HHS Public Access

Author manuscript

Birth Defects Res. Author manuscript; available in PMC 2024 November 01.

Published in final edited form as:

Birth Defects Res. 2023 November 01; 115(18): 1693–1707. doi:10.1002/bdr2.2245.

A machine learning model for predicting congenital heart defects from administrative data

Haoming Shi[1], **Wendy Book**[2,3], **Cheryl Raskind-Hood**[3], **Karrie F. Downing**[4], **Sherry L. Farr**[4], **Mary N. Bell**[1], **Reza Sameni**[5], **Fred H. Rodriguez III**[2,6], **Rishikesan Kamaleswaran**[1,5]

[1]Department of Biomedical Engineering, Georgia Institute Technology, Atlanta, Georgia, USA

[2]Division of Cardiology, Emory University School of Medicine, Atlanta, Georgia, USA

[3]Department of Epidemiology, Emory University, Rollins School of Public Health, Atlanta, Georgia, USA

[4]National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[5]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, USA

[6]Children's Healthcare of Atlanta, Atlanta, Georgia, USA

Abstract

Introduction: International Classification of Diseases (ICD) codes recorded in administrative data are often used to identify congenital heart defects (CHD). However, these codes may inaccurately identify true positive (TP) CHD individuals. CHD surveillance could be strengthened by accurate CHD identification in administrative records using machine learning (ML) algorithms.

Methods: To identify features relevant to accurate CHD identification, traditional ML models were applied to a validated dataset of 779 patients; encounter level data, including ICD-9-CM and CPT codes, from 2011 to 2013 at four US sites were utilized. Five-fold cross-validation determined overlapping important features that best predicted TP CHD individuals. Median values and 95% confidence intervals (CIs) of area under the receiver operating curve, positive predictive value (PPV), negative predictive value, sensitivity, specificity, and F1-score were compared across four ML models: Logistic Regression, Gaussian Naive Bayes, Random Forest, and eXtreme Gradient Boosting (XGBoost).

Correspondence Haoming Shi, Department of Biomedical, Engineering, Georgia Institute, Technology, Atlanta, GA, USA. haoming.shi@emory.edu.

AUTHOR CONTRIBUTIONS

Haoming Shi: investigation, analysis, writing original draft and review and editing. **Wendy M. Book**: conceptualization, data curation, review and editing, preparation, supervision, project administration, funding acquisition. **Cheryl Raskind-Hood**: data curation, preparation, supervision, analysis, writing original draft, review and editing. **Karrie Downing**: data curation, review and editing. **Sherry Farr**: conceptualization, data curation, review and editing. **Mary N. Bell**: conceptualization, methodology, investigation, analysis. **Reza Sameni**: methodology, investigation, preparation, supervision, writing original draft, review and editing, project administration. **Fred H. Rodriquez III**: conceptualization, methodology, review and editing, project administration. **Rishikesan Kamaleswaran**: conceptualization, methodology, investigation, data curation, writing original draft and review and editing, preparation, supervision, project administration, funding acquisition.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

Results: Baseline PPV was 76.5% from expert clinician validation of ICD-9-CM CHD-related codes. Feature selection for ML decreased 7138 features to 10 that best predicted TP CHD cases. During training and testing, XGBoost performed the best in median accuracy (F1-score) and PPV, 0.84 (95% CI: 0.76, 0.91) and 0.94 (95% CI: 0.91, 0.96), respectively. When applied to the entire dataset, XGBoost revealed a median PPV of 0.94 (95% CI: 0.94, 0.95).

Conclusions: Applying ML algorithms improved the accuracy of identifying TP CHD cases in comparison to ICD codes alone. Use of this technique to identify CHD cases would improve generalizability of results obtained from large datasets to the CHD patient population, enhancing public health surveillance efforts.

Keywords

congenital heart disease; machine learning; population health

1 | INTRODUCTION

Congenital heart defects (CHD) are the most common birth defect, with a prevalence of 80 per 1000 live births, causing 25% of infant mortality in developed countries (Botto & Correa, 2003; Gilboa et al., 2016; Marelli et al., 2014; Müller et al., 2022; Oster et al., 2013; Warnes et al., 2001; Yang et al., 2002). Despite improved survival of individuals with CHD and increasing prevalence of CHD across the lifespan (Gilboa et al., 2016), individuals with CHD still experience significant late morbidity and premature mortality (Müller et al., 2022). Surveillance of this population is constrained by difficulties identifying patients not included in birth defect registries (e.g., because they were born outside of a catchment area or diagnosed too late) and limited longitudinal surveillance of those who are in birth defect registries (Liberman et al., 2023; Massin & Dessy, 2006). Nevertheless, surveillance is important because individuals with CHD can experience a wide spectrum of long-term health outcomes. Compared to the general population, those with CHD have increased risk of developing cardiovascular comorbidities including atrial fibrillation, hypertension, heart failure, and other cardiac-related conditions (Billett et al., 2008). However, CHD is not a homogeneous disease, even among those with the same anatomic defect; thus long-term health outcomes are affected by many factors, including defect anatomy, the type of repair (if applicable), the number and nature of interventions, age, social determinants of health, access to care, and the therapeutic plan (Bhatt et al., 2015; Brida & Gatzoulis, 2019; Stout et al., 2019).

CHD surveillance on a population level often relies upon International Classification of Diseases, Ninth and Tenth Revision, Clinical Modification codes (ICD-9-CM and ICD-10-CM) in large administrative and clinical datasets to estimate prevalence of CHD, healthcare utilization, and various health outcomes, yet this methodology may not accurately identify the population of interest (Mylotte et al., 2014). Some CHD-related ICD-9-CM codes have high false positive (FP) rates and cannot sufficiently distinguish individuals with a true positive (TP) CHD from those who do not have a CHD (Agarwal et al., 2016; Broberg et al., 2015; Khan et al., 2018; Rodriguez et al., 2018). ICD-9-CM code 745.5 that identifies secundum atrial septal defect (ASD) and patent foramen ovale has a FP for CHD as high as 76% in an adult population (Rodriguez et al., 2018). High FP rates have also been shown

for subaortic stenosis and pulmonary valve stenosis (Khan et al., 2018) and the FP rate for ICD-9-CM codes for shunt lesions has been reported at 50% (Broberg et al., 2015). Though restricting datasets to patients with severe CHD codes can improve the FP rate, it also limits the understanding of outcomes in patients with non-severe defects such as bicuspid aortic valves (Udholm et al., 2019; Wallace et al., 2022), that can still cause considerable morbidity.

Machine learning methods can offer an alternative solution to improve TP CHD case detection in administrative and clinical data without restricting to the most severe cases. Using data from combined clinical and administrative sources, the current study aimed to develop and test algorithms that improve accuracy of detecting CHD using various machine learning models (Logistic Regression [LR], Gaussian Naive Bayes [GaussianNB], Random Forest [RF], and XGBoost), and identify the most salient features that aid in the accurate detection of a CHD.

2 | METHODS

2.1 | Data source

A de-identified validated dataset was sourced from the Centers for Disease Control and Prevention's (CDC's) Surveillance of Congenital Heart Disease Across the Lifespan project (CDC RFA DD15-1506). The dataset combined clinical and administrative data for patients suspected of having CHD over 3 years, 2011–2013, from four sites: Georgia (GA), North Carolina (NC), New York (NY), and Utah (UT). For CHD diagnosis validation through chart review, 200 cases were selected from each of four sites for a planned total of 800 cases. The total number of cases was based on the feasibility for chart abstraction and validation at the sites. Cases were selected randomly, while ensuring an approximately even distribution across four mutually exclusively CHD ICD code groups (severe, shunt, valve, and other) and, within those, an approximately even distribution by age groups. Anatomic groups were defined by a multi-site clinician group based on ICD-9-CM CHD codes. Severe CHD was defined as CHD that typically requires surgery in the first year of life to permit survival. For GA, NC, and NY, the age groups were 1–10-year-olds, 11–19-year-olds, 20–64-year-olds, and >64-year-olds, while ages for UT were 11–19-year-olds and 20–64-year-olds. Further methodological details can be found in Rodriguez et al. (2022). During the study period, all contributing healthcare systems utilized ICD-9-CM codes. Patients having at least 1 of 55 ICD-9-CM codes for a CHD were reviewed; all patients had encounter level data spanning 3 years, including all associated ICD-9-CM and Current Procedural Terminology (CPT) codes. Those with only a 745.5 code in isolation or in conjunction with other non-specific CHD ICD-9-CM codes were omitted from the original dataset due to known poor PPV of this code for secundum atrial septal defect (Rodriguez et al., 2018). From this validation dataset, features (predictive variables) were identified and used by machine learning algorithms to identify TP CHD cases. Race and ethnicity were self-reported in the contributing healthcare system data sources. Race and ethnicity data were collected to understand applicability of developed models to a population. The final data set available for analysis had five additional cases not included in Rodriguez et al. (2022).

