



Published in final edited form as:

Birth Defects Res. 2025 February ; 117(2): e2440. doi:10.1002/bdr2.2440.

A Generalized Machine Learning Model for Identifying Congenital Heart Defects (CHD) Using ICD Codes

Haoming Shi¹, Wendy M. Book^{2,3}, Lindsey C. Ivey³, Fred H. Rodriguez III^{2,4}, Cheryl Raskind-Hood³, Karrie F. Downing⁵, Sherry L. Farr⁵, Courtney E. McCracken⁶, Vinita O. Leedom⁷, Susan E. Haynes⁸, Sandra Amouzou⁶, Reza Sameni^{1,9}, Rishikesan Kamaleswaran^{10,11,12,13}

¹Department of Biomedical Engineering, Georgia Institute Technology, Atlanta, GA

²Department of Medicine, Division of Cardiology, Emory University School of Medicine, Atlanta, GA

³Department of Epidemiology, Emory University, Rollins School of Public Health, Atlanta, GA

⁴Children's Healthcare of Atlanta, Atlanta, GA

⁵National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA

⁶Center for Research and Evaluation, Kaiser Permanente Georgia, Atlanta, GA

⁷South Carolina Department of Health and Environmental Control

⁸Prisma Health Upstate, South Carolina

⁹Department of Biomedical Informatics, Emory University School of Medicine, Atlanta GA

¹⁰Department of Surgery, Duke University School of Medicine, Durham NC

¹¹Department of Anesthesiology, Duke University School of Medicine, Durham NC

¹²Department of Biomedical Engineering, Duke University, Durham NC

¹³Department of Electrical and Computer Engineering, Duke University, Durham NC

Abstract

Background: *International Classification of Diseases* (ICD) codes utilized for congenital heart defect (CHD) case identification in datasets have substantial false positive (FP) rates. Incorporating machine learning (ML) algorithms following case selection by ICD codes may improve accuracy of CHD identification, enhancing surveillance efforts.

Corresponding Author: Haoming Shi, haoming.shi@emory.edu.
Reza Sameni and Rishikesan Kamaleswaran are both co-senior authors

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

DBDID Replication Statement: Analyses have been replicated by Cheryl Raskind-Hood, CDC/NCBDDD and Emory University

Disclosures

The authors have no conflicts to declare.

Methods: Traditional ML methods were applied to four encounter-level datasets, 2010–2019, for 3,334 patients with validated diagnoses and with at least one CHD ICD code identified. A 5-fold cross-validation approach was applied to the dataset to determine the set of overlapping important features best classifying CHD cases. Training and testing combinations were explored to determine the approach yielding most accurate CHD classification.

Results: CHD ICD positive predictive values (PPVs) by site ranged from 53.2% to 84.0%. The ML algorithm achieved a PPV of 95% (1273/1340) for the four-site dataset with a false negative (FN) rate of 33% (639/1912) by choosing an operating point prioritizing PPV from the PPV-FN rate curve. XGBoost reduced 2,105 Clinical Classification Software (CCS) features to 137 that identified those with true positive (TP) CHD and false positive FP classification.

Conclusion: Applying machine learning algorithms following case selection by CHD-related ICD codes improved the accuracy of identifying TP true positive CHD cases.

Keywords

Machine Learning; Congenital Heart Disease; Population Health

Introduction

Congenital heart defects (CHD), characterized by structural heart and thoracic vessel abnormalities, are the most common birth defects in infants (Mai et al., 2019), with a prevalence of 8 per 1,000 live births in the U.S. (Hoffman & Kaplan, 2002; Reller et al., 2008), representing a leading cause of infant mortality (Almli et al., 2017). CHD represents a spectrum of structural heart defects affecting valves, chambers, and blood vessels, leading to a heterogenous spectrum of cardiac and non-cardiac complications (Lui et al., 2019). Some present with CHD in infancy (Oster et al., 2013), and others present in adulthood (van der Bom et al., 2011). Approximately 1.4 million adults are living with CHD in the U.S., which is more than estimated pediatric CHD cases (Gilboa et al., 2016). CHD patients may develop late complications including atrial fibrillation, hypertension, heart failure, and other cardiac-related conditions, which significantly increase morbidity and mortality (Billett et al., 2008; Kuehl et al., 1999; Lewis et al., 2023; Mahle et al., 2009; Menachem et al., 2020; Yang et al., 2002). While advancements in medical screenings and treatments have improved survival of individuals with CHD, these individuals face substantial late morbidity and premature mortality (Gilboa et al., 2016; Liberman et al., 2023; Massin & Dessy, 2006; Müller et al., 2022). Numerous factors affect late outcomes, including defect anatomy, type of repair, number and type of interventions, age, access to healthcare, and the treatment plan (Bhatt et al., 2015; Brida & Gatzoulis, 2019; Stout et al., 2019).

