JACC: ADVANCES VOL. 3, NO. 2, 2024

ORIGINAL RESEARCH

Machine Learning Informed Diagnosis for Congenital Heart Disease in Large Claims Data Source

Ariane J. Marelli, MD, MPH,[a] Chao Li, MEng,[a] Aihua Liu, PhD,[a] Hanh Nguyen, MD,[a] Harry Moroz, BS,[a] James M. Brophy, MD, PhD,[b] Liming Guo, MSc,[a] David L. Buckeridge, PhD,[b] Jian Tang, PhD,[c] Archer Y. Yang, PhD,[d] Yue Li, PhD[e]

ABSTRACT

BACKGROUND With an increasing interest in using large claims databases in medical practice and research, it is a meaningful and essential step to efficiently identify patients with the disease of interest.

OBJECTIVES This study aims to establish a machine learning (ML) approach to identify patients with congenital heart disease (CHD) in large claims databases.

METHODS We harnessed data from the Quebec claims and hospitalization databases from 1983 to 2000. The study included 19,187 patients. Of them, 3,784 were labeled as true CHD patients using a clinician developed algorithm with manual audits considered as the gold standards. To establish an accurate ML-empowered automated CHD classification system, we evaluated ML methods including Gradient Boosting Decision Tree, Support Vector Machine, Decision tree, and compared them to regularized logistic regression. The Area Under the Precision Recall Curve was used as the evaluation metric. External validation was conducted with an updated data set to 2010 with different subjects.

RESULTS Among the ML methods we evaluated, Gradient Boosting Decision Tree led the performance in identifying true CHD patients with 99.3% Area Under the Precision Recall Curve, 98.0% for sensitivity, and 99.7% for specificity. External validation returned similar statistics on model performance.

CONCLUSIONS This study shows that a tedious and time-consuming clinical inspection for CHD patient identification can be replaced by an extremely efficient ML algorithm in large claims database. Our findings demonstrate that ML methods can be used to automate complicated algorithms to identify patients with complex diseases.

From the [a]McGill University Health Centre, McGill Adult Unit for Congenital Heart Disease Excellence, Montreal, Québec, Canada; [b]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Québec, Canada; [c]Department of Decision Sciences HEC, Université de Montréal, Montreal, Québec, Canada; [d]Department of Mathematics and Statistics, McGill University, Montreal, Québec, Canada; and the [e]School of Computer Science, McGill University, Montreal, Québec, Canada.
The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received October 6, 2022; revised manuscript received August 10, 2023, accepted October 20, 2023.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2023.100801

ABBREVIATIONS AND ACRONYMS

AUPRC = area under the precision recall curve

AUROC = area under the receiver operating characteristic curve

CHD = congenital heart disease

GBDT = gradient boosting decision tree

ML = machine learning

SVM = support vector machine

VSD = ventricular septal defect

Harnessing large claims databases in medical practice and research has potential and merits. These databases provide large samples and less referral bias than other sources of information. Efficient identification of patients with the disease of interest is the first and an essential step. For complex disease, it usually requires substantial time-consuming work including development of hierarchical algorithm for disease classification and validation through tedious manual audits. For example, congenital heart disease (CHD) presents great varieties in cardiac lesion manifestation and diagnostic and treatment regimens. There are 24 CHD classification codes based on the International Classification of Diagnosis (ICD)-version 9. The upcoming ICD 11 will have even more diagnostic codes.[1] All these portend great challenges in identifying CHD patients in large claims databases. To overcome the obstacle for using large claims database in CHD research, we have developed an empirical algorithm based on clinician expertise and substantial clinician audits.[2] To verify the algorithm, substantial manual audits were done. Out of the original 61,386 patients in the database, data files for 17,474 patients were manually reviewed by 2 cardiologists (A.J.M and A.S.M). To make sure that the algorithm performs well with different patient profiles, the subsets of patients selected for audits cover all categories of subjects, including those excluded, with severe and other CHD, with unspecified defects, hospitalized (operated and unoperated), and outpatients. It has been proven to be valid and has been well established in literature to classify CHD patients based on claim and hospitalization data.[2-8] However, the algorithm is extremely time-consuming to implement and therefore calls for an automated process. We now turn to machine learning (ML) methods to learn the latent rules that can automate the clinician decision-making process.

ML approaches such as tree-based models[9] and deep learning[10] are able to identify complex data patterns among variables by exploring the manifold of high order of complex interactions. These ML methods have potential to help physicians to explain the complex cardiovascular disease mechanisms.[11] Effectively using these ML algorithms, we can generate more accurate prediction and classification, and benefit cardiovascular medicine and beyond.[12-14] Previous studies have used ML in medical field for tasks such as patient classification[15] or predicting emergency admission.[16,17]

In this study, we aim to apply existing ML approaches to large claims databases to accurately identify CHD patients. We evaluated a variety of ML approaches including both generalized linear approaches such as regularized logistic regression and completely nonlinear approaches such as decision tree and gradient tree boosting on the CHD phenotyping task.