2.2 | Exclusions

Of the 800 person planned cohort, a total of 21 patients were excluded leaving a total of 779 patients in the validated analytic dataset: 15 cases did not meet inclusion criteria (i.e., 12 cases had ICD-9-CM code 745.5 in isolation and three cases only had a fetal echocardiogram performed); three cases did not have any clinical data to review for validation; and three cases were inadvertently reviewed twice.

2.3 | Feature generation

A variety of features (independent variables), totaling 29,693, for each patient were initially created by summarizing demographics, healthcare encounter types, ICD-9-CM codes, and CPT codes across all encounters. Discrepant values across data sources were reconciled, and healthcare utilization variables were summarized to the patient level by counting the number of encounters of a given type the patient had over the three-year surveillance window from 2011 to 2013, using days as the primary measure. To tabulate counts of diagnoses and symptoms as well as comorbidities and complications, several existing diagnostic classification schemes were applied. For instance, the Healthcare Cost and Utilization Project's (HCUP) Clinical Classification Software (CCS), a categorization scheme that collapses over 15,000 ICD-9-CM and CPT codes into 259 diagnostic categories, was used. The best categorization scheme for patient diagnoses and procedures was not determined a priori; instead, all considered schemes and resulting features were included in the dataset for training, testing, and algorithm development to determine which groupings were most helpful in identifying patients with CHD. Fully zero-features (i.e., no one had that feature) and features related to geographic location were removed from the dataset; this included 15,638 ICD-9-CM codes, 76 CCS categorical codes, and 6821 CPT codes that did not appear on any encounters over the three-year study period as well as 20 demographic features. After excluding a total of 22,555 fully zero-features and geographic location features, the final analytic dataset contained 7138 features, including nine demographics, 3200 ICD-9-CM codes, 15 health encounter types, 1056 CCS categorical codes, and 2858 CPT codes, and was de-identified for machine learning training and validation.

2.4 | Feature selection

Feature selection was applied to identify the subset of the most relevant features from the analytic set of 7138 features by removing redundant and irrelevant features, which allowed those that best predicted a true CHD to be retained. First, a random search method and a five-fold cross-validation that split the whole dataset into five non-overlapping splits and trained on 4 splits, tested on 1 split then repeated for five times such that each split was included in the train and test datasets were conducted using the pooled data for all four sites to optimize XGBoost hyperparameters of the 7138 features. Then, XGBoost was applied on each site's data independently to evaluate feature importance. Specifically, the leave-one-site-out strategy, which used three sites for model training and retained the remaining site for testing, was utilized. Features with less than 1% importance contribution score from XGBoost feature importance evaluations were removed for that site; feature importance scores were generated through gradient boosting after boosted trees were constructed, and they are indicative of how useful a given feature is within the model. Remaining features

were compared across sites and all features that overlapped for 2 or more sites (n = 10 features) were selected and retained for algorithm development. Mean Shapley Additive exPlanation (SHAP) values for these 10 features were generated using the pooled data from all four sites and ranked by relative magnitude to show the impact of each selected feature on CHD prediction (Lundberg & Lee, 2017). This same set of 10 features was used for all four machine learning models (LR, GaussianNB, RF, and XGBoost) during algorithm development.

2.5 | Algorithm development

2.5.1 | Model development and cross-validation—

The dataset was split using a five-fold cross validation approach (illustrated in Figure 1) to demonstrate generalizability of the learning algorithms, which split the dataset into five non-overlapping splits, took one unique split as test set and remaining splits as train set before repeated for five times such that each split was included in the train and test datasets. Since the available data was from four different sites, the four selected machine learning models were applied on each site's data independently to evaluate the performance of each model, again using the leave-one-site-out strategy with three sites for model training and the remaining site for testing. Five-fold cross validation was conducted on data from three sites for training the dataset to optimize the model parameters resulting in a less biased model by outliers, and then tested on the remaining site to estimate performance metrics including AUROC, PPV, NPV, sensitivity, specificity, and F1-score; this five-fold cross validation step was conducted a total of four times such that each site was left out once as the test dataset. The model operating point was selected when applying the algorithm to datasets to prioritize either PPV or to minimize false negative cases. By generating a plot of PPV versus false negative rate evaluated at different models, the trade-off between increasing in PPV and false negative classification can be visualized. In addition to assessing algorithm generalizability to distinguish between TP CHD and false positive (FP) CHD cases from datasets unseen (other sites), this strategy elucidated the statistical similarities and differences between the TP CHD and FP CHD populations from the studied sites. Medians and corresponding 95% confidence interval (CI) values were used to summarize the performance metrics across the five-folds for each of the four sites (i.e., 20 sets of performance metrics) for each model.

In addition, we evaluated the best performing machine learning model, XGBoost, on site-specific data. The leave-one-site-out strategy was again adopted for each site, in which five-fold cross validation was used for training three sites with the XGBoost model and the remaining fourth site was used for testing. This strategy ensured a reliable and unbiased estimate of performance between the four sites. A visual on the process for testing inter-site capacity of the XGBoost model can be seen in the model development diagram (Figure 1).

3 | RESULTS

3.1 | Sample demographics

Table 1 displays the demographic characteristics of the analytic sample, overall and by CHD classification. A total of 779 patients across four sites comprised the analytic sample with a PPV of 76.5% (596/779) based on ICD-9-CM code classification. Overall, 48.8% were

male; the distribution of sex among TP and FP was similar in the bivariate analysis ($p = .29$). Mean age of the sample was 30.97 years (standard deviation [SD] ± 25.02 years). However, those with a FP CHD were significantly older compared to those with a TP CHD, $\bar{X} = 43.86$ years (SD ± 27.64 years) versus $\bar{X}\ X = 27.01$ years (SD ± 22.76 years), respectively ($t = 8.31$, $p < .0001$). The majority of the sample were White ($n = 522$; 67.0% overall, 78% excluding those with unknown race), non-Hispanic ($n = 547$; 70.2% overall, 86% excluding those with unknown ethnicity), and covered by public health insurance ($n = 429$; 55.1% overall, 60% excluding those with unknown insurance type).

3.2 | Feature importance

Mean absolute SHAP values for the 10 features with the most influence on CHD prediction across data from all sites are presented in Figure 2; this bar plot depicts how much each selected feature contributes to the prediction of CHD, ordered from most influential to least (Explaining Machine Learning Models: A Non-Technical Guide to Interpreting SHAP Analyses, 2021). Number of outpatient healthcare encounters with at least one documented CHD code (hereafter referred to as “CHD-coded”) had the highest mean absolute SHAP value, with a magnitude of 0.5. Other relevant features contributing to CHD positive or negative classification prediction included: having diagnosis codes for a CHD in the “other” anatomic complexity group (as defined in Appendix A); number of CHD-coded healthcare encounters overall; age as continuous variable (older age less predictive of CHD); having ICD-9-CM diagnosis codes belonging to the CCS categories for musculoskeletal system and connective tissue, circulatory diseases, and respiratory system comorbidity groups; having a documented electrocardiogram; and number of emergency department (ED) visits.

Another visualization of the 10 features is displayed in a SHAP summary plot (Figure 3), with features seen along the *y*-axis and SHAP values displayed along the *x*-axis. Each dot represents the SHAP value of that feature for one of the 779 individuals in the dataset. SHAP values closer to zero suggest that that instance contributes little to the prediction of a TP or FP, SHAP values closer to one suggest greater influence toward the prediction of a TP, and SHAP values closer to negative one suggest greater influence toward the prediction of a FP. The blue to red color range represents the value of the feature for that person. For example, a case with 0 CHD-coded outpatient healthcare encounters is represented by a blue dot, a case with 3 CHD-coded outpatients encounters is represented by a shade of purple, and a case with 13 CHD-coded outpatient healthcare encounters (i.e., the maximum) would be represented by a red dot. As number of CHD-coded outpatient healthcare encounters, overall CHD-coded healthcare encounters, and ED visits increased, the model more often predicted the case to be a TP. On the other hand, the model more often predicted cases to be FP as age increased or if “other” CHD codes were detected—that is, these features are negative predictors of correct classification.