Population-based CHD surveillance often relies upon *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) and *Tenth Revision, Clinical Modification* (ICD-10-CM) codes, within extensive administrative and clinical datasets to estimate CHD prevalence, healthcare utilization, and various health outcomes (Ivey et al., 2024; Glidewell et al., 2021; Lui et al., 2022). The accuracy of ICD codes to identify CHD in administrative data can differ depending on age, medical practices, healthcare systems, and geographic regions; consequently, CHD ICD codes can have low positive predictive

values (PPV), which describe the percentage of people who actually have CHD among those who have documented CHD codes. Some CHD ICD codes lack specificity and correspond to more than one CHD diagnosis, and some have high false positive (FP) rates ineffectively distinguishing individuals with a heart defect from those without a CHD (Agarwal et al., 2016; Broberg et al., 2015; Khan et al., 2018; Rodriguez et al., 2018, 2022). For example, the ICD-9-CM code 745.5, used to identify both secundum atrial septal defects (ASD) and patent foramen ovale (PFO), exhibits a high FP rate for CHD where only 3 in 10 adults with that code have a true CHD (Rodriguez et al., 2018). Similarly, high FP rates have been observed for subaortic stenosis and pulmonary valve stenosis (Khan et al., 2018), and the FP rate for ICD-9-CM codes related to shunt lesions has been reported to be around 50% (Broberg et al., 2015). From a healthcare system perspective, FPs may result from use of “rule out” codes when ordering tests, applying CHD ICD codes to other heritable conditions or from data entry errors. Limiting datasets to patients with severe CHD codes improves the PPV of ICD codes, but also restricts the understanding of outcomes in patients with non-severe defects, such as bicuspid aortic valves, which can still lead to considerable morbidity (Udholm et al., 2019; Wallace et al., 2022).

Our previous research has found that the ability to identify people with CHD in administrative and clinical data can be significantly improved by performing machine learning (ML) methods without restricting to the most severe cases (Shi et al., 2023). ML involves the creation of computer algorithms that enhance task performance by learning and adapting from data. ML has found numerous applications in healthcare delivery, including computer-assisted diagnostics, risk prediction and stratification, clinical decision support, deep phenotyping, and precision medicine (Jone et al., 2022). The recent exponential growth of ML applications in healthcare can be attributed to the advancements in computational power, which now enable the processing of vast amounts of complex reference data at high speeds.

Using data from combined U.S.-based clinical and administrative sources, we aimed to develop and test generalized ML models that improve the accuracy of identifying CHD using various training approaches. Furthermore, this study aimed to identify the most crucial features that contribute to accurate identification of CHD.

Key contributions of this work include (1) the development and validation of a generalized ML pipeline for CHD prediction, which has shown robust performance across multiple clinical sites and holds significant promise for application on external datasets. By incorporating a transfer learning approach, the model dynamically adapts as more data becomes available, improving generalization and mitigating the risk of overfitting. This adaptability enhances the model’s robustness and scalability, making it more effective in diverse clinical settings and with expanding datasets. Furthermore, the pipeline (2) addresses key limitations of traditional ICD-based CHD classifications, which often have low PPV and high FP rates.

Methods

This study utilized training and validation data from four datasets, collected under a Cooperative Agreement with the Centers for Disease Control and Prevention (CDC-RFA-DD19–1202B) for CHD surveillance during 2010–2019. The study received Institutional Review Board approval on August 26, 2020 (STUDY00001030) and was granted a complete waiver of Health Insurance Portability and Accountability Act (HIPAA) authorization and informed consent. Participating sites were responsible for their respective Institutional Review Board approval. The cohort for case validation consisted of individuals born between 1/1/1955 and 12/31/2019, who were identified with a healthcare encounter between 1/1/2010–12/31/2019. The ICD codes used to define the CHD cohort are in Supplemental Table 1. Throughout the manuscript, the term “cases” will refer to identified individuals with CHD-related healthcare encounters between 1/1/2010–12/31/2019, including those who may not actually have a CHD.

Data Sources

Data sources included 4 datasets with validated CHD diagnoses from 4 sites.

1. Previously validated cohort from Emory Healthcare and Children’s Healthcare of Atlanta (EHC/CHOA), a Georgia-statewide adult and a statewide pediatric hospital system, respectively;
2. From a cohort of 15,504 individuals, identified in EHC meeting inclusion criteria based on CHD ICD codes entered by any provider type and at least 18 years of age as of 1/1/2010, and from a cohort of 36,399 individuals identified at CHOA, 1,497 cases were randomly selected, split evenly between pediatric and adult cases, and diagnoses were manually validated as previously described (Ivey et al., 2023);
3. *Emory’s Center for Heart Failure Therapies (CHFT), a legacy database with cardiac provider-entered diagnoses for individuals with cardiovascular disease seen in the CHFT and Adult Congenital Heart Center at Emory Healthcare between 1998 and 2010.* From 3,135 patients in the database, 1,213 cases with at least one CHD ICD-coded encounter between 2010 and 2019 were included.
4. *Kaiser Permanente of Georgia (KPGA):* From 8,299 patients with a CHD ICD code, entered by any provider type, 2,319 individuals were identified who met age criteria and had an encounter between 1/1/2010 and 12/31/2019. Of these, 350 were randomly sampled of which 301 who had sufficient information in clinical notes to assess CHD diagnosis during validation in their medical records.
5. *Revenue and Fiscal Affairs data from the state of South Carolina (SC):* From 11,082 individuals, a random sample of 323 cases who had CHD ICD-coded encounters between 01/01/2010 – 12/31/2019 were validated and included for analysis.

