International Journal of Population Data Science

Journal Website: www.ijpds.org

A Data Science Approach to Predictive Analytic Research and Knowledge Translation

Fisher, S[1,2,3], Talarico, R[2], Sequeira, Y[1], Bennett, C[1,5], Spruin, S[1], Hsu, A[4,5,6,7], Tanuseputro, P[4,5,6,7], and Manuel, D[4,5,6,7]

[1]Ottawa Hospital Research Institute
[2]Institute for Clinical Evaluative Sciences
[3]University of Ottawa
[4]Ottawa Hospital Research Institute, Clinical Epidemiology Program
[5]Institute for Clinical Evaluative Sciences, UOttawa
[6]Bruyère Research Institute
[7]University of Ottawa, School of Epidemiology and Public Health
[8]University of Ottawa, Department of Family Medicine

Introduction

Current approaches to the development and application of predictive studies is inefficient and difficult to reproduce. Thousands of predictive health algorithms have been developed; however, less than 2% have been assessed outside their original setting and even fewer have been applied and evaluated in practice.

Objectives and Approach

Objective: To develop a standardized workflow for algorithm development, dissemination and implementation.

Existing predictive analytics workflow and open standards were adapted and expanded for health research and health care settings. The approach was designed to work within multidisciplinary teams and to improve research transparency, reproducibility, quality, efficiency and application. Key components include standardized algorithm description files, documentation and code libraries. All libraries and programming packages, which were created for/with open-source software, can be used for a wide range of statistical and machine learning models. Publicly-available repositories contain the algorithms, validation data, R code and other supporting infrastructure.

Results

Algorithm development involves variable pre-specification and documentation of model variables, followed by creation of data preprocessing code to generate model variables from the study dataset. Preprocessing uses algorithm specification documentation and a function library, building upon and integrating with existing algorithms when possible to preventing code duplication. Models are output as a Predictive Modelling Markup Language (PMML) file, a portable industry standard for describing and scoring predictive models. A separate scoring "engine" is used to implement PMML-described algorithms in a range of settings, including algorithm validation at other research institutions. Algorithm applications currently include the Project Big Life (www.projectbiglife.ca) online calculators, population, health services and public health planning uses and an algorithm visualization tool. An API permits use of the calculator engine by other organizations.

Conclusion/Implications

Barriers to the implementation of predictive analytics in real-world settings—such as within electronic medical records or decision aid applications—can be mitigated with well described algorithms that are easy to replicate and implement, especially as access to big health data increases and algorithms become increasingly complex.

http://dx.doi.org/10.23889/ijpds.v3i4.797