3.3 | Model performance

Median performance metrics summarizing the 20 metrics across the five folds per site for the four machine learning models, LR, GaussianNB, RF, and XGBoost, are presented in Table 2. Median PPV in each test dataset varied from a high of 0.94 in XGBoost to a low of 0.78 in the LR model, and the F1-score, a combined measure of precision and recall, varied

from 0.84 for XGBoost to 0.52 for GaussianNB. Having outperformed the other models with the test datasets, XGBoost was then used on the entire dataset and yielded a PPV of 0.94 (437/465) and an NPV of 0.49 (155/314). In Figure 4, we present the receiver operating curve (ROC) analyses corresponding to the median AUROC values in Table 2 for the four models on the test datasets. In Figure 5, we present the area under the precision-recall curves (AUPRC) for the four models on the test datasets; the XGBoost model again had the best performance with an AUPRC = 0.88, despite the imbalanced dataset in which TP and FP cases were unevenly split (Figure 5). The average PPV-false negative rate curve with CI for the XGBoost model is in Figure 6. The red dot denotes the optimal operating point selected using the PPV from our final XGBoost model (PPV = 0.94), which has a corresponding false negative rate = 0.27. Further, increases in PPV are associated with exponentially higher false negative rates; thus this point, and the curve, represents a trade-off between PPV and false negative classification.

Table 3 includes metrics describing XGBoost's performance on the four site-specific datasets. PPV varied across sites from a high of 0.94 (116/124) in GA to a low of 0.86 (101/117) in NY, and F1 score varied from a high of 0.86 for GA to 0.79 for NY. We found XGBoost for GA to have the best performance with an AUROC value of 0.84, which was larger than NC (0.81), NY (0.82), and UT (0.81) (Table 3). While there are some slight differences in terms of AUROC, PPV, NPV, sensitivity, specificity, and F1 score values, performance metric values were similar across sites.

4 | DISCUSSION

Using ICD-9-CM codes alone, the PPV of CHD in the administrative and clinical dataset was 76.5%. When the ‘other’ CHD category is excluded, the PPV increased to 86.5% but with 83 false negatives (10.7% of the data set, 14.0% of the TP). Using an XGBoost machine learning model, we improved this PPV to 94% while keeping the NPV at 49%, demonstrating the utility of machine learning in increasing the accuracy of administrative data for surveillance of individuals with CHD. However, with the increase in PPV to 94%, 22% of TP CHD cases (131/596) were incorrectly labeled as not having CHD. An operating point can be selected when applying the algorithm to datasets to prioritize either PPV or to minimize false negative cases; in our case, we decided that the threshold of PPV = 0.94, corresponding to a false negative rate = 0.27, is an optimal operating point when applying the XGBoost model to future analyses to prioritize PPV. Figure 6 illustrates the trade-off between increases in PPV and false negative classification; each point on the curve created by a model. Increases in PPV beyond this operating point are associated with exponentially higher false negative rates.

Public health surveillance of CHD is important to understand factors contributing to short-term and long-term health-related outcomes following diagnosis and repair, defect-specific survival, comorbidities, and healthcare utilization. Capturing as many CHD cases as possible while ensuring accuracy of CHD datasets is important to public health CHD surveillance efforts. Administrative and clinical data are often used for CHD surveillance efforts because they offer large quantities of readily available longitudinal information on patients with CHD. However, unlike some homogenous populations readily identified by ICD codes,

CHD represents a heterogenous spectrum of native anatomy complexity, with surgical repairs that vary by surgical era and geography, and differences in billing practices and coding styles that impact how the dozens of CHD-related ICD-9-CM codes are used. Though excluded from our data, non-specific CHD codes, codes for extra-cardiac vascular anomalies, and the code used to indicate a normal patent foramen ovale variant (745.5) are often included in ICD-9-CM code group (745.XX–747.XX) used to identify CHD in administrative and clinical data despite evidence that the PPV of code group 745.XX–747.XX for a CHD is as low as 48.7% (Khan et al., 2018). After excluding the non-CHD codes in ICD-9-CM code group 745.XX–747.XX, and excluding isolated 745.5 from our data, the PPV was still less than ideal at 76.5%, although this was further improved to 86.5% by excluding the entire group of 'other' CHD codes in Rodriguez et al. (2022). While PPV can be increased by restricting the data to certain codes, or to specific encounter types and numbers of encounters with CHD codes, entire subsets of TP CHD cases (e.g., all individuals with atrial septal defects) would be excluded and the resulting data would be skewed toward more severe cases. Our results show that machine learning algorithms can be applied to large national datasets to increase accuracy of case identification and improve surveillance of CHD across the lifespan as a better alternative.

The performance of XGBoost was compared with LR, GaussianNB, and RF. LR is a statistical tool that helps identify relationships between multiple variables; it is a relatively fast model to apply compared to other models such as RF and XGBoost, which are often time consuming (Logistic Regression Analysis—an Overview | ScienceDirect Topics, 2020). However, LR may be limited in performance accuracy, especially for multiple complex variables (Logistic Regression Analysis—an Overview | ScienceDirect Topics, 2020). GaussianNB is a model that is based on Gaussian distribution presumptions of the feature sets and is considered an efficient machine learning model (Jahromi & Taheri, 2017). Although GaussianNB is efficient in time (faster in processing and training time) to classify, it suffers from weak conditional independence, which impacts model performance (Jahromi & Taheri, 2017). RF classification is an ensemble learning technique that uses combinations of decision trees for classification (Nguyen et al., 2013). It is a popular model for diagnosis classification because it incorporates large amounts of data, while limiting data overfitting which is a common, yet undesirable phenomenon inherent to machine learning models that can occur when the algorithm learns the details of a particular training dataset so well that it fails to be predictive and generalizable when tested with novel datasets (Dreiseitl & Ohno-Machado, 2002; Jabbar & Khan, 2014; Nguyen et al., 2013; Zhang et al., 2017). However, RF can be challenging as large datasets require high amounts of memory (Santur et al., 2016). Therefore, gradient tree boosting, specifically XGBoost, is often used to reduce overfitting and increase classification efficiency (Chen & Guestrin, 2016). XGBoost combines boosting, which applies classifications to reweighted versions of training data to increase performance, with incorporated regularized modeling to prevent overfitting of the model (Chen & Guestrin, 2016; Friedman et al., 2000). Due to issues associated with multiple complex variables, we did not expect favorable results from the GaussianNB or the LR models, and we expected comparable results from the RF and XGBoost models (Lundberg, 2020). Ultimately, we found that XGBoost outperformed all the other models considered.

While some researchers have utilized machine learning to evaluate risk factors for CHD and predict a pregnant person's risk of having an infant with CHD (Luo et al., 2017; Rani & Masood, 2020), the present study is the first, to our knowledge, to use machine learning to detect individuals who actually have a CHD in administrative records and distinguish them from individuals who have CHD codes but do not have CHD, for the purpose of enhancing CHD surveillance across the lifespan. We have demonstrated improved accuracy of CHD case identification in administrative data from diverse sources by developing a machine learning algorithm using XGBoost to detect TP CHD cases, yielding a PPV of 94%, and an F1-score of 0.84. PPV can be further increased at the expense of sensitivity, but optimization of F1-score is needed to represent the desired population most accurately.

In this analysis, the overlapping features most helpful in distinguishing TP and FP cases in development of the machine learning algorithm, and those that were included in the model were as follows: number of outpatient CHD-coded encounters, having a CHD of “other” anatomic group, number of CHD-coded healthcare encounters, age, a diagnosis in the CCS musculoskeletal group, circulatory group, electrocardiography group, factors influencing healthcare (medication management) group, respiratory system group, and number of ED visits were. Different features could be positive or negative predictors of true CHD classification. For example, increasing CHD-coded outpatient healthcare encounters, overall CHD-coded healthcare encounters, and ED visits led to more predictions of TP. On the other hand, the model more often predicted cases to be FP as age increased or if “other” CHD codes were detected. This feature reduction process helps avoid a ‘black box’ where there is uncertainty about the variables driving the algorithmic results. By using overlapping features identified as important in all datasets, we could ensure generalizability of the model and avoid reliance on a model that performs well in one dataset but whose results cannot be replicated in a different dataset. Principal component analysis (PCA) was applied to all the features, but a large number of principal components are required to account for 90% of total variance contribution, which means PCA would not be a good method for feature reduction and the dataset is too complex to be described by a few principal components. After reducing the model to features deemed important for each site, the XGBoost model had an AUPRC of 0.88 and a PPV of 94%. When further trained on data from three sites and applied to a fourth site, accuracy was maintained.