Abstractor Training—Two board certified CHD clinicians (WMB and FHR) and a trained abstractor (LCI) performed abstraction for EHC/CHOA and CHFT and trained the

abstractors at the SC and KPGA sites to record the correct CHD classification (yes/no) and the correct CHD diagnosis for a case through a previously-described iterative process that used standardized CHD classification methods and a common REDCap database (Ivey et al., 2023). For each case, abstractors determined CHD classification based on review of text clinical documents of any available records, including healthcare system-associated health network records and scanned outside records. “Not CHD” was defined in the medical record documentation as the absence of heart defects recorded in imaging or provider notes. Cases with insufficient documentation to determine either the presence or absence of CHD (such as no encounter notes) were excluded. The team reviewed cases weekly for accuracy.

Feature Generation

Each case had encounter level data from 2010–2019, excluding identifiers and geographic variables, which were used for the ML component. Initially, a set of features (independent variables), totaling 33,460, was generated for each patient by summarizing only their ICD-9-CM codes, ICD-10-CM codes, and CPT codes across all their medical visits, the specifics of which are noted in Supplemental Table 2, and the relevant features are captured in Figures 2 and 3. Demographics were excluded to ensure data deidentification and to improve generalizability of the ML model to other datasets. Since racial and ethnicity data are often missing or unavailable in administrative datasets, inclusion of these in the machine learning model development was also omitted as to not further hinder generalizability to other datasets. In sum, features included aimed to capture a comprehensive view of a patient’s medical history and healthcare utilization, ensuring the model could effectively identify CHD.

For example, racial and ethnicity data are often missing or unavailable in administrative datasets, and inclusion in the machine learning model development would hinder generalizability to other datasets. From 1/1/2010 to 9/30/2015, all healthcare systems involved used ICD-9-CM diagnosis codes. Subsequently, from 10/1/2015 to 12/31/2019, ICD-10-CM diagnosis codes were used. Healthcare utilization variables were calculated at the patient level by counting the number of encounters of specific types a patient had over the ten-year observation period (2010 to 2019), with days as the primary unit of measurement. To aggregate counts of diagnoses, symptoms, as well as comorbidities and complications, three established diagnostic classification system schemes were used: the Healthcare Cost and Utilization Project’s (HCUP) Clinical Classification Software (CCS) for ICD-9-CM, Clinical Classifications Software Refined (CCSR) for ICD-10-CM, and CCS for Services and Procedures for procedural ICD codes. HCUP software categorizes codes by clinical body systems, and by three levels of increasing granularity. For this analysis, we looked at all categories, and at all levels of classification, as the optimal scheme for categorizing patient diagnoses and procedures was not predetermined. Instead, all considered CCS schemes and the resulting features were incorporated into the dataset for the purpose of training, testing, and developing algorithms to ascertain which groupings were most effective in identifying patients with CHD. After categorizing the 33,460 variables using CCS, CCSR, and CCS-Services and Procedures, the final analytical dataset consisted of 2,105 features, comprising 1,132 CCS codes, 753 CCSR categorical codes, and 220 CCS-Services and Procedures codes, and was de-identified for ML training and validation.

Feature Selection

Feature selection was applied to identify the subset of features that are most effective at distinguishing TP that have CHD from FP who do not. Initially, a random search method and a 5-fold cross-validation strategy were employed. The dataset was divided into 5 non-overlapping splits, with training taking place on 4 splits and testing on 1 split. This process was repeated five times to ensure that each split was included in both the training and testing datasets. The pooled data from all four sites were used to optimize the XGBoost hyperparameters for the set of 2,105 features. Following this, XGBoost was applied to the data from each site to assess feature importance. Specifically, the leave-one-site-out strategy was utilized, where each site is used once as the test data, and the three remaining sites are used for model training. Feature importance scores were derived through gradient boosting after constructing boosted trees, serving as an indicator of the utility of each feature within the model. Features with an importance contribution score of less than 0.1%, as determined by XGBoost feature importance evaluations, were excluded for that site. Of the remaining features, those that were common to two or more sites were selected and retained for algorithm development, resulting in a total of 137 features. *Shapley Additive exPlanation* (SHAP) values measure the impact of selected features on CHD identification (Lundberg & Lee, 2017). Using combined data from all four sites, SHAP values for the 137 features were calculated and summary data for the top 10 most impactful features (having the highest mean absolute SHAP values) for CHD identification presented.