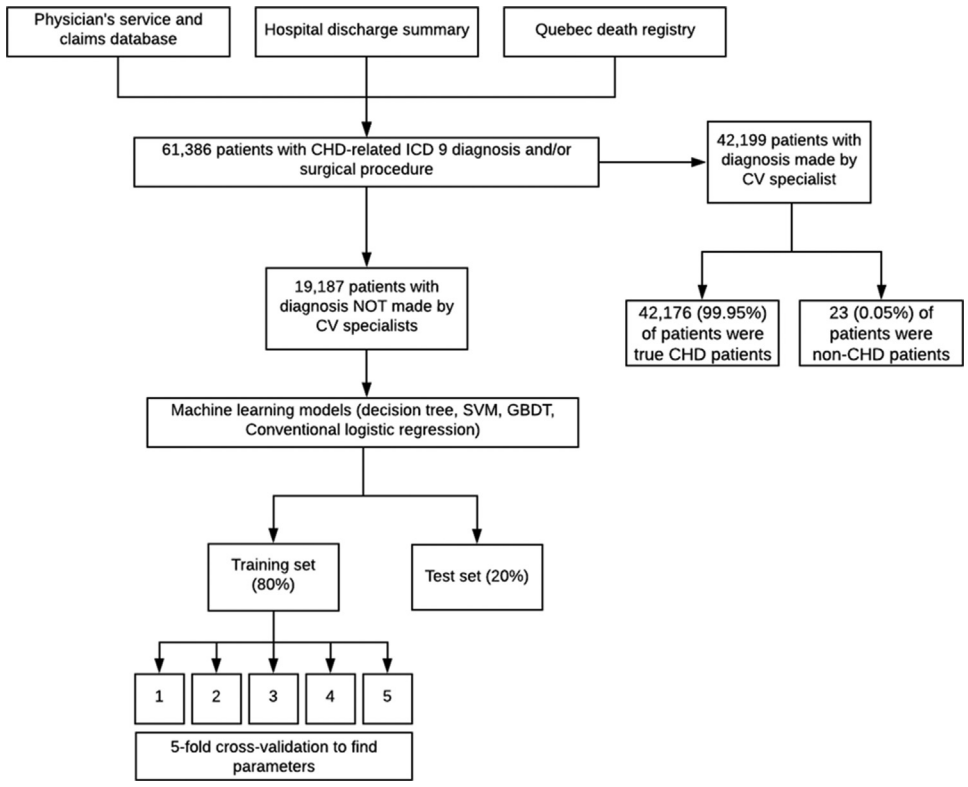
METHODS

DATA SOURCE. In Quebec (Canada) where universal health care is provided, every resident is assigned a unique Medicare number and all health services rendered are systematically recorded in administrative databases until death. In this study, 3 administrative databases were merged by the unique patient Medicare number and used: the physician's services and claims database (Regie de l'Assurance Maladie du Quebec) from l983 to 2000, the hospital discharge summary database (Med-Echo) from l987 to 2000, and the demographics and vital status database from 1983 to 2000. Demographic information included age, sex, and death information. The gold standard labels for CHD patients were generated using our clinician-developed, hierarchical algorithm for CHD classification to derive the Quebec CHD database.[2]

STUDY POPULATION. As illustrated in **Figure 1**, a total of 61,386 Quebec residents were diagnosed with at least 1 CHD between 1983 and 2000. Of them, 42,199 were diagnosed by a cardiovascular (CV) specialist (cardiologist, thoracic surgeon, cardiac surgeon, cardiovascular and thoracic surgeon). We opted to exclude these patients whose diagnosis was coded by a CV specialist because our preliminary analyses showed that having at least 1 CHD-related diagnosis by CV specialist was a strong predictor of correct identification of true CHD patients, and therefore lessened the utility of ML tools. In fact, 99.95% of patients whose CHD diagnosis was made by a CV specialist were true CHD patients, whereas only 19.72% of patients whose diagnosis was made by a non-CV specialist were true CHD patients. Thus, the study population included 19,187 patients with at least 1 CHD-related ICD-9 diagnosis and/or a surgical procedure between 1983 and 2000, whose CHD diagnosis was not made by a CV specialist, but rather by a primary physician (general practitioner, pediatrician, internist) or other physicians (all other fields of practice). **Figure 1** depicts our study design.

OUTCOME. The study outcome was the binary label of CHD diagnosis. Using the clinician-developed algorithm, we generated the gold standard binary labels

FIGURE 1 Study Population Derivation



for the true CHD diagnosis with 1 for true CHD patients and 0 for non-CHD patients. We used these labels to train our ML models on the training set (where the labels were known to each model) and then evaluated these trained models on a separate testing set (where the labels were unknown to the model) to evaluate their diagnostic accuracy. Among the study cohort, 3,784 patients out of 19,184 in total were identified as true CHD patients.

FEATURES. Congenital heart lesions occur during embryonic development and consist of abnormal formations of the heart walls, valves, or blood vessels. They are generally classified into 24 different types.[2] For example, transposition of great artery is 1 type of CHD diagnosed with ICD-9 codes of 745.10, 745.11, 745.12, and 745.19. Types of CHD lesions, CHD-related surgical procedures from which a specific CHD diagnosis could be inferred, source of the CHD diagnosis (outpatient or hospitalization records), and specialty of the physicians who made the diagnosis were also factors affecting the accuracy of CHD diagnoses.[2] To decrease data sparsity and take into account the varying accuracies in diagnosing CHD lesions by ICD codes, procedural codes, data source, and physician specialties, we grouped all relevant ICD codes by the 24 CHD lesion types.[2] Details could be found in **Table 1**. CHD-related surgical procedures were grouped into 4 complexity levels.[18] Physicians were classified and grouped into 1) primary physicians referring to general practitioner, family physician, pediatrician, and internist; and 2) other physicians referring to allergist, pathologist, anesthetist, microbiologist, biochemist, general surgeon, orthopedic surgeon, etc.[18] To decrease data sparsity due to the long follow-up period of the study (up to 18 years) and capture multiple diagnoses/surgical procedures over time, for each type of CHD lesions, we calculated the yearly average number of diagnoses by data source and physician specialty groups, yielding a total of 72 features of which 48 were from outpatient records and rest from hospitalization records. For CHD-related surgical procedures at each complex level, we also calculated the average yearly counts, yielding additional 4 features (**Table 2**).