5 | LIMITATIONS

The small size of the validated dataset limited the number of methods that could be applied in model development. Even after feature reduction, the reduced feature set, totaling 7138 features, was larger than the cohort size of 779 patients. This limits the rank of the feature space to the number of patients, which technically limits the use of statistical feature selection algorithms (e.g., based on eigen-analysis). We had a large number of missing values on race and ethnicity variables, which limits understanding of the performance of the models by race or ethnicity. This may limit the robustness and generalizability of the selected feature set when applied to other cohorts. We utilized ICD-9-CM codes for this project, which would not be directly applicable to datasets with ICD-10-CM codes. The original dataset had excluded 745.5 (secundum ASD), thus results are not applicable to datasets that include this code. Future applications of ML to datasets inclusive of 745.5

would improve usefulness of ML in developing more accurate CHD datasets. Some of the selected features (e.g., number of ED visits, a diagnosis in respiratory system group) may not perform as expected during the coronavirus disease 2019 pandemic.

6 | CONCLUSIONS

Accuracy of administrative datasets to detect CHD can be improved with machine learning techniques, which may be further improved with larger validated datasets. Use of machine learning techniques for CHD case classification in administrative data has the potential to enhance public health surveillance efforts. More machine learning studies conducted using different datasets from other locations and other time spans with a more variety of patients will be needed to verify this finding.

ACKNOWLEDGMENTS

The authors would like to thank participating sites and CDC.

FUNDING INFORMATION

Centers for Disease Control and Prevention, Cooperative Agreement: Congenital Heart Defects Surveillance across Time And Regions (CHD STAR) – Component B (DD19-1902B) | 9/30/2020-9/29/2023 | Book (PI).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Centers for Disease Control and Prevention. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of Centers for Disease Control and Prevention.

APPENDIX A: Mutually exclusive congenital heart defect severity categories by International Classification of Disease version 9.0 Clinical Modification codes.

Category	ICD-9-CM[a] code	Code description
Severe (if case has a severe code, regardless of presence of shunt, valve, or other codes)	745.0	Common truncus
	745.1	Transposition of the great arteries (TGA)
	745.10	Complete TGA (dextro-TGA), not otherwise specified (NOS) or classical
	745.11	Double outlet right ventricle, or incomplete TGA
	745.12	Corrected TGA (levo-TGA)
	745.19	TGA other
	745.2	Tetralogy of Fallot
	745.3	Single ventricle, or cor triloculare

Category	ICD-9-CM[a] code	Code description
	745.6	Endocardial cushion defect
	745.60	Endocardial cushion defect unspecified
	745.69	Endocardial cushion defect, other
	746.01	Pulmonary valve atresia or absence
	746.1	Tricuspid atresia, stenosis or absence
	746.7	Hypoplastic left heart syndrome
	747.11	Interrupted aortic arch
	747.41	Total anomalous pulmonary venous return
Shunt + valve (case has shunt AND valve codes)	A combination of the shunt/valve codes below	A combination of the shunt/valve defects below
Shunt (case has at least one shunt code, no valve or severe codes)	745.4	Ventricular septal defect (VSD)
	745.61	ASD primum
	745.8	Other specified defect of septal closure
	745.9	Unspecified defect of septal closure
	747.0	Patent ductus arteriosus (PDA)
	747.42	Partial anomalous venous return
Valve (case has at least one valve code, no shunt or severe codes)	746.0	Anomalies of pulmonary valve
	746.00	Pulmonary valve anomaly, unspecified
	746.02	Pulmonary valve stenosis
	746.09	Pulmonary valve anomaly, other
	746.2	Ebstein Anomaly
	746.3	Aortic valve stenosis
	746.4[b]	Aortic insufficiency or bicuspid/unicuspid aortic valve[b]
	746.5	Mitral stenosis or mitral valve abnormalities
	746.6[b]	Mitral insufficiency[b]
	764.81	Subaortic stenosis
	746.83	Infundibular or subvalvar pulmonary stenosis
	747.1/747.10	Coarctation of aorta
	747.22	Atresia or stenosis of aorta
	747.3[b]	Anomalies of pulmonary artery[b]
	747.31	Pulmonary artery atresia, coarctation, or hypoplasia
	747.39	Anomalies of pulmonary artery, other
Other only (case only has one or more codes in this category)	745.7	Cor biloculare

Category	ICD-9-CM[a] code	Code description
	746.8	Other specified anomalies of heart
	746.82	Cor triatriatum
	746.84	Obstructive anomalies of heart
	746.85	Coronary artery anomaly
	746.87	Malposition of heart or apex
	746.89[b]	Other specified anomaly of heart (various types)[b]
	746.9[b]	Unspecified defect of heart[b]
	747.2	Other anomalies of the aorta
	747.20	Anomalies of aorta, unspecified
	747.21	Anomaly of aortic arch
	747.29	Other anomalies of aorta, other specified
	747.4	Anomalies of great veins
	747.40	Anomalies of great veins, unspecified
	747.49	Other anomalies of great veins
	747.9	Unspecified anomalies of circulatory system
	648.5x	Congenital cardiovascular disorders in the mother
	V13.5	Personal history of (corrected) congenital malformations of heart and circulatory system

Note: Individuals were ascertained by the site-specific surveillance system if they had any ICD-9-CM CHD diagnosis codes between 745.XX and 747.XX documented in a healthcare encounter between January 1, 2011 and December 31, 2013, excluding: Atrial septal defect (ASD) secundum or Patent Foramen Ovale (745.5) congenital heart block (746.86), absent/hypoplastic umbilical artery (747.5), pulmonary arteriovenous malformation (747.32), other anomalies of peripheral vascular system (747.6X), and other specified anomalies of circulatory system (747.8X).

[a] International Classification of Disease version 9.0 Clinical Modification.

[b] Codes considered to be minor. All other listed codes considered to be major.

REFERENCES

Agarwal S, Sud K, & Menon V (2016). Nationwide hospitalization trends in adult congenital heart disease across 2003–2012. Journal of the American Heart Association, 5(1), e002330. 10.1161/JAHA.115.002330 [PubMed: 26786543]

Bhatt AB, Foster E, Kuehl K, Alpert J, Brabeck S, Crumb S, Davidson WR, Earing MG, Ghoshhajra BB, Karamlou T, Mital S, Ting J, & Tseng ZH (2015). Congenital heart disease in the older adult: A scientific statement from the American Heart Association. Circulation, 131(21), 1884–1931. 10.1161/CIR.0000000000000204 [PubMed: 25896865]

Billett J, Cowie MR, Gatzoulis MA, Vonder Muhll IF, & Majeed A (2008). Comorbidity, healthcare utilisation and process of care measures in patients with congenital heart disease in the UK: Cross-sectional, population-based study with case-control analysis. Heart, 94(9), 1194–1199. 10.1136/hrt.2007.122671 [PubMed: 17646191]

Botto LD, & Correa A (2003). Decreasing the burden of congenital heart anomalies: An epidemiologic evaluation of risk factors and survival. Progress in Pediatric Cardiology, 18(2), 111–121. 10.1016/S1058-9813(03)00084-5

Brida M, & Gatzoulis MA (2019). Adult congenital heart disease: Past, present and future. Acta Paediatrica, 108(10), 1757–1764. 10.1111/apa.14921 [PubMed: 31254360]

Broberg C, McLarry J, Mitchell J, Winter C, Doberne J, Woods P, Burchill L, & Weiss J (2015). Accuracy of administrative data for detection and categorization of adult congenital heart disease patients from an electronic medical record. Pediatric Cardiology, 36(4), 719–725. 10.1007/s00246-014-1068-2 [PubMed: 25428778]

Chen T, & Guestrin C (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785–794). Association for Computing Machinery. 10.1145/2939672.2939785

Dreiseitl S, & Ohno-Machado L (2002). Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics, 35(5–6), 352–359. 10.1016/s1532-0464(03)00034-0 [PubMed: 12968784]