Algorithm Development

Model Development and Cross-validation

After feature selection (Figure 1a), in preparation for model development, three variations of a leave-one-site-out strategy for splitting the data into training and testing datasets for each site were evaluated: 1) training with data from all sites except one, testing on data from the remaining; 2) training with data from three sites and 20% of cases from the remaining site, testing on the other 80% of samples from the remaining site; and 3) training with data from three sites and 200 stratified cases from the remaining site, testing with the rest of the cases from the remaining site. These three approaches were compared to determine the most effective train and test splitting method that optimize the trade-off between maximizing sensitivity and minimizing the FP rate, or the area under the receiver-operating curve (AUROC), for each site. The final, optimal approach for each site was dependent on the size of the site's dataset (Figure 1b). In summary, if 20% of the remaining site's data equated to more than 200 cases (i.e., it is a "large" dataset per Figure 1b), then 20% of the cases would be allocated for training. Conversely, if 20% of the remaining site's data equated to 200 cases or less (i.e., it is a "small" dataset per Figure 1b), then exactly 200 cases from the remaining site would be designated for training. Ultimately, each site's training dataset had at least 200 cases, ensuring an adequate number of cases for training while also preserving enough for testing and validation.

XGBoost was selected for model development because it outperformed other algorithms (i.e., logistic regression, Gaussian Naive Bayes, and Random Forest) in our previous related work (Shi et al., 2023). Once cases were divided into training and testing datasets, a

5-fold cross-validation was conducted on each training dataset to refine the XGBoost model parameters, with a specific emphasis on optimizing AUROC to mitigate the model's bias and improve overall performance. Then, each model was tested on the testing data from the remaining site to estimate performance metrics, including AUROC, PPV, negative predictive value (NPV), sensitivity, specificity, and F1-score (which balances PPV with sensitivity). Medians and corresponding 95% confidence intervals (95% CI) for performance metrics were calculated across the 5 models resulting from cross validation for each site separately. This approach ensured a reliable and unbiased estimation of performance across the sites. By implementing a 5-fold cross-validation within a leave-one-site-out strategy, the algorithm's generalizability to discriminate between TP and FP CHD cases when used on data from entirely new sites was evaluated (i.e., sites that were not originally involved in the model's development), and insights into the statistical similarities and differences between the TP CHD and FP CHD populations in the analyzed sites was conducted.

To calculate performance metrics for all four sites combined, the entire dataset underwent evaluation through the training of 20% stratified samples from each site, followed by testing the remaining 80% of samples from those sites. A 5-fold cross-validation was then employed on the training data to refine XGBoost model parameters, still focusing on optimizing AUROC. Medians and the corresponding (95% CI) values were again used to summarize performance metrics across the 5 folds of cross validation for the entire dataset. Metrics from models with thresholds set at 0.5 were reported. Using a threshold of 0.5 in a binary classification model is motivated by its symmetry around the probability scale, which treats both positive and negative classes as equally likely. Additionally, it assumes that the cost of misclassifying a positive instance as negative (FN) should be equal to the cost of misclassifying a negative instance as positive (FP). Therefore, this threshold is commonly used as the default choice initially, especially when there is no specific preference or imbalance between the classes, and when equal weight is assigned to both types of errors. However, given our interest in prioritizing a higher PPV, we re-ran the models while setting the thresholds such that PPV = 95% instead.

Results

Sample Demographics

Table 1 shows sample characteristics, overall and by CHD classification, for each of the four sites. The analytic sample consisted of 3,334 total patients (48.9% male, mean age of the entire sample 27.7 years (standard deviation [SD] $\pm$ 18.7 years), across four sites, with a PPV of 71.6% (2,387/3,334) based on ICD-9-CM and ICD-10-CM code classification. Mean age at FQE for TP was 26.8 (17.53) and 30.1 (21.1) for FP ($p < 0.001$). The majority of the sample were White ($n = 2,084$; 62.4% overall, 66.5% excluding those with unknown race), non-Hispanic ($n = 1,929$; 57.9% overall, 92.5% excluding those with unknown ethnicity). Site datasets varied by race, ethnicity, and age group distributions, as well as by CHD severity group as identified from medical record abstraction and by PPV of CHD ICD codes documented for each case ($p < 0.001$), a significant difference that persisted with exclusion of unknown race and ethnicity.

Feature Importance

Figure 2 presents the mean absolute SHAP values for the top ten features with the most influence on CHD identification across data from all sites. This bar plot provides a visual representation of the extent to which each selected feature contributes to the identification of CHD, with features arranged from the most influential to the least influential. Relevant features most contributing to CHD classification included: having ICD-9-CM diagnosis codes belonging to the CCS categories for other congenital anomalies of the heart, atrial septal defect, congenital insufficiency of aortic valve, hypertension, factors influencing healthcare, cardiac dysrhythmias, ventricular septal defect, and congenital anomalies group and having ICD-10-CM diagnosis codes belonging to the CCSR categories for cardiac and circulatory congenital anomalies.

Figure 3 shows more detailed SHAP summary plots of the top ten features across data from all sites. SHAP plots for each site separately are shown in Supplemental Figures 1a–1d. Each feature is denoted along the y-axis, while the x-axis displays the SHAP values. SHAP values closer to 0 indicate that the instance has a minimal impact on the classification as a TP or FP; SHAP values closer to +1 suggest a greater influence towards classification as a TP; whereas SHAP values closer to –1 suggest a greater influence towards classification as a FP. The blue to red color range represents the value of the feature for each person. For example, a case with 0 encounters with ICD9/10 codes belonging to the “Cardiac Anomalies” CCS category is represented by a blue dot, a case with 5 encounters with codes belonging to the “Cardiac Anomalies” CCS category is shown as a shade of purple, and a case with 691 encounters with codes belonging to the “Cardiac Anomalies” CCS category (i.e., the maximum) is depicted by a red dot. The model more frequently classified a case as a TP when the case had a higher number of encounters with codes belonging to the CCSR category for “Cardiac and Circulatory Congenital Anomalies” or any of the following CCS categories: other congenital anomalies of the heart, congenital insufficiency of aortic valve, ventricular septal defect, or congenital anomalies. On the other hand, the model more often classified a case as FP when the case had a higher number of encounters with codes belonging to these CCS categories: atrial septal defect, hypertension, factors Influencing healthcare, and cardiac dysrhythmias.