It is worth noting that using the same database, we have observed a temporal change in the prevalence of CHD diagnoses and rates of congenital and valvular

TABLE 1 Congenital Heart Disease-Related ICD-9 Codes

No.	Disease	ICD-9 Codes
1	Endocardial cushion defect	7456
2	Tetralogy of Fallot	7452
3	Univentricular heart	7453
4	Transposition complex including complete and congenitally corrected	7451
5	Truncus arteriosus	7450
6	Ebstein anomaly	7462
7	Hypoplastic left heart syndrome	7467
8	Atrial septal defect	7455
9	Ventricular septal defect	7454
10	Patent ductus arteriosus	7470
11	Aortic coarctation	7471
12	Unspecified defect of septal closure	7459
13	Anomalies of the pulmonary artery	7473
14	Anomalies of the pulmonary valve	7460
15	Congenital tricuspid valve disease	7461
16	Congenital aortic stenosis	7463
17	Congenital aortic insufficiency	7464
18	Congenital mitral stenosis	7465
19	Congenital mitral insufficiency	7466
20	Anomalies of great veins	7474
21	Other unspecified anomalies of the heart	7469
22	Other unspecified anomalies of the heart	7468
23	Other unspecified anomalies of the aorta	7472
24	Other unspecified anomalies of the circulation	7479

surgical operations over the past decades.[2,7,18] Thus, the inference of using CHD-related ICD and surgical procedures for CHD diagnosis should consider the calendar year of the diagnoses and operations. In light of this, we included patient's year of birth and age at the first record of CHD (ICD code or surgical procedure) as 2 additional features for phenotype CHD. Together with sex, a total of 3 features for demographics were created and included in the models. Given the goal of our study was to compare the performance of different ML methods, all the models used the same set of features, that is, all the 79 features listed in **Table 2**.

CHD-DIAGNOSTIC MODEL CONSTRUCTION. Gradient boosting decision tree (GBDT), support vector machine (SVM), decision tree, and regularized logistic regression were chosen for identifying CHD patients in this study considering several factors including the specific research objectives, the nature of the data set, available computational resources and interpretability. Detailed explanations were included in Supplemental Appendix 1.

For GBDT model, we selected the optimal tree depth and the total number of boosting trees in a grid search that maximize the 5-fold cross-validation performance. If the number of trees increases, the model could fit the data better. However, if the number of trees is too large, the model tends to overfit. The tree depth represents the maximum order of interactions that the model can capture, that is, the model complexity. The number d of splits in each tree, or equivalently the interaction depth, controls the complexity of the boosted ensemble, in which d splits can capture at most the interactions of order d-1.[19] A tree depth of 2 means that there is no interaction between features, while tree depth of 3 suggests interaction between features up to second orders.

SVM is another popular approach for classification by exploiting a linear or nonlinear separation surface in the feature space depending on the user-defined kernels.[19] The performance of SVM classifier largely depends on the kernel.[20] SVM classifiers are tuned using 5-fold cross-validation method and looks for optimal hyperplane as a decision function to classify observations. In the testing stage, the test set was fed into the SVM model for CHD classification.

In parallel to GBDT and SVM models, regularized logistic regression and decision tree models were developed for comparison purpose and to help with results interpretation. Decision trees can be easily visualized due to their inherent structure and representation. The decision tree algorithm recursively partitions the data based on different features and creates a tree-like structure. Each node in the tree corresponds to a specific feature and a threshold value for partitioning the data, and the branches indicate the decision path taken based on the values of the features. The visual representation of a decision tree allows for a clear understanding of the decision-making process within the model. By following the branches from the root node to the leaf nodes, 1 can easily interpret the sequence of decisions and conditions that lead to a particular classification outcome. Additionally, decision trees provide information about the importance and contribution of different features in the decision-making process, as features closer to the root node are deemed to have higher importance. Regularized logistic regression was chosen due to its simplicity and interpretability, whereas decision tree is highly interpretable and could be easily visualized.

EVALUATION. To assess if our sample was sufficient for the proposed ML models, we evaluated the performances of the models with different training samples ranging from a size of 2,000 to 15,375 which was the size of our training sample. The results were shown in Supplemental Figure 1. The figure demonstrated that the Area Under the Precision Recall Curves (AUPRCs) for all the models reached plateaus

after the training sample size equaled 10,000, with some fluctuations around this point. These findings in Supplemental Appendix 2 supported the conclusion that our sample was sufficient for the study.

We split the study cohort randomly into training set with 80% of the sample and a test set with the rest 20%. The training set was further randomly split into 5 folds, and the model trained on 4 of the folds and performance validated against the fifth to establish optimal free parameters for each GBDT (ie, number of trees and tree depth) and different kernels for SVM models (ie, different kernels). The AUPRC from the 5-fold cross-validation was used to choose the optimum parameters.

Selection of cutoff value would influence the results of sensitivity and specificity, as decreasing the cutoff value would increase sensitivity but decrease specificity, and vice versa. A cutoff value of 0.5 was adopted in this study for the classification of patients as CHD or not considering equally the importance of sensitivity and specificity. Furthermore, area under the receiver operating characteristic curve (AUROC) was computed based on the specificity and sensitivity values at all thresholds. We used AUPRC as our metric because our data are unbalanced: there are more negative examples than positive examples. Compared to AUROC, AUPRC focuses on the precision at each recall as opposed to the sensitivity at each false positive rate and therefore more suitable to our application.

In addition to the common statistics for assessing model performance such as AUROC, we added F1 score which combines precision and recall into a single value. The F1 score provides a balanced measure of the model's ability to simultaneously optimize precision (minimizing false positives) and recall (minimizing false negatives). It is widely used in medical research as a robust evaluation metric for imbalanced data sets. Besides AUPRC, we also chose an optimal threshold based on the 5-fold cross-validation in order to obtain F1 score, accuracy, specificity, and sensitivity on the test set. We repeated evaluation 10 times. For each repetition, we used a specific random seed for generating random split of 80/20 training/testing. We then reported the median performance for each method. Details about identifying the final model could be found in Supplemental Appendix 3 (Supplemental Table 1).

The whole study cohort included 19,187 patients with 3,784 true CHD patients. We split the study cohort randomly into a training set with 80% of the patients and a test set with the rest 20%. There were 15,400 patients in the training set with 3,059 true CHDs, while 3,787 patients were included in the test set with 725 true CHDs.

TABLE 2 Features Included in the Development of Machine Learning Model for Identifying Congenital Heart Disease Patients in Large Claims Database

Category (Number of Variables)	Description
Demographics (3)	Patient sex (male/female)
	Year of birth
	Age at the first CHD-related bill or the first CHD-related procedure
Outpatient records (48)	Yearly average number of CHD-related diagnosis by primary physicians[a] and specialists in diagnostic radiology and ultrasonography. One variable for each of the 24 CHD-related diagnoses from Table 1
	Yearly average number of CHD-related diagnosis by other physicians[a]. One variable for each of the 24 CHD-related diagnoses
Inpatient records (24)	Yearly average number of CHD-related diagnoses during hospitalization. One variable for each of the 24 CHD-related diagnoses from Table 1
Surgical procedures (4)	Yearly average number of CHD-related surgery billed by cardiovascular specialists[a]. One variable for each of the 4 different complex level surgeries[b].