Explaining machine learning models: A non-technical guide to interpreting SHAP analyses. (2021). Impromptu Engineer. Retrieved July 6, 2023, from https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/

Friedman J, Hastie T, & Tibshirani R (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337–407. 10.1214/aos/1016218223

Gilboa SM, Devine OJ, Kucik JE, Oster ME, Riehle-Colarusso T, Nembhard WN, Xu P, Correa A, Jenkins K, & Marelli AJ (2016). Congenital heart defects in the United States: Estimating the magnitude of the affected population in 2010. Circulation, 134(2), 101–109. 10.1161/CIRCULATIONAHA.115.019307 [PubMed: 27382105]

Jabbar HK, & Khan RZ (2014). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices, 70(10.3850), 978–981. 10.3850/978-981-09-5247-1_017

Jahromi AH, & Taheri M (2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. Artificial Intelligence and Signal Processing Conference, 2017, 209–212. 10.1109/AISP.2017.8324083

Khan A, Ramsey K, Ballard C, Armstrong E, Burchill LJ, Menashe V, Pantely G, & Broberg CS (2018). Limited accuracy of administrative data for the identification and classification of adult congenital heart disease. Journal of the American Heart Association, 7(2), e007378. 10.1161/JAHA.117.007378 [PubMed: 29330259]

Liberman RF, Heinke D, Lin AE, Nestoridi E, Jalali M, Markenson GR, Sekhavat S, & Yazdy MM (2023). Trends in delayed diagnosis of critical congenital heart defects in an era of enhanced screening, 2004–2018. The Journal of Pediatrics, 257, 113366. 10.1016/j.jpeds.2023.02.012 [PubMed: 36858148]

Logistic Regression Analysis—An overview | ScienceDirect Topics. (2020). Elsevier B. V. Retrieved September 24, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis

Lundberg S. (2020). October 6, Interpretable Machine Learning with XGBoost. Medium. https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Lundberg S, & Lee S-I (2017). A unified approach to interpreting model predictions. arXiv, 30, 4768–4777. 10.48550/arXiv.1705.07874

Luo Y, Li Z, Guo H, Cao H, Song C, Guo X, & Zhang Y (2017). Predicting congenital heart defects: A comparison of three data mining methods. PLoS One, 12(5), e0177811. 10.1371/journal.pone.0177811 [PubMed: 28542318]

Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, & Kaouache M (2014). Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. Circulation, 130(9), 749–756. 10.1161/CIRCULATIONAHA.113.008396 [PubMed: 24944314]

Massin MM, & Dessy H (2006). Delayed recognition of congenital heart disease. Postgraduate Medical Journal, 82(969), 468–470. 10.1136/pgmj.2005.044495 [PubMed: 16822925]

Müller MJ, Norozi K, Caroline J, Sedlak N, Bock J, Paul T, Geyer S, & Dellas C (2022). Morbidity and mortality in adults with congenital heart defects in the third and fourth life decade. Clinical Research in Cardiology, 111(8), 900–911. 10.1007/s00392-022-01989-1 [PubMed: 35229166]

Mylotte D, Pilote L, Ionescu-Ittu R, Abrahamowicz M, Khairy P, Therrien J, Mackie AS, & Marelli A (2014). Specialized adult congenital heart disease care: The impact of policy on mortality. Circulation, 129(18), 1804–1812. 10.1161/CIRCULATIONAHA.113.005817 [PubMed: 24589851]

Nguyen C, Wang Y, & Nguyen HN (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Journal of Biomedical Science and Engineering, 6(5), 551–560. 10.4236/jbise.2013.65070

Oster ME, Lee KA, Honein MA, Riehle-Colarusso T, Shin M, & Correa A (2013). Temporal trends in survival among infants with critical congenital heart defects. Pediatrics, 131(5), e1502–e1508. 10.1542/peds.2012-3435 [PubMed: 23610203]

Rani S, & Masood S (2020). Predicting congenital heart disease using machine learning techniques. Journal of Discrete Mathematical Sciences and Cryptography, 23(1), 293–303. 10.1080/09720529.2020.1721862

Rodriguez FH, Ephrem G, Gerardin JF, Raskind-Hood C, Hogue C, & Book W (2018). The 745.5 issue in code-based, adult congenital heart disease population studies: Relevance to current and future ICD-9-CM and ICD-10-CM studies. Congenital Heart Disease, 13(1), 59–64. 10.1111/chd.12563 [PubMed: 29266726]

Rodriguez FH, Raskind-Hood CL, Hoffman T, Farr SL, Glidewell J, Li JS, D'Ottavio A, Botto L, Reeder MR, Hsu D, Lui GK, Sullivan AM, & Book WM (2022). How well do ICD-9-CM codes predict true congenital heart defects? A Centers for Disease Control and Prevention-based multisite validation project. Journal of the American Heart Association, 11(15), e024911. 10.1161/JAHA.121.024911 [PubMed: 35862148]

Santur Y, Karaköse M, & Akin E (2016). Random forest based diagnosis approach for rail fault inspection in railways. In 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO). Institute of Electrical and Electronics Engineers (IEEE). (pp. 745–750).

Stout KK, Daniels CJ, Aboulhosn JA, Bozkurt B, Broberg CS, Colman JM, Crumb SR, Dearani JA, Fuller S, Gurvitz M, Khairy P, Landzberg MJ, Saidi A, Valente AM, & Van Hare GF (2019). 2018 AHA/ACC guideline for the management of adults with congenital heart disease: Executive summary: A report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Circulation, 139(14), e637–e697. 10.1161/CIR.0000000000000602 [PubMed: 30586768]

Udholm S, Rex C, Eckerström F, Onat M, Nyboe C, & Hjortdal VE (2019). Small unrepaired atrial septal defects display impaired exercise capacity compared with healthy peers. Congenital Heart Disease, 14(3), 372–379. 10.1111/chd.12740 [PubMed: 30582285]

Wallace F, Buratto E, Schulz A, d'Udekem Y, Weintraub RG, Brizard CP, & Konstantinov IE (2022). Long-term outcomes of primary aortic valve repair for isolated congenital aortic stenosis in children. The Journal of Thoracic and Cardiovascular Surgery, 164(5), 1263–1274.e1. 10.1016/j.jtcvs.2021.11.097 [PubMed: 35430079]

Warnes CA, Liberthson R, Danielson GK, Dore A, Harris L, Hoffman JI, Somerville J, Williams RG, & Webb GD (2001). Task force 1: The changing profile of congenital heart disease in adult life. Journal of the American College of Cardiology, 37(5), 1170–1175. 10.1016/s0735-1097(01)01272-4 [PubMed: 11300418]

Yang Q, Rasmussen SA, & Friedman JM (2002). Mortality associated with Down's syndrome in the USA from 1983 to 1997: A population-based study. Lancet, 359(9311), 1019–1025. 10.1016/s0140-6736(02)08092-3 [PubMed: 11937181]

Zhang C, Bengio S, Hardt M, Recht B, & Vinyals O (2017). Understanding deep learning requires rethinking generalization. arXiv, 64(3), 107–115. 10.48550/arXiv.1611.03530

