Submitted by: Thomas Nyberg, Postdoctoral Researcher in Statistics, Columbia University.

Research Objectives and Aims: There is currently no established practice for testing the validity of large-scale computational statistical algorithms. Existing software does not have well-documented testing frameworks, and those applications do not scale to the problems faced in observational health research, where we desire to fit large-scale models with millions of observations and hundreds of thousands of covariates. We plan to develop a statistical testing framework for CCD and other modeling applications within OHDSI (Observational Health Data Sciences and Informatics). The goal is to understand the kinds of inputs needed to verify the validity of implementations of statistical algorithms under the restriction that there is no known comparable reference implementation.

Scope/Proposed Approach: We plan to research and develop such a framework, likely borrowing from different disciplinary thinking on testing, ranging from a computer science orientation toward deterministic boundary conditions to small-scale statistics concepts of convergence and simulation to assess expected behavior in stochastic systems. We plan to use the Research Lab to test large-scale statistical software as we develop it, and to publish both the software and the testing paradigm that we develop to ensure that the algorithm functions as expected.

Many modeling applicationss in R, like glmnet or rpart, will fall over once we get beyond 100,000 records or 10,000 covariates. An example of a typical problem we may want to solve is to predict treatment assignment (e.g. estimate propensity score) between two commonly used drugs (ex: metformin vs. gliburide, each with greater than 1 million new users). There is a near-infinite number of potential predictors, but even doing something simple like using indicator variables for all conditions, all drugs, all procedures, and all lab tests, would give you roughly 50k covariates and millions of rows. Access to the Research Lab is essential due to the large amounts of data involved.

Impact: As ever more complex statistical algorithms are implemented, the accuracy of their implementations requires increasing scrutiny. A general framework for statistical testing would increase confidence in all statistical software and would build more confidence when further implementing statistical algorithms.

Experience: Thomas Nyberg, PhD in Mathematics, Columbia University, 2014.

Timeline: 12 months.