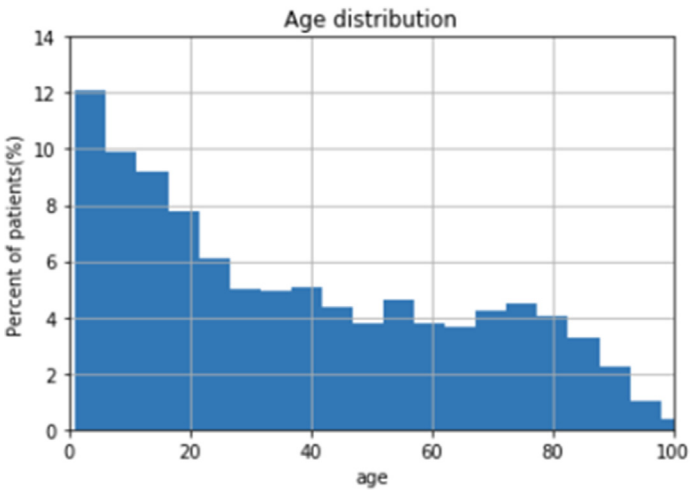
Score = 1.5 to 5.9: complexity level 1. Score = 6 to 7.9: complexity level 2. Score = 8 to 9.9: complexity level 3. Score = 10 to 15: complexity level 4. [a]Primary physicians refer to general practitioner, family physician, pediatrician, and internist. Other physicians refer to allergist, pathologist, anesthetist, microbiologist, biochemist, general surgeon, orthopedic surgeon, etc. [b]Definition of complex levels was defined based on Mean Aristotle Basic Score.[18]

CHD = congenital heart disease.

L1 LASSO PENALTY. Although linear model is limited to discovery only the additive effects of the feature predictors, it has better interpretability than the tree-based and kernel-based models described above. This advantage is important in understanding the contribution of each feature.

Least Absolute Shrinkage and Selection Operator (Lasso) is a regularization technique in ML for linear regression. The purpose of Lasso is to achieve a parsimonious model by shrinking the coefficients of less important features to zero, effectively removing them from the model.[19] Features with coefficients closer to zero are regarded as less important. To this end, we adopted the L1 Lasso regularization method to study the potential of representing the data with fewer predictors by penalizing the absolute sum of predictors' weights used in the model (ie, the L1-norm) along with the cross entropy of the difference between the predicted CHD risk and the true CHD label. Because the gradient of each predictor coefficient is defined unless it is zero, using regularized logistic regression along with Lasso regularization could shrink the coefficients of the unimportant variables to be exactly zero and therefore reduce the complexity of the model. In this way, the Lasso performs simultaneous feature selection and model estimation of the original database.[21] We used the scikit-learn implementation of Lasso. To determine the optimal Lasso regularization parameter, often referred to as the tuning parameter, we employed a cross-validation. We evaluated the performance of

FIGURE 2 Age Distribution of the Study Cohort



the Lasso model across a range of regularization values and selected the value that yielded the best performance based on our evaluation metric (ie., AUPRC) procedure (results shown in Supplemental Appendix 4 and Supplemental Table 2).

EXTERNAL VALIDATION. We performed external validation on the developed GBDT model using an updated data set (1983-2010) from the same sources, that is, Quebec administrative databases of outpatient data, hospitalization records, and vital statistics. Subjects that were included in the training and test sets for developing the GBDT model were excluded from external validation.

IMPLEMENTATION. We performed all statistical analyses using Python version 3.6, Python-based scikit-learn package[22] and visualized using matplotlib.[23]

RESULTS

Age distribution of study population is shown in Figure 2. Consistent with what has been reported in the literature, the majority of the CHD population were adults.

OPTIMUM CHD DIAGNOSIS MODEL CONSTRUCTION. We adopted 5-fold cross-validation method for evaluation. Possible overfitting was assessed by comparing the prediction accuracy between training and validation sets. We repeated the evaluation 10 times, each time with a different random split of 80/20 training/testing. The median performance and interquartile range for each method was included in Table 3. As the interquartile range was very low suggesting minimal variations, we opted to select the model demonstrating the median performance for test data set as the final model.

Based on 5-fold cross-validation, the best tree depth and tree number for GBDT model are 6 and 195, respectively, which constitute the optimum GBDT model. This means there are up to fifth order of interactions between the features. This indicates the nonlinearity of the data as well as the capability of GBDT handling different types of features and capturing interactions among features.

The optimal parameters for SVM model, Radial Basic Function kernel was selected over the linear kernel, indicating nonlinearity in the feature space. These results were consistent to the results from the optimum GBDT model which suggested high-order interactions between the features.

Figure 3 shows the first 3 features about how the Decision Tree model classified CHD patients. The number of ventricular septal defect (VSD) by primary physicians was the most important feature. If a patient had no diagnosis of VSD made by primary physicians, the chance of non-CHD was 85.4%. The second and third most important features were the number of complex level 3 surgeries by CV specialists and number of atrial septal defect by primary physicians. Figures 4 and 5 show AUROC and AUPRC, respectively. While all of the ML models demonstrated good performance, GBDT clearly leads the performance across all of the 5 metrics including AUPRC, AUROC, F1 score, accuracy, sensitivity, and specificity.

We further assessed the performance of the 4 models using the Precision-Recall Plot. Figure 5 shows pairs of recall and precision values. As seen from the plot, the GBDT model outperforms all the other 3 models with nearly perfect precision and recall values at all the thresholds. The largest difference between any 2 models is GBDT and regularized logistic regression with 0.993 and 0.936 AUPRC for the testing set, respectively (ie, 5.7% improvement). This suggests that some nontrivial features interactions can only be captured by the more sophisticated GBDT. AUROC in Figure 4 also showed that GBDT outperformed the other 3 models.