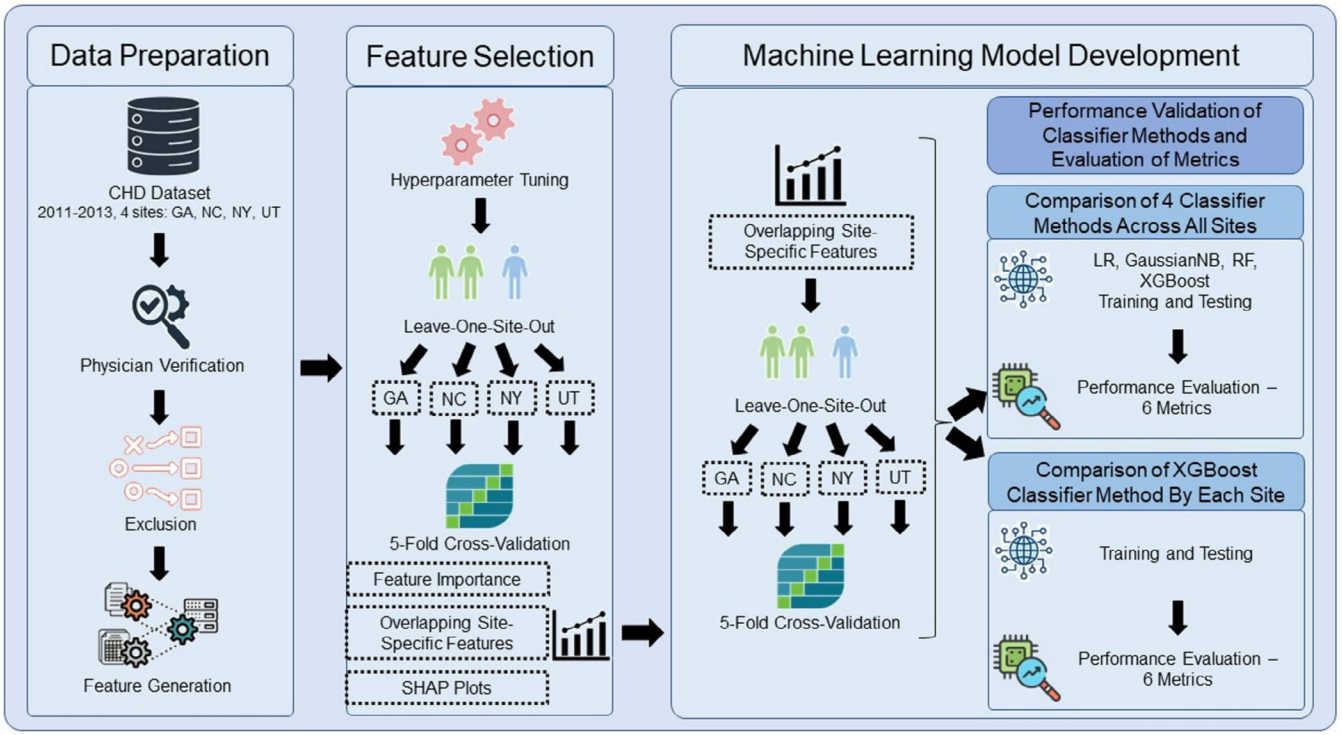

FIGURE 1.

Machine learning model development for CHD prediction. Depiction of algorithm development and cross-validation procedures. CHD, congenital heart defects; GA, Georgia; GaussianNB, Gaussian Naive Bayes; LR, Logistic Regression; NC, North Carolina; NY, New York; RF, Random Forest; SHAP, SHapley Additive exPlanation values generated using XGBoost; UT, Utah. *Six performance metrics*: AUROC, area under the receiver operating characteristic curve; PPV (positive predictive value), 100 × TP/(TP + FP); NPV (negative predictive value), 100 × TN/(TN + FN); sensitivity, 100 × TP/(TP + FN); specificity, 100 × TN/(TN + FP); F1 score, 2 × (PPV × sensitivity)/(PPV + sensitivity).

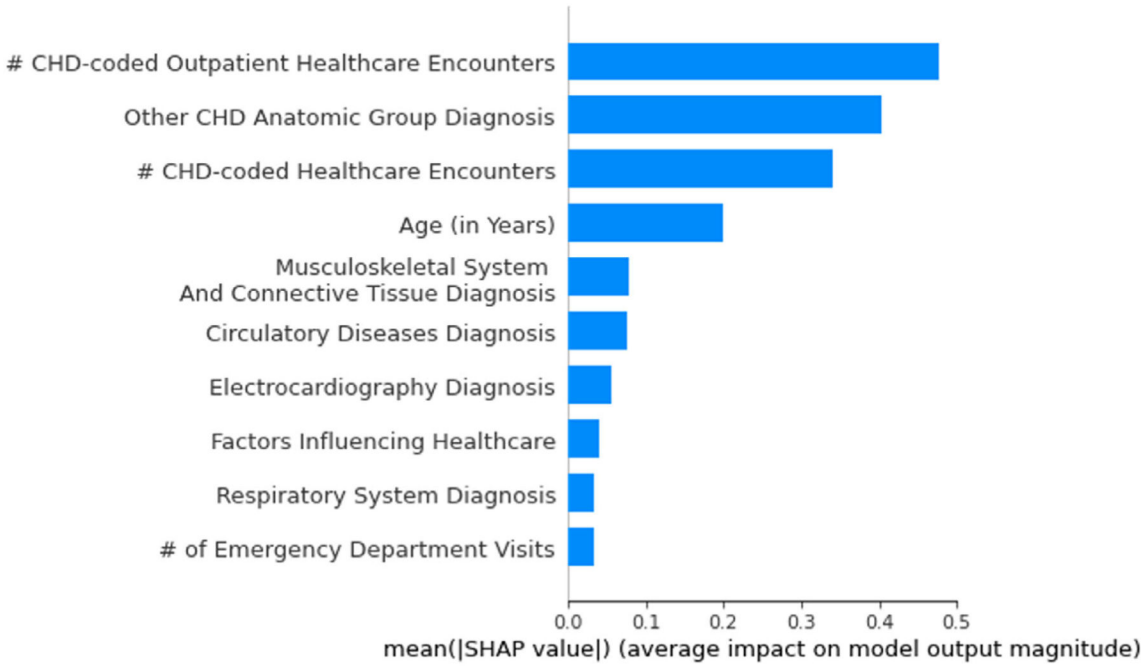

FIGURE 2.

Bar plot of mean absolute SHAP values for common relevant features for congenital heart defect (CHD) prediction using the XGBoost model. This bar plot shows the mean absolute SHAP values for common relevant features (variable) across all the data using the XGBoost model for CHD prediction. The key aspects of this bar plot are the ordering of features and the relative magnitude (positive or negative) of the mean absolute SHAP values; the mean absolute SHAP value quantifies, on average, how much the feature impacts prediction in the positive or negative direction. Features with higher mean absolute SHAP values like # CHD-coded outpatient healthcare encounters, 'other' CHD anatomic group diagnosis, and # CHD-coded healthcare encounters are more influential. Other features that influence prediction include several CCS categories: having a musculoskeletal system and connective tissue diagnosis; a circulatory disease diagnosis; an electrocardiography diagnosis; "factors influencing healthcare" which includes a subcategory "medication management"; and a respiratory system diagnosis (Rodriguez et al., 2022). CHD, congenital heart defects; SHAP, SHapley Additive exPlanation values.

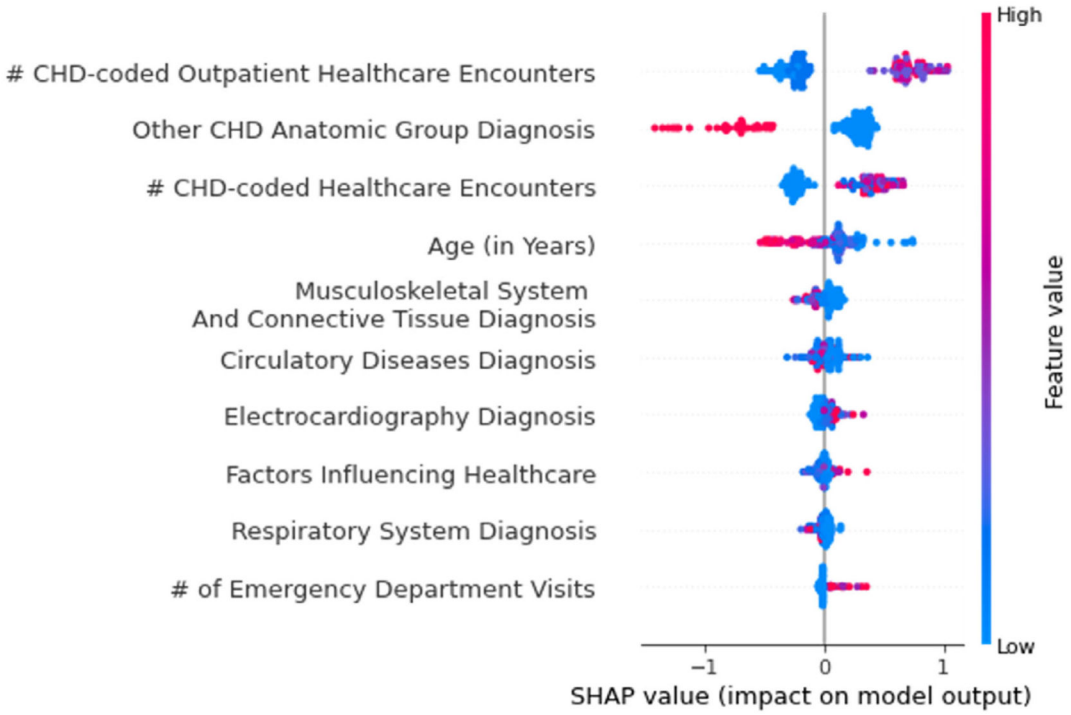

FIGURE 3.