Model Performance

In Supplemental Figure 2, stacked percentage bar graphs illustrate PPVs and NPVs at four different sites for all three variations of the leave-one-site-out strategy. The initial PPVs and NPVs for ICD codes alone are provided as the baseline for comparison. With initial “large” sample sizes of 1,213 and 1,497, respectively, the optimal leave-one-site-out strategy for sites CHFT and EHC/CHOA involved using 80% of the site dataset for testing and the remaining 20% with three other sites for training; the optimized AUROCs are depicted in Figure 4. In doing so, the PPV for CHFT increased from 84.1% to 92.1% and the PPV for EHC/CHOA increased from 68.1% to 81.5% (Supplemental Figure 2). With “small” sample sizes of 301 and 387, the optimal leave-one-site-out strategy for sites KPGA and SC involved adding (n-200) samples of the site dataset for testing and the remaining 200 with three other sites for training; the optimized AUROCs are depicted in Figure 4. In doing so,

PPV for KPGA increased from 53.2% to 67.2% and PPV for SC increased from 58.2% to 75.3% (Supplemental Figure 2).

Finally, after training on a dataset combining 20% of samples from each site and testing on a dataset combining 80% of cases from each site, each of the five models output PPV-FN rate (FNR) curves with confidence intervals, summarized as an average in Figure 5. If XGBoost model thresholds were set at 0.5, the resulting average PPV was 84.4% (an improvement from the baseline PPV of 71.6%) and FNR was 9.4%. When changing the threshold to PPV=95%, the corresponding FNR equaled 33%; this is the optimal operating point as denoted by the red dot on Figure 5. Further, increases in PPV are associated with exponentially higher FNR; thus this point and the curve represents a desirable trade-off between PPV and false negative classification.

The medians of performance metrics across the five folds for each site and for the entire dataset are presented in Table 2. As described above, the median PPV in each test dataset ranged from a high of 92% in the models for site CHFT to a low of 67% in the models for site KPGA; the median NPV ranged from 63% for CHFT to 74% for SC; and the F1-score, a metric that combines precision (or PPV) and recall (or sensitivity), ranged from 70% for site KPGA to 94% for site CHFT. We include median metrics from the models with thresholds set at 0.5 in the table, but when thresholds were set such that PPV=95%, median NPV was 54% and median F1-score was 80%. With this optimized XGBoost model, all PPVs exhibited an improvement of more than 10 percentages points compared to the baseline PPV values using ICD codes alone.

Discussion

The findings from this study are clinically relevant for CHD surveillance as FP cases in a dataset may lead to inaccurate conclusions. For example, if a CHD code for secundum atrial septal defect/patent foramen ovale is used to order an echocardiogram for someone with a stroke who turns out not to have a CHD, one could draw erroneous conclusions about the prevalence of stroke in the CHD population. FN cases may lead to an incomplete picture of the population. The PPV of CHD ICD codes alone (baseline) varied from 53.2% (KPGA) to 84.0% (CHFT) across site datasets and a PPV of 71.6% for a dataset combining all sites. By developing and employing an XGBoost machine learning model for the present data, we were able to increase the PPV from 72% to 95% overall, while maintaining the NPV below 54%, and for each site, PPV improved by at least 10 percentage points. Therefore, ML revealed to be an effective tool for improving the accuracy and reducing FP, of identifying individuals with CHD in clinical and administrative data using only features based on ICD-9 or ICD-10 codes. A comprehensive set of clinical and procedural features (2010–2019 encounter-level data) was developed (Supplemental Table 2; Figures 2 and 3) to capture patients' medical histories and healthcare utilization, facilitating accurate CHD identification. The highest baseline PPV occurred for cases from a tertiary care referral system with legacy medical records with cardiologist-entered diagnoses, which perhaps lead to a more accurate diagnosis by specialists or a more complex CHD population within a tertiary referral dataset

In a previous paper, XGBoost outperformed logistic regression, Gaussian Naive Bayes, and random forest models when using clinical and administrative data to identify CHD cases (Shi et al., 2023). Using an XGBoost model and the transfer learning strategy for the current data, each site's PPV was improved by at least 10 percentage points while maintaining F1-scores between 70% and 93%, suggesting that the model can adapt to data from new sites if a fraction of validated samples are available for training. Future studies implementing this strategy to introduce new site data should ensure an appropriate balance between data added to the training set and data reserved for testing. The cutoffs for splitting the data may need to be adjusted accordingly.