FEATURE IMPORTANCE. After training the GBDT model, we could acquire the feature importance of each feature, which is computed as the total reduction of the criterion brought by that feature. A higher value indicates a more important feature.

We performed 10 experiments each with a random split of training and testing sets. We obtained consistent features among the top 10 most important features across these experiments although there

TABLE 3 Comparison Between GBDT, SVM, Decision Tree, and Regularized Logistic Regression in Model Diagnostic Performance (Median Values From 10 Repetitions With Various Random Seeds)

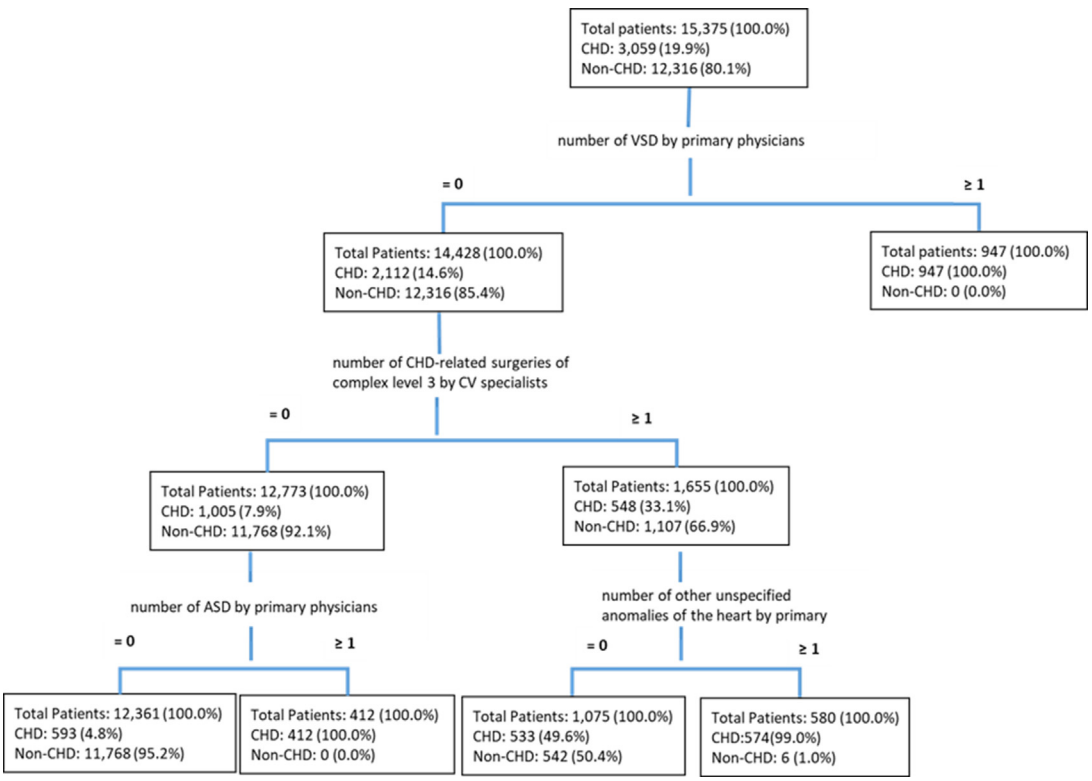
AUPRC	Regularized Logistic Regression		Decision Tree		SVM		GBDT	
Run	Train	Test	Train	Test	Train	Test	Train	Test
1	0.943	0.936	0.942	0.943	0.986	0.962	0.999	0.997
2	0.950	0.926	0.941	0.944	0.998	0.966	1.000	0.982
3	0.942	0.936	0.943	0.939	0.951	0.950	0.999	0.994
4	0.944	0.936	0.940	0.949	0.992	0.971	0.999	0.995
5	0.947	0.936	0.945	0.930	0.996	0.932	0.999	0.992
6	0.945	0.933	0.941	0.946	0.988	0.951	0.999	0.987
7	0.945	0.927	0.941	0.944	0.988	0.958	1.000	0.993
8	0.945	0.933	0.942	0.941	0.988	0.952	1.000	0.997
9	0.945	0.920	0.943	0.938	0.991	0.960	1.000	0.990
10	0.946	0.946	0.943	0.940	0.989	0.956	0.999	0.998
Median	0.945	0.935	0.942	0.942	0.989	0.957	0.999	0.994
Q1	0.94425	0.9285	0.941	0.9393	0.988	0.95125	0.999	0.9905
Q3	0.94575	0.936	0.943	0.944	0.99175	0.9615	1.000	0.9965
IQR	0.0015	0.0075	0.002	0.0048	0.00375	0.01025	0.001	0.006

AUPRC = area under the precision recall curve; GBDT = gradient boosting decision tree; SVM = support vector machine.

were slight variations in their ranks between experiments.

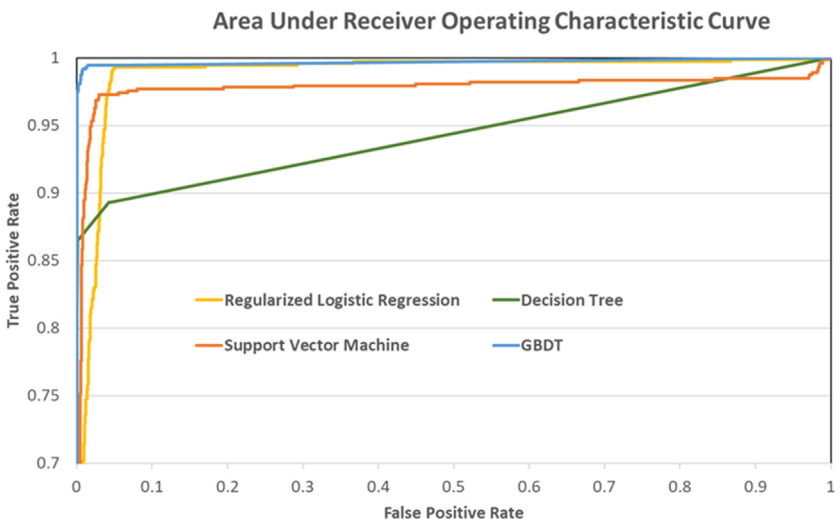
Top 10 features in 1 run extracted from the acquired GBDT model is shown in Figure 6. Number of VSD diagnosis by primary physicians was the most important feature, followed by number of complex level 3 surgeries by CV specialists, and number of atrial septal defect diagnosis by primary physicians,

FIGURE 3 Visualization of Decision Trees in Classifying CHD Patients



CHD = congenital heart disease.

