Machine Learning to Automate Clinician Designed Empirical Manual for Congenital Heart Disease Identification in Large Claims Database

Ariane J. Marelli, MD, MPH;[1] *Chao Li, MEng;*[1] *Aihua Liu, PhD;*[1] *Hanh Nguyen, MD;* [1] *James M Brophy, MD, PhD;*[2]
Liming Guo, MSc;[1] *David L Buckeridge, PhD;*[2] *Jian Tang, PhD;*[3] *Joelle Pineau, PhD;*[4] *Yi Yang, PhD;* [5] *Yue Li, PhD* [4]
[1]*McGill Adult Unit for Congenital Heart Disease Excellence, Montreal, Québec, Canada*
[2]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Québec, Canada*
[3]*Department of Decision Sciences HEC, Université de Montréal, Montreal, QC, Canada*
[4]*School of Computer Science, McGill University, Montreal, Québec, Canada*
[5]*Department of Mathematics and Statistics, McGill University, Montreal, Québec, Canada*

Background.
Large claims databases are increasingly used in medical research. Correct identification of patients with the disease of interest is an essential step to producing evidence-based health services in large populations. For complex diseases such as congenital heart disease (CHD), we have created manuals that require clinician audits and intensive labor to achieve internally valid results. To automate this process, machine learning (ML) methods can be employed. This study aims to evaluate ML approaches in automating a clinician designed manual to identify patients with CHD in large claims databases.

Methods.
We first adopted a clinician-developed empirical manual that is well established in literature to classify CHD patient based on claim and hospitalization data. This manual is extremely time consuming to implement and therefore calls for an automated process. To this end, we sought an efficient ML method to learn the latent rules that can automate the clinician-decision making process. To train such ML model, we harnessed data from the Quebec claims, hospitalization, and vital status databases. The study cohort included 19,187 patients. Of them, 3,784 patients were labelled as true CHD patients using the clinician-developed empirical manual. We extracted a set of 82 features including patient demographics, diagnosis codes and their respective physician/specialist types, and procedure/surgery codes. For each of the 25 CHD-related ICD-9 diagnoses, we summarized the number of the diagnoses during the entire follow-up period to reflect the changes over time in CHD management guidelines with the adjustment for birth year. The study cohort was randomly divided into training (80%) and testing sets (20%). ML approaches including Gradient Boosting Decision Tree (GBDT) and Support Vector Machine (SVM) were adopted. For GBDT model, 5-fold cross-validation method was employed to select the optimum solution on tree depth as well as the total number of boosting trees. To choose optimal solution of SVM model, the linear or non-linear kernels by the 5-fold cross validation was adopted. The Area Under the Precision Recall Curve (AUPRC) was used to choose the optimum parameters in the process of cross-validation on the training set. In parallel to GBDT and SVM models, traditional logistic regression and decision tree models were implemented for comparison purpose and to help with results interpretation. Diagnostic performance was evaluated on AUPRC, F1-score, accuracy, specificity, sensitivity using the test set. Evaluations were repeated 10 times each time with different random split of 80/20 training/testing to achieve the median performance statistics.

Results.
Here we demonstrate for the very first time a tedious and time consuming clinical inspection for CHD patient identification can be replaced by an extremely efficient and completely automated ML algorithm. Indeed, both the GBDT and SVM models showed excellent performance with >=95% in all performance statistics and outperformed decision tree and logistic regression in identifying true CHD patients. The GBDT model led the performance with a 99.3% for both AUPRC and accuracy, 98.8% for F1-score, 98.0% for sensitivity, and 99.7% for specificity. The top 10 important features were similar between the GBDT and the decision tree models.

Conclusion.
This is the first study to show that ML models especially GBDT can automate the clinician designed manual to identify true CHD patients in large claims database. The high precision and recall suggests that our approach is able to capture the hierarchical regularities that were derived based on long-term expertise of clinicians. Our findings are important to a field

where automation of clinically derived manual for a complex medical condition will greatly facilitate the acquisition of empirical evidence to support health services policy.