**Study Title: Benchmarking the accelerated oblique random survival forest**

**Principal Investigator, Co-investigator(s):**

Byron C. Jaeger, PhD

**Other staff to have on IRB**

None

**Sponsor or funding source:** Wake Forest Biomedical Informatics Institute (PI Jaeger)

**Background, Rationale and Context**

The oblique random survival forest (ORSF) is a supervised learning algorithm that has been shown to have excellent prediction accuracy in general benchmarks and has also proven to be the most accurate prediction model for incident heart failure in an analysis of several epidemiological cohorts.1,2 The primary limitation of ORSF is its low computational efficiency. We have developed an accelerated ORSF model that is over 500 times faster than the original ORSF. To show that the accelerated ORSF model can predict outcomes just as accurately as its predecessor, we propose to run a general benchmark using multiple datasets from the BIOLINCC repository managed by the NHLBI.

**Methods and Measures**

**Design**

Observational analysis of data publicly available in BIOLINCC.

**Setting**

Wake Forest Baptist Health

**Subjects Selection Criteria**

* **Inclusion Criteria**

Participants in the BIOLINCC repositories for the Jackson Heart Study (JHS), Atherosclerosis Risk in Communities Study (ARIC), and Multi-Ethnic Study of Atherosclerosis (MESA)

* **Exclusion Criteria**
* We will restrict the population to patients who consented to have cardiovascular events and all-cause mortality events adjudicated during the follow-up periods of their respective trials.
* **Sample Size**

The JHS has a sample size of roughly 4,000, ARIC 15,500, and MESA 6,800.

**Interventions and Interactions**

This project will not involve any interaction with patients. It will consist entirely of secondary analyses of public data

**Outcome Measure(s)**

We will evaluate incident cardiovascular events and mortality following the baseline visit of each study.

**Analytical Plan**

We will benchmark the accelerated ORSF and 9 other machine learning algorithms in the requested datasets. Our analysis will split the data into a 50/50 training and testing set and apply each machine learning algorithm to create a prediction model with the training data. Next, we will evaluate the accuracy of each prediction model in the testing data, and we will repeat this split-test procedure 100 times to gather an adequate amount of data on the performance of each machine learning algorithm. Last, we will fit a hierarchical Bayesian linear mixed model to the performance data of the algorithms, accounting for correlation within datasets and within each split of the data, to estimate the overall performance of each algorithm and also estimate the posterior expected differences in performance between models.

**Human Subjects Protection**

**Subject Recruitment Methods**

As this is a secondary data analysis, not applicable.

**Informed Consent**Written informed consent will not be obtained.  The risk of harm or discomfort that may occur as a result of taking part in this research study is not expected to be more than in daily life or from routine physical or psychological examinations or tests.  The rights and welfare of study will be protected through the use of measures to maintain the confidentiality of study information.  Study results will be presented or published in lieu of providing individual subjects additional information regarding the study.

**Confidentiality and Privacy**

Confidentiality will be protected by transferring only deidentified data from the BIOLINCC repositories to secure workstations managed by the primary investigator of the study.

**Data and Safety Monitoring**

The principal investigator will be responsible for the overall monitoring of the data and safety of study participants.  The principal investigator will be assisted by other members of the study staff.

**Reporting of Unanticipated Problems, Adverse Events or Deviations**

Any unanticipated problems, deviations or protocol changes will be promptly reported by the principal investigator or designated member of the research team to the IRB and sponsor or appropriate government agency if appropriate.

**References**

1 Segar MW, Jaeger B, Patel KV, *et al.* Development and Validation of Machine Learning-based Race-specific Models to Predict 10-year Risk of Heart Failure: A Multi-cohort Analysis. *Circulation* 2020; **142**: A196–A196.

2 Jaeger BC, Long DL, Long DM, *et al.* Oblique random survival forests. *The Annals of Applied Statistics* 2019; **13**: 1847–83.