FIGURE 4 AUROC Plots of the 4 Constructed Models for CHD Diagnosis



AUROC = area under the receiver operating characteristic curve; CHD = congenital heart disease; GBDT = gradient boosting decision tree.

consistent with the top 3 features identified by decision tree.

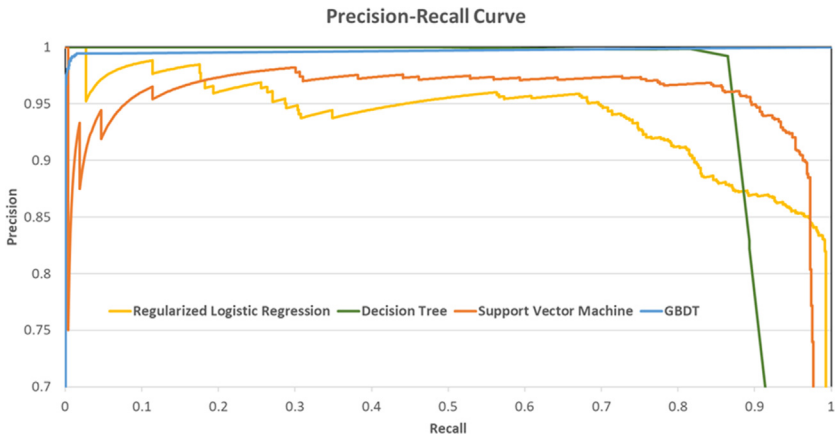
Furthermore, the LASSO coefficients for 90% of variables were nonzero, indicating that most of the variables play a role in computing the CHD probability. Details about Lasso diagnostics could be found in Supplemental Appendix 5 (Supplemental Figure 2).

EXTERNAL VALIDATION. The data set for external validation included 68,192 patients with at least 1 CHD-related ICD-9 diagnosis and/or a surgical procedure between 1983 and 2010, whose CHD diagnosis was not made by a CV specialist. The model showed excellent performance: accuracy = 0.993, F1 score = 0.990, sensitivity = 0.978, and specificity = 0.998.

DISCUSSION

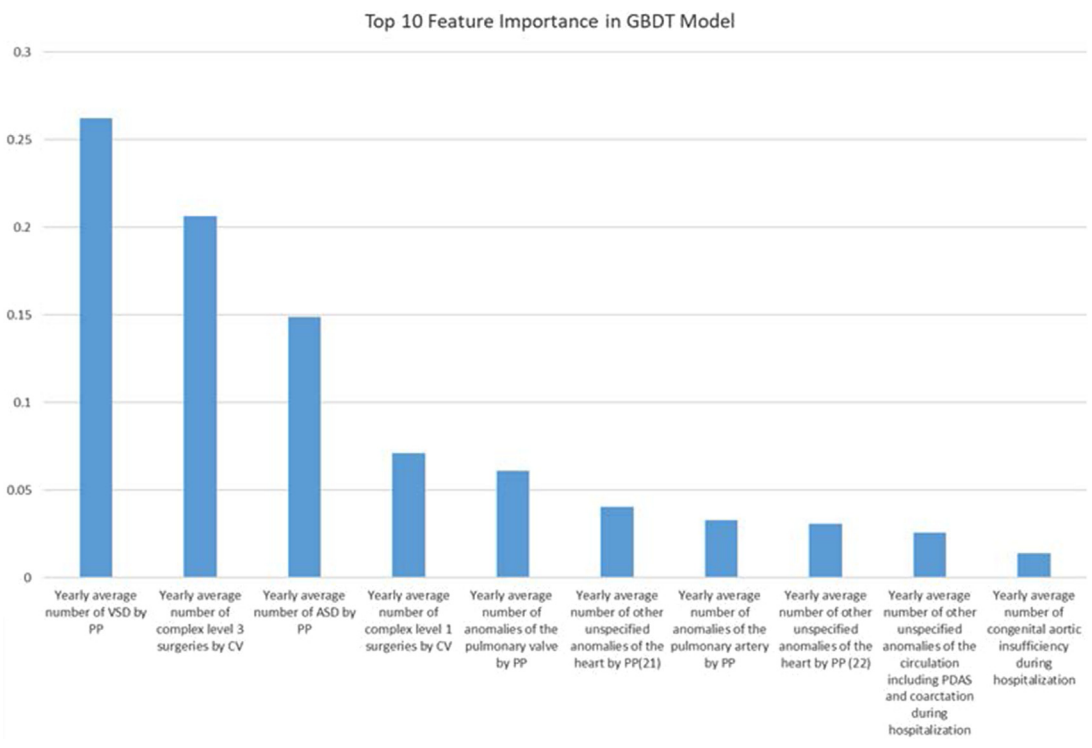
This is the first study to show that ML models especially GBDT can automate the clinician designed

FIGURE 5 Precision-Recall Plots of the 4 Constructed Models for CHD Diagnosis



CHD = congenital heart disease; GBDT = gradient boosting decision tree.