An operational threshold can be chosen to emphasize improvement in PPV or the reduction of FNs. In this scenario, the PPV was set at 95% on the curve. Beyond this operational point, elevating PPV results in an exponential rise in FN rates. However, future studies using this model can choose other operation points along the PPV-FNR curve in Figure 5 to set as the thresholds depending on the trade-off between PPV and FNR that are deemed acceptable for surveillance or research purposes.

While prior research has applied ML to assess CHD risk factors and predict the risk of CHD in infants during pregnancy (Luo et al., 2017; Rani & Masood, 2020), this study follows previous work on development of a generalized ML pipeline for the detection of true CHD cases within administrative records using both ICD-9-CM and ICD-10-CM across 4 distinct validated datasets (Shi et al., 2023). In both studies, the primary aim was to create a standardized ML pipeline to distinguish individuals with true CHD from those with CHD codes but no CHD, with the ultimate goal of enhancing CHD surveillance, understanding healthcare utilizations, and improving patient outcomes. In this study, a transfer learning strategy was added to demonstrate how models can potentially adapt to data from new sites.

One other recent publication sought to improve the identification of CHD cases in administrative data through ML (Marelli et al., 2024). Marelli et al. developed a ML model using administrative data collected continuously from 1983 to 2000 for 19,187 Canadian patients with at least 1 CHD-related ICD-9 diagnosis and/or a surgical procedure documented by any clinician other than a cardiovascular specialist. Models were trained on a dataset with labels (true CHD or no CHD) that had been generated by a time-consuming, but highly accurate algorithm previously validated in the same data and developed by clinicians. By improving timeliness of CHD classification and achieving a 99% F1-score in evaluation, this study similarly concluded that ML is a useful tool to identify CHD cases in administrative data. However, generalizability of the model to U.S.-based data may present a challenge given the differences in patient access to healthcare and in the continuity of collected healthcare data when comparing a universal healthcare system (Canada) to the U.S. healthcare system. Additionally, some U.S. administrative databases may not reliably capture cardiac specialty provider type.

By leveraging diverse clinical and administrative data sources, the model improves CHD identification accuracy. Unlike previous methods, which were limited to severe cases, this approach broadens the scope of CHD detection, capturing both severe and milder forms of the condition. This broad applicability not only enhances the accuracy of CHD

detection, but also lays the groundwork for extending the pipeline to other conditions that require multi-site data integration. In addition to improving prediction accuracy, this study identifies key predictive features that contribute to accurate CHD classification, enhancing the interpretability and clinical relevance of the model. The generalized nature of the pipeline also provides a flexible framework for adaptation to other healthcare datasets, underscoring its potential for widespread utility in CHD surveillance and the detection of similar complex conditions. This work establishes a scalable, high-performance solution for CHD identification that can drive more effective CHD surveillance, facilitate early intervention, and support improved long-term care strategies for affected populations.

Traditional machine learning (ML) models were chosen for this study, despite the current popularity of advanced techniques like deep learning (DL) and artificial intelligence (AI), due to practical demands of clinical applications. First, traditional ML models offer greater interpretability, which is especially important in healthcare contexts where the transparency of diagnostic criteria is crucial. Models like XGBoost or Random Forest allow clinicians to understand the specific features influencing the classification of CHD cases, fostering trust and aiding in clinical decision-making. Additionally, the dataset used in this study, although substantial, does not reach the size or complexity that typically warrants the use of deep learning, which thrives on vast and often unstructured data. Traditional ML methods are more suitable for medium-sized, structured datasets, and algorithms like XGBoost are particularly adept at feature selection, efficiently narrowing down a large set of variables—such as reducing 2,105 Clinical Classification Software (CCS) features to 137 critical ones. This dimensionality reduction is essential for maintaining high performance without sacrificing interpretability. Finally, the objective of maximizing PPV for CHD identification aligns well with the strengths of traditional ML methods. By tuning these models to prioritize PPV while balancing the FN rate, the study achieved a PPV of 95% across data from 4 different sites, demonstrating the effectiveness of traditional ML in meeting the high-accuracy demands of clinical applications.

ML considers features in the context of other features and whether those features have a positive or negative impact on the classification, as quantified by the SHAP value. These features act as either positive or negative predictors for accurate CHD classification, with some related to more frequent classification as a TP and some as a FP. Cases were more likely to be classified as TP when the number of encounters with codes related to CCS categories for cardiac anomalies, other congenital anomalies of the heart, congenital insufficiency of aortic valve, ventricular septal defect, and congenital anomalies and CCSR category codes for circulatory congenital anomalies increased. Conversely, cases were more likely to be classified as FP as the number of encounters with CCS category codes for atrial septal defect, hypertension, factors influencing healthcare, and cardiac dysrhythmias increased. This feature reduction process prevents the model from being a ‘black box,’ where the variables influencing the algorithm’s outcomes are uncertain. By utilizing overlapping features identified as important across all datasets, the model’s generalizability is reserved and reliance on a model that performs well in one dataset, but cannot replicate results in another dataset is avoided.

Principal component analysis (PCA) was attempted on all features, but more than 60 principal components would be required to account for 90% of the total variance contribution, rendering PCA an ineffective method for feature reduction; the dataset proved to be too complex to be adequately represented by a few principal components. Therefore, an alternative approach was taken to reduce the model to the essential features specific to each site.