SHAP summary plot of common relevant features for congenital heart defect (CHD) prediction using the XGBoost model. This figure shows the SHAP summary plot of the 10 common relevant features across all the data using an XGBoost model for CHD prediction. The color bars represent raw SHAP values for each feature. Each dot represents one person; red color represents cases are present, while blue color represents cases who are absent. Features with higher predicted CHD risk, including # outpatient CHD healthcare encounters, # CHD healthcare encounters, circulatory diseases diagnosis, electrocardiography diagnosis, factors influencing healthcare (medication management), and emergency department visit, are denoted by SHAP values whose bars have red portions to the right of '0', whereas features with lower predicted CHD risk, including having a CHD 'other' group diagnosis, age (in years), musculoskeletal system and connective tissue diagnosis, and respiratory system diagnosis, are denoted by SHAP values whose bars have red portions to the left of '0'. CHD, congenital heart defects; SHAP, SHapley Additive exPlanation values.

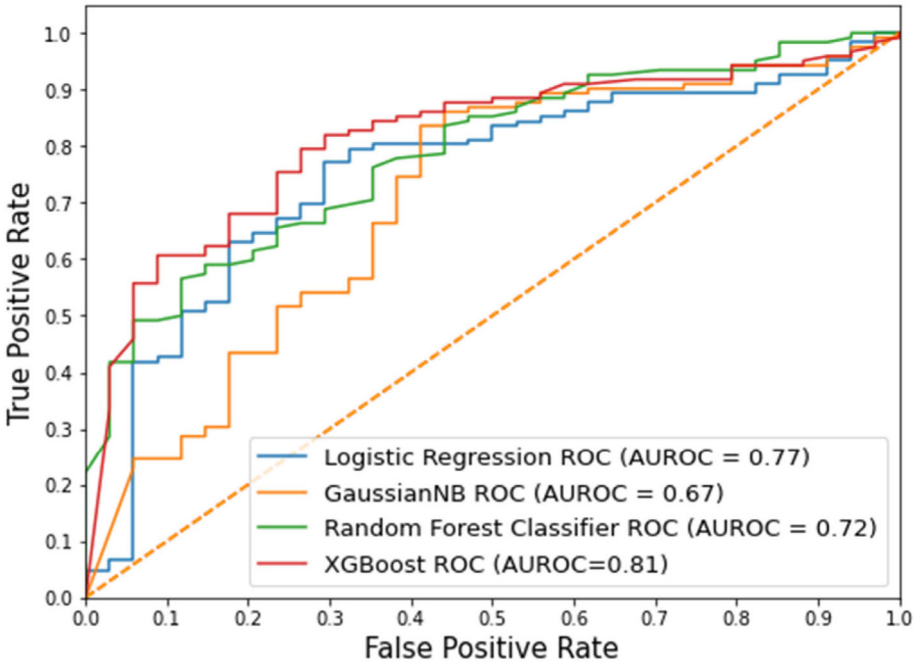

FIGURE 4.

Receiver operating curve (ROC) analyses with area under the receiver operating curve (AUROC) values for four machine learning models. This figure shows the ROC curve for four different machine learning models for CHD prediction. XGBoost model has the highest AUROC, 0.81, compared to other models. AUROC, area under the receiver operating curve; GaussianNB, Gaussian Naïve Bayes; ROC, receiver operating curve.

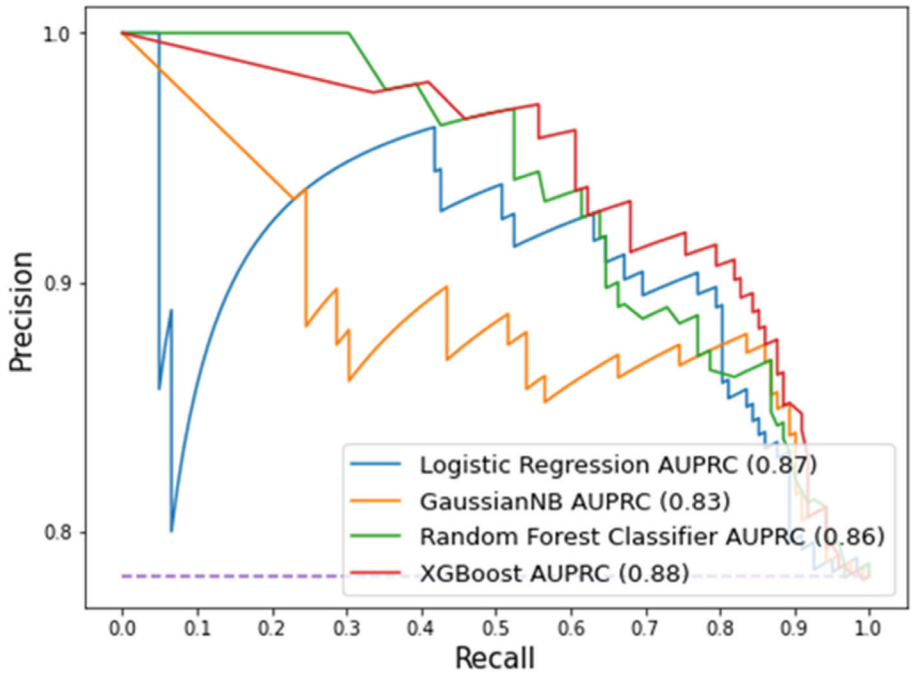

FIGURE 5.

Area under the precision-curve (AUPRC) analyses for four machine learning models. This figure shows the precision-recall curve for four machine leaning models for CHD prediction. XGBoost model has the highest AUPRC, 0.88, compared to other machine learning models. AUPRC, area under the precision-recall curve; GaussianNB, Gaussian Naïve Bayes.

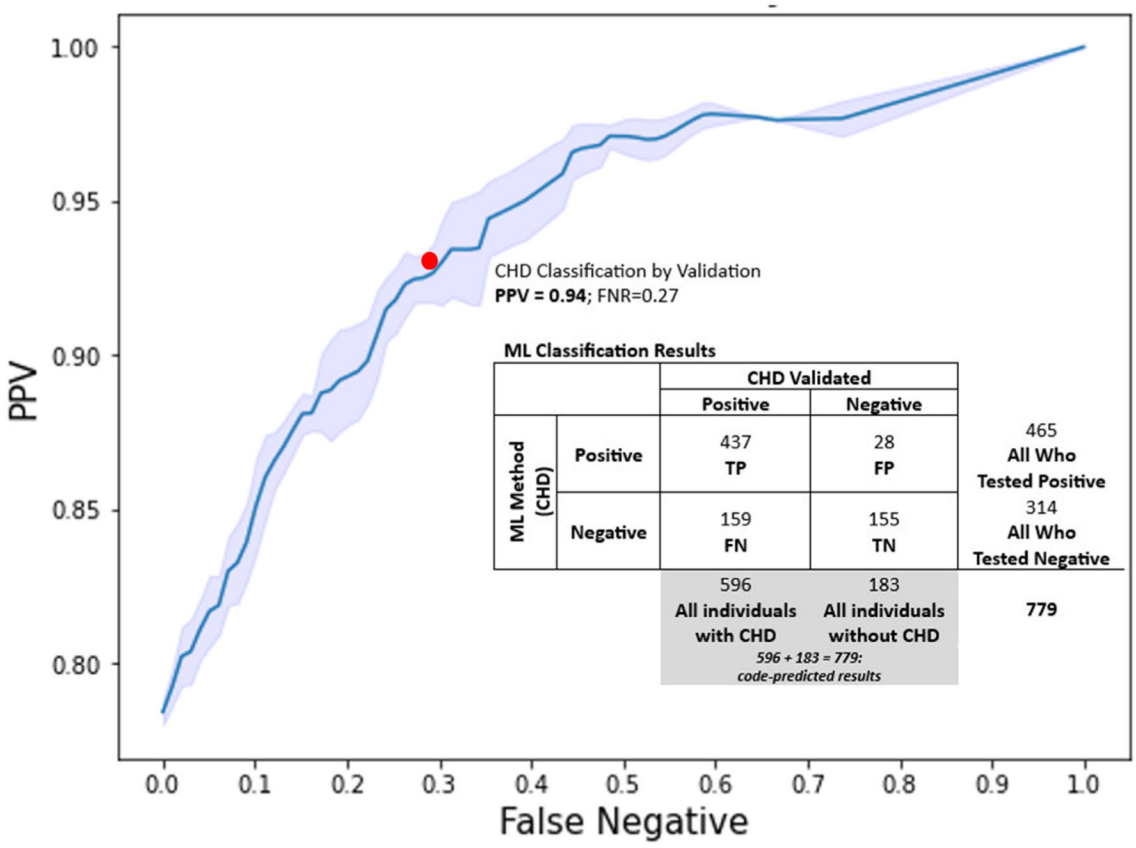

FIGURE 6.