FIGURE 6 Selected Top 10 Features in the Optimum GBDT Model for CHD Diagnosis



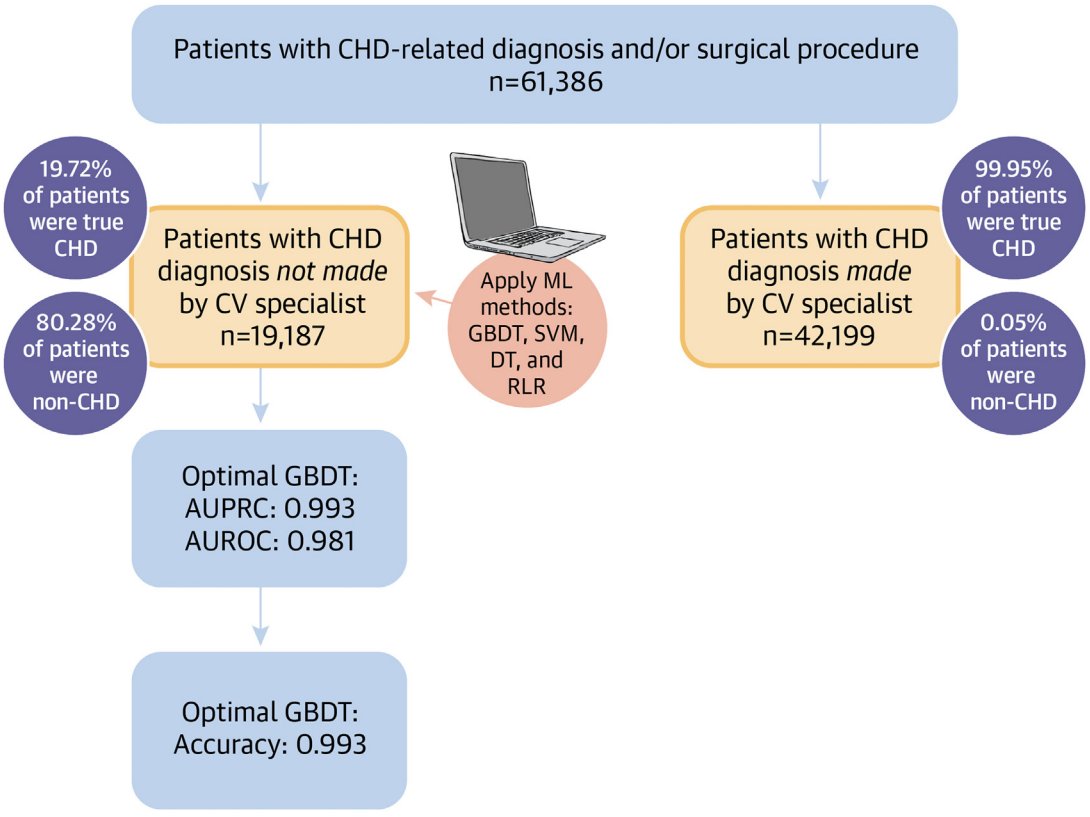
CHD = congenital heart disease; GBDT = gradient boosting decision tree.

algorithm to identify true CHD patients in large claims database. The high precision and recall suggest that our approach could capture the hierarchical regularities that were derived based on long-term expertise of clinicians (**Central Illustration**).

CHD diagnosis is challenging and requires specialist's manual audits to increase accuracy of claims database analyses. To facilitate this process, we propose a computational pipeline that leverages ML models to identify CHD patients based on a large claim database containing over 19,000 patients. To this end, we evaluated several popular ML methods including the regularized logistic regression model and more sophisticated tree-based and kernel-based methods. Compared to regularized logistic regression, GBDT and SVM could identify true CHD patients from large administrative databases with excellent diagnostic performance. GBDT model demonstrated the highest AUPRC around 0.99 for the testing sets. Lasso regularization method indicated the importance of majority of the features we used in the CHD identification. Our data set is imbalanced, with a higher number of negative examples (non-CHD patients) compared to positive examples (true CHD patients). The class proportion reflects the population prevalence. Manually down-sample the negative examples to match the positive examples will lead to a biased model with the prior probability at 50% CHD prevalence. Therefore, it is desirable to have a model with high precision at the same recall rate. To this end, we used AUPRC as our metric, which focuses on the precision at each recall as opposed to the recall (or sensitivity at each false positive rate [ie, ROC]).

The near-perfect AUPRC suggests that our pipeline can capture the hierarchical regularities that were derived based on expertise of clinicians. The excellent results from external validation further prove a great generalizability of our ML approach in predicting CHD patients. We envision that our approach can be employed in other claims database for identifying CHD patients. The results presentation was limited by the fact that the feature importance from the GBDT model was from 1 rather than all the 10 runs. However, given the fact that the top 10 most important

CENTRAL ILLUSTRATION **Pipeline for CHD Identification Using Machine Learning Methods in Large Claims Data Source**



Marelli AJ, et al. JACC Adv. 2024;3(2):100801.

AUPRC = area under the precision recall curve; AUROC = area under the receiver operating characteristic curve; CHD = congenital heart disease; CV = cardiovascular; DT = decision tree; GBDT = gradient boosting decision tree; RLR = regularized logistic regression; SVM = support vector machine.

features were consistent across the 10 runs with minor variations in ranking, the presented feature importance was informative.

The result of our study is in line with many prior studies demonstrating the predictive superiority of ML over more traditional statistical models. For example, ML was better at predicting heart failure readmissions and cardiovascular events[24,25] than well-established risk assessment scores. In the realm of CHD, ML models were developed to stratify patients' risks and gauge their prognosis, as well as to forecast the level of risk associated with congenital heart surgery.

The relevance of our study is reflected in the exponential increase in the use of administrative databases in the past decade. They also facilitate investigation of regional variations and favor large-scale collaborative initiatives.

STUDY STRENGTHS. Strengths of our study include the characteristics of the databases, the rigorous internal validation previously done and the potential for generalizability. First, the database source is administrative database. In a country with universal access to health care and where physician's renumeration is claims-based, the database is inclusive and exhaustive. Moreover, the manual audit done on around one-third of the population by 2 specialized clinicians contributes to the robustness of the internal validity of the database. Finally, the features kept after regularization of the data allow for the potential to refine the standardized electronic health records coding system (eg, ICD-10) and ultimately the generalization of this model to other databases.