Limitations

The limited sample size for both SC and KGPA may be a factor contributing to the potential for further improvement in performance. Limitations in the availability and diversity of medical records for abstraction may have resulted in misclassification of some TP as FP, especially within the KPGA data since its system does not incorporate data from external records. Unlike CHFT and Emory/CHOA, the KPGA data do not draw medical record information from other data sources that the medical record abstractors use when confirming that case has CHD; thus, PPV could be underestimated for this dataset. Additionally, the KPGA dataset had the highest proportion of older cases and cases without severe CHD than the other datasets, which are both associated with lower PPV (Rodriguez et al., 2022). The SC site, with a similarly low baseline PPV of 58.2%, included data from several diverse community-based data sources. Therefore, less accurate coding by non-cardiac specialists in this sample might explain the low PPV for this dataset. The algorithm classified patients as having any CHD or not having any CHD, but did not determine PPV of individual CHD codes. Nevertheless, the PPV for all sites' cases were substantially improved from baseline. The constraint on the feature space ranking, corresponding to the number of patients, may technically limit the application of statistical feature selection algorithms like those based on eigen-analysis.

Additionally, certain features selected, such as the number of emergency department visits and a diagnosis in the respiratory system group, may exhibit variations during the coronavirus disease 2019 pandemic that could impact performance for future datasets with subsequent years of surveillance data.

Conclusion

The accuracy of administrative datasets in detecting CHD can be enhanced through the application of machine learning techniques, with potential for further improvement through the utilization of larger, validated datasets. Employing a generalized machine learning pipeline for CHD case classification within administrative data has the capacity to enhance public health surveillance endeavors, understand healthcare utilizations, and improve patient outcomes. To validate this discovery, additional machine learning studies involving diverse datasets from various locations, time periods, and patient demographics would be informative.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378 and the Medical Imaging, Informatics, and Artificial Intelligence Core (MIIAI), which is supported by the Department of Biomedical Informatics, Emory University School of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding Sources

Centers for Disease Control and Prevention Cooperative Agreement, *Congenital Heart Defects Surveillance across Time And Regions (CHD STAR)* Grant/Award Number: CDC-RFA-DD19-1902B

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]
- Almli LM, Alter CC, Russell RB, Tinker SC, Howards PP, Cragan J, Petersen E, Carrino GE, & Reefhuis J (2017). Association Between Infant Mortality Attributable to Birth Defects and Payment Source for Delivery—United States, 2011–2013. *Morbidity and Mortality Weekly Report*, 66(3), 84–87. 10.15585/mmwr.mm6603a4 [PubMed: 28125575]
- 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 (British Cardiac Society)*, 94(9), 1194–1199. 10.1136/hrt.2007.122671 [PubMed: 17646191]
- 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]
- 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]
- Hoffman JIE, & Kaplan S (2002). The incidence of congenital heart disease. *Journal of the American College of Cardiology*, 39(12), 1890–1900. 10.1016/s0735-1097(02)01886-7 [PubMed: 12084585]
- Ivey LC, Ahmad A, Chen J, Rodriguez FH, Raskind-Hood C, & Book WM (2024). Anatomic and Physiologic Classification of Adults with Congenital Heart Disease to Predict Adverse Outcomes: Use of Administrative Codes Compared to Clinical Staging. *American Heart Journal*, S0002–8703(24)00034–6. 10.1016/j.ahj.2024.02.007
- Ivey LC, Rodriguez FH, Shi H, Chong C, Chen J, Raskind-Hood CL, Downing KE, Farr SL, & Book WM (2023). Positive Predictive Value of International Classification of Diseases, Ninth Revision, Clinical Modification, and International Classification of Diseases, Tenth Revision, Clinical Modification, Codes for Identification of Congenital Heart Defects. *Journal of the American Heart Association*, 12(16), e030821. 10.1161/JAHA.123.030821 [PubMed: 37548168]
- Jill Glidewell M, Farr SL, Book WM, Botto L, Li JS, Soim AS, Downing KE, Riehle-Colarusso T, D'Ottavio AA, Feldkamp ML, Khanna AD, Raskind-Hood CL, Sommerhalter KM, & Crume TL (2021). Individuals aged 1–64 years with documented congenital heart defects at healthcare encounters, five U.S. surveillance sites, 2011–2013. *American Heart Journal*, 238, 100–108. 10.1016/j.ahj.2021.04.007 [PubMed: 33951414]