Positive predictive value (PPV)–false negative (FN) rate analyses for the XGBoost model for four sites. This figure shows the PPV–FN rate curve for CHD prediction for pooled data across four sites. The operating point is at the threshold of PPV = 0.94 and FN rate = 0.35, as seen by the red dot. Less increase in PPV is associated with per unit FN rates increase since PPV reaches a plateau. For any point beyond 0.94, substantially more FN will be sacrificed to achieve a higher PPV. The dark blue line shows the mean of the five-fold cross validation, while the light blue shaded area represents the range of values within 1 SD of the mean. The dark blue line is the mean of the five-fold cross validation rather than real model performances. The dot is the model with the highest PPV and the lowest false negative rate among five-fold cross validation, thus does not appear on the dark blue line. FN, false negative; PPV, positive predictive value; SD, standard deviation. TP (true positive), individual correctly identified as having CHD FP (false positive), individual incorrectly identified as having CHD (FP/FP + TN); TN (true negative), individual correctly identified as not having CHD FN (false negative), individual incorrectly identified as not having CHD (FN/FN + TP); PPV = 100 × TP/(TP + FP) = 100 × 437/(437 + 28) = 94%; NPV = 100 × TN/(TN + FN) = 100 × 155/(155 + 159) = 49%; accuracy = 100 × (TP + TN)/N = 100 × (437 + 155)/779 = 80%; ensitivity = 100 × TP/(TP + FN) = 100 × 437/(437 + 159) = 73%; pecificity = 100 × TN/(TN + FP) = 100 × 155/(28 + 155) = 85%; CHD prevalence = 100 × (TP + FN)/N = 100 × (437 + 159)/779 = 75%.

TABLE 1

Sample demographics.

		Validated (based on medical record validation)		
	Total, *N* = 779	**True positive CHD, *n* = 596 (76.5%)**	**False positive CHD, *n* = 183 (23.5%)**	X^2
Validated CHD anatomic grouping				
Severe	154	154 (100.0%)		–
Shunt	194	194 (100.0%)		
Valve	178	178 (100.0%)		
Shunt + valve	<11	<11 (−)		
Other	64	64 (100.0%)		
Not CHD	183	0 (0.0%)	183 (100.0%)	
Validated CHD anatomic grouping				
Severe	154	154 (100.0%)		–
Non-severe	442	442 (100.0%)		
Not CHD	183		183 (100.0%)	
Site				
GA	195	153 (78.5%)	42 (21.5%)	5.61 *p = .1322*
NC	198	159 (80.3%)	39 (19.7%)	
NY	191	135 (70.7%)	56 (29.3%)	
UT	195	149 (76.4%)	46 (23.6%)	
Sex				
Male	380	297 (78.2%)	83 (21.8%)	1.12
Female	399	299 (74.9%)	100 (25.1%)	*p = .2892*
Age group[a] (in years)				
Mean age (SD) (in years)	30.97 (25.02)	27.01 (22.76)	43.86 (27.64)	*t* = 8.31 *p < .0001*
Race and ethnicity				
Non-Hispanic White	399 (51.2%)	310 (77.7%)	89 (22.3%)	6.67
Non-Hispanic Black	116 (14.9%)	78 (67.2%)	38 (32.8%)	*p = .0831*
Hispanic	86 (11.0%)	67 (77.9%)	19 (22.1%)	
Other[b]	178 (22.9%)	141 (79.2%)	37 (20.8%)	

		Validated (based on medical record validation)		
	Total, *N* = 779	**True positive CHD, *n* = 596 (76.5%)**	**False positive CHD, *n* = 183 (23.5%)**	***X*** ^{2}
Insurance				
Private only	282	236 (83.7%)	46 (16.3%)	–
Any public	429	307 (71.6%)	122 (28.4%)	
Selfpay/Uninsured	<11	<11 (−)	<11 (−)	
Unknown	63	49 (77.8%)	14 (22.2%)	

Note: Column percentages in cells. X^2 does not include the total column, only the true positive and false positive categories. Abbreviations: CHD, congenital heart defects; GA, Georgia; NC, North Carolina; NY, New York; SD, standard deviation; UT, Utah.

[a]Age group: UT included 11–64-year-olds.

[b]Other race/ethnicity group: Asian, American Indian/Native American, native Hawaiian/Pacific Islander, and multi-racial.

TABLE 2

Performance validation of four machine learning models across four sites.

Model	AUROC	PPV	NPV	Sensitivity	Specificity	F1-score
LR	0.77 [0.69, 0.84]	0.78 [0.71, 0.82]	0.76 [0.71, 0.82]	0.76 [0.67, 0.84]	0.76 [0.72, 0.83]	0.77 [0.69, 0.83]
GaussianNB	0.67 [0.57, 0.74]	0.85 [0.68, 0.98]	0.60 [0.56, 0.65]	0.37 [0.27, 0.48]	0.92 [0.87, 0.98]	0.52 [0.40, 0.63]
RF	0.72 [0.68, 0.77]	0.82 [0.76, 0.87]	0.76 [0.70, 0.82]	0.72 [0.63, 0.82]	0.84 [0.80, 0.87]	0.76 [0.69, 0.84]
XGBoost	0.81 [0.78, 0.82]	0.94 [0.91, 0.96]	0.46 [0.34, 0.56]	0.76 [0.64, 0.88]	0.81 [0.68, 0.88]	0.84 [0.76, 0.91]
XGBoost entire dataset	0.82 [0.81, 0.83]	0.94 [0.94, 0.95]	0.49 [0.48, 0.53]	0.73 [0.70, 0.77]	0.85 [0.83, 0.89]	0.82 [0.81, 0.85]

Note: Four machine learning models conducted with five-fold cross-validation across four sites; median and 95% confidence intervals (95% CI) displayed for six performance metrics.

Abbreviations: AUROC, area under the receiver operating characteristics; CI, confidence interval; GaussianNB, Gaussian Naïve Bayes; LR, Logistic Regression; NPV, negative predictive value; PPV, positive predictive value; RF, Random Forest.

TABLE 3

Performance validation of the XGBoost model for each of four sites.

	AUROC	PPV	NPV	Sensitivity	Specificity	F1-score
XGBoost for GA	0.84 [0.80, 0.86]	0.94 [0.92, 0.96]	0.49 [0.38, 0.53]	0.77 [0.68, 0.84]	0.82 [0.72, 0.88]	0.86 [0.79, 0.89]
XGBoost for NC	0.81 [0.80, 0.82]	0.92 [0.90, 0.92]	0.43 [0.41, 0.45]	0.73 [0.71, 0.75]	0.76 [0.70, 0.78]	0.81 [0.80, 0.82]
XGBoost for NY	0.82 [0.81, 0.83]	0.86 [0.84, 0.89]	0.52 [0.51, 0.55]	0.73 [0.70, 0.77]	0.71 [0.65, 0.78]	0.79 [0.78, 0.81]
XGBoost for UT	0.81 [0.80, 0.83]	0.89 [0.87, 0.91]	0.45 [0.42, 0.47]	0.72 [0.69, 0.76]	0.72 [0.65, 0.76]	0.80 [0.78, 0.82]

Note: The XGBoost model was conducted with five-fold cross-validation by each of four sites; median and 95% confidence intervals (95% CI) are displayed for six performance metrics.

Abbreviations: AUROC, area under the receiver operating characteristics; CI, confidence interval; GA, Georgia; NC, North Carolina; NPV, negative predictive value; NY, New York; PPV, positive predictive value; UT, Utah.