STUDY LIMITATIONS. We note some limitations of the current study that highlight opportunities for

future development. Firstly, there is a need to explore whether the true CHD patients with no CHD-related diagnosis from CV specialists were ever seen by a CV specialist and not given a diagnosis of CHD; or were never seen and therefore not given a CHD diagnosis. The former case suggests that our ML tool would decrease false negative diagnoses among CV specialist whereas the latter case suggests that it would serve to signal which patients to properly refer to a CV specialist for follow-up. Secondly, significant unrecognized predictors contributing to CHD classification may exist. Thirdly, despite the use of regularization methods to decrease overfitting and the absence of overfitting supported by the results, our models need to be reproduced on other CHD databases to test its external validity. Fourthly, our study used data from claims data sources. This is not immune to the potential flaws that are commonly seen in billing records motivating the need for models that enhance diagnostic accuracy. Fifthly, we aggregated patients' medical data across their records and trained the ML models by treating each aggregated record as an independent training example. There are alternative computational strategies that directly model the longitudinal electronic health records data using recurrent neural network with long short-term memory architecture[26,27] that may provide further insights into the extent of how well a ML method can perform on the CHD phenotyping task with longitudinal data. Limited by the scope of this article, we will leave model extension as well as the comparison with the physician-developed method as future work.

CONCLUSIONS

Using an administrative database of patients with a CHD diagnosis from which CHD patients were previously manually extracted, we showed that ML models could identify the true CHD patients whose diagnosis was made by a non-CV physician with much higher accuracy than regularized logistic regression. The GBDT model with the 79 carefully selected features showed robustness and effectiveness in identifying CHD cases. As the model is easy to be constructed and the features are commonly found in claims database, it provides significant potential for practical implementation. We also described the relative importance of each predictor. These results are promising as more accurate diagnosis will ultimately lead to better characterization of CHD thereby improving the care of CHD patients across the lifespan.

ACKNOWLEDGMENTS The authors acknowledge the staff at the Régie de l'assurance maladie du Québec for their helpfulness in assisting us with the laborious and meticulous extractions required for the reliable raw data to be sent to the MAUDE Unit (McGill Adult Unit for Congenital Heart Disease Excellence) for the creation of the Quebec CHD database.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

The study was supported by the Canadian Institutes of Health Research Foundation Grant (#148462) and Heart and Stroke Foundation Grant-In-Aid (G-21-0031574) awarded to Dr Marelli. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Ariane J. Marelli, Department of Cardiology, McGill Adult Unit for Congenital Heart Disease, McGill University Health Centre, McGill University, D055108, 1001 Decarie Boulevard, Montreal, H4A 3J1 Quebec, Canada. E-mail: ariane.marelli@mcgill.ca.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE: Machine learning models can be applied to large claim databases for efficient disease classification.

TRANSLATIONAL OUTLOOK 1: Machine learning models could be used to automate manual classification system for congenital heart disease in claims database.

TRANSLATIONAL OUTLOOK 2: Machine learning models could be generalized for classification of other complex diseases in administrative databases.

REFERENCES

1. Beland MJ, Harris KC, Marelli AJ, Houyel L, Bailliard F, Dallaire F. Improving quality of congenital heart disease research in Canada: standardizing nomenclature across Canada. *Can J Cardiol*. 2018;34:1674–1676.

2. Marelli AJ, Mackie AS, Ionescu-Ittu R, Rahme E, Pilote L. Congenital heart disease in the general population: changing prevalence and age distribution. *Circulation*. 2007;115:163–172.

3. Beausejour Ladouceur V, Lawler PR, Gurvitz M, et al. Exposure to low-dose Ionizing radiation from cardiac procedures in patients with congenital heart disease: 15-year data from a population-based longitudinal cohort. *Circulation*. 2016;133:12–20.

4. Bouchardy J, Therrien J, Pilote L, et al. Atrial arrhythmias in adults with congenital heart disease. *Circulation*. 2009;120:1679–1686.

5. Cohen S, Liu A, Gurvitz M, et al. Exposure to low-dose Ionizing radiation from cardiac procedures and malignancy risk in adults with congenital heart disease. *Circulation*. 2018;137:1334–1345.

6. Lanz J, Brophy JM, Therrien J, Kaouache M, Guo L, Marelli AJ. Stroke in adults with congenital heart disease: incidence, cumulative risk, and predictors. *Circulation*. 2015;132:2385–2394.

7. Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. *Circulation*. 2014;130:749–756.

8. Mylotte D, Pilote L, Ionescu-Ittu R, et al. Specialized adult congenital heart disease care: the impact of policy on mortality. *Circulation*. 2014;129:1804–1812.

9. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:1–10.

10. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.

11. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res*. 2017;121:1092–1101.

12. Miao KH, Miao JH. Coronary heart disease diagnosis using deep neural networks. *Int J Adv Comput Sci Appl*. 2018;9:1–8.

13. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc*. 2013;20(e2):e206–e211.

14. Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22:993–1000.

15. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal lab tests. *arXiv*. 2016. https://doi.org/10.48550/arXiv.1608.00647

16. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. *PLoS Med*. 2018;15:e1002695.

17. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13:395–405.

18. Ionescu-Ittu R, Mackie AS, Abrahamowicz M, et al. Valvular operations in patients with congenital heart disease: increasing rates from 1988 to 2005. *Ann Thorac Surg*. 2010;90:1563–1569.

19. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Springer; 2013.

20. Smola AJ, Schölkopf B, Müller K-R. The connection between regularization operators and support vector kernels. *Neural Netw*. 1998;11:637–649.

21. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*. Springer; 2001.

22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.

23. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9:90–95.

24. Shameer K, Johnson KW, Yahi A, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. *Pac Symp Biocomput*. 2017;22:276–287.

25. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;9:629–640.

26. Harutyunyan H, Khachatrian H, Kale DC, Galstyan A. Multitask learning and benchmarking with clinical time series data. *arXiv*. 2017. https://doi.org/10.1038/s41597-019-0103-9

27. Zhao J, Feng Q, Wu P, et al. Learning from longitudinal data in electronic health record and genetic data to Improve cardiovascular event prediction. *Sci Rep*. 2019;9:717.

KEY WORDS congenital heart disease, large administrative claims database, machine learning

APPENDIX For a supplemental appendix, tables, and figures, please see the online version of this paper.