorithm underlying our published risk calculators. Our approach, XGBoost-Elbow (XGBoost-E), sequentially removes the least important variable (i.e., smallest SHAP value) until the reduced set shows worse performance than the previous larger set in all three evaluation metrics: discrimination (AUROC), precision (AUPRC), and overall performance (sBrS), i.e., an inflection point or "elbow" in performance. A simulation study was conducted where the simulated outcome Y depended on a function of 5 key inputs. Five noise inputs were added to the dataset to assess whether XGBoost-E could remove noise inputs and keep key inputs. We then applied XGBoost-E for a BCC risk prediction model using real-world data in CCSS followed by external validation in the St. Jude Lifetime Cohort (SJLIFE), using the full set of 19 inputs as a benchmark.

RESULTS: In 500 simulation datasets (n=5365, mean event rate=0.27), XGBoost-E kept on average 4.4 variables, of which <1 (mean=0.76) was a noise input. Three of five key inputs were always kept; the remaining two key inputs were rarer exposures (9% and 3% exposed) and were kept inconsistently (55% and 9%, respectively). Compared to the XGBoost models with the full input set, the XGBoost-E models with reduced inputs had similar or better prediction performance with respect to the three metrics: 71% (AUROC), 50% (AUPRC), and 48% (sBrS), respectively, using a nested cross-validation framework. In the BCC real-world example, XGBoost-E selected 12 inputs from the original 19 predictors resulting in a reduced model with similar or better prediction performance than the full XGBoost model evaluated at age 40, internally in CCSS (reduced vs. full: AUROC 0.747 vs. 0.745; AUPRC 0.153 vs. 0.155; sBrS 0.049 vs. 0.050) and externally in SJLIFE (AUROC 0.762 vs. 0.760; AUPRC 0.321 vs. 0.308;

sBrS 0.079 vs. 0.074). Dropped inputs included radiation therapy doses to the limbs and pelvis, and specific chemotherapies (anthracyclines, vinca alkaloids, platinum, corticosteroids).

CONCLUSION: Our new approach XGBoost-E consistently drops noise inputs. The models with reduced input sets perform comparably to full-set XGBoost models in our simulation study. In our use case, BCC risk prediction with XGBoost-E selected reduced inputs that performed better than the full-set XGBoost model in external validation data. Updating our survivor-specific medical risk calculators with this approach will improve their clinical usability.