- Jone P-N, Gearhart A, Lei H, Xing F, Nahar J, Lopez -Jimenez Francisco, Diller G-P, Marelli A, Wilson L, Saidi A, Cho D, & Chang AC (2022). Artificial Intelligence in Congenital Heart Disease. *JACC: Advances*, 1(5), 100153. 10.1016/j.jacadv.2022.100153 [PubMed: 38939457]
- 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]
- Kuehl KS, Loffredo CA, & Ferencz C (1999). Failure to diagnose congenital heart disease in infancy. *Pediatrics*, 103(4 Pt 1), 743–747. 10.1542/peds.103.4.743 [PubMed: 10103296]
- Lewis MJ, Reardon LC, Aboulhosn J, Haeffele C, Chen S, Kim Y, Fuller S, Forbess L, Alshawabkeh L, Urey MA, Book WM, Rodriguez F, Menachem JN, Clark DE, Valente AM, Carazo M, Egbe A, Connolly HM, Krieger EV, ... Rosenbaum M (2023). Morbidity and Mortality in Adult Fontan Patients After Heart or Combined Heart-Liver Transplantation. *Journal of the American College of Cardiology*, 81(22), 2161–2171. 10.1016/j.jacc.2023.03.422 [PubMed: 37257951]
- 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]
- Lui GK, McGarry C, Bhatt A, Book W, Riehle-Colarusso TJ, Dunn JE, Glidewell J, Gurvitz M, Hoffman T, Hogue CJ, Hsu D, Obenhaus S, Raskind-Hood C, Rodriguez FH, Zaidi A, & Zutphen ARV (2019). Surveillance of Congenital Heart Defects among Adolescents at Three U.S. Sites. *American Journal of Cardiology*, 124(1), 137–143. 10.1016/j.amjcard.2019.03.044 [PubMed: 31030970]
- Lui GK, Sommerhalter K, Xi Y, Botto LD, Crume T, Farr S, Feldkamp ML, Glidewell J, Hsu D, Khanna A, Krikov S, Li J, Raskind-Hood C, Sarno L, Van Zutphen AR, Zaidi A, Soim A, & Book WM (2022). Health Care Usage Among Adolescents With Congenital Heart Defects at 5 Sites in the United States, 2011 to 2013. *Journal of the American Heart Association*, 11(18), e026172. 10.1161/JAHA.122.026172 [PubMed: 36102252]
- Lundberg S, & Lee S-I (2017). A Unified Approach to Interpreting Model Predictions (arXiv:1705.07874). arXiv 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]
- Mahle WT, Newburger JW, Matherne GP, Smith FC, Hoke TR, Koppel R, Gidding SS, Beekman RH, Grosse SD, American Heart Association Congenital Heart Defects Committee of the Council on Cardiovascular Disease in the Young, Council on Cardiovascular Nursing, and Interdisciplinary Council on Quality of Care and Outcomes Research, & American Academy of Pediatrics Section on Cardiology and Cardiac Surgery, and Committee on Fetus and Newborn. (2009). Role of pulse oximetry in examining newborns for congenital heart disease: A scientific statement from the American Heart Association and American Academy of Pediatrics. *Circulation*, 120(5), 447–458. 10.1161/CIRCULATIONAHA.109.192576 [PubMed: 19581492]
- Mai CT, Isenburg JL, Canfield MA, Meyer RE, Correa A, Alverson CJ, Lupo PJ, Riehle-Colarusso T, Cho SJ, Aggarwal D, Kirby RS, & Network NBDP (2019). National population-based estimates for major birth defects, 2010–2014. *Birth Defects Research*, 111(18), 1420–1435. 10.1002/bdr2.1589 [PubMed: 31580536]
- Marelli AJ, Li C, Liu A, Nguyen H, Moroz H, Brophy JM, Guo L, Buckeridge DL, Tang J, Yang AY, & Li Y (2024). Machine Learning Informed Diagnosis for Congenital Heart Disease in Large Claims Data Source. *JACC: Advances*, 3(2), 100801. 10.1016/j.jacadv.2023.100801 [PubMed: 38939385]
- 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]
- Menachem JN, Schlendorf KH, Mazurek JA, Bichell DP, Brinkley DM, Frischhertz BP, Mettler BA, Shah AS, Zalawadiya S, Book W, & Lindenfeld J (2020). Advanced Heart Failure in Adults With Congenital Heart Disease. *JACC: Heart Failure*, 8(2), 87–99. 10.1016/j.jchf.2019.08.012 [PubMed: 31838031]

- 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]
- 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), e1502–1508. 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
- Reller MD, Strickland MJ, Riehle-Colarusso T, Mahle WT, & Correa A (2008). Prevalence of Congenital Heart Defects in Metropolitan Atlanta, 1998–2005. *The Journal of Pediatrics*, 153(6), 807–813. 10.1016/j.jpeds.2008.05.059 [PubMed: 18657826]
- 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]
- Shi H, Book W, Raskind-Hood C, Downing KF, Farr SL, Bell MN, Sameni R, Rodriguez FH III, & Kamaleswaran R (2023). A machine learning model for predicting congenital heart defects from administrative data. *Birth Defects Research*, n/a(n/a). 10.1002/bdr2.2245
- 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). 10.1161/CIR.0000000000000602
- 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]
- van der Bom T, Zomer AC, Zwinderman AH, Meijboom FJ, Bouma BJ, & Mulder BJM (2011). The changing epidemiology of congenital heart disease. *Nature Reviews. Cardiology*, 8(1), 50–60. 10.1038/nrcardio.2010.166 [PubMed: 21045784]
- 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]
- 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 (London, England)*, 359(9311), 1019–1025. 10.1016/s0140-6736(02)08092-3 [PubMed: 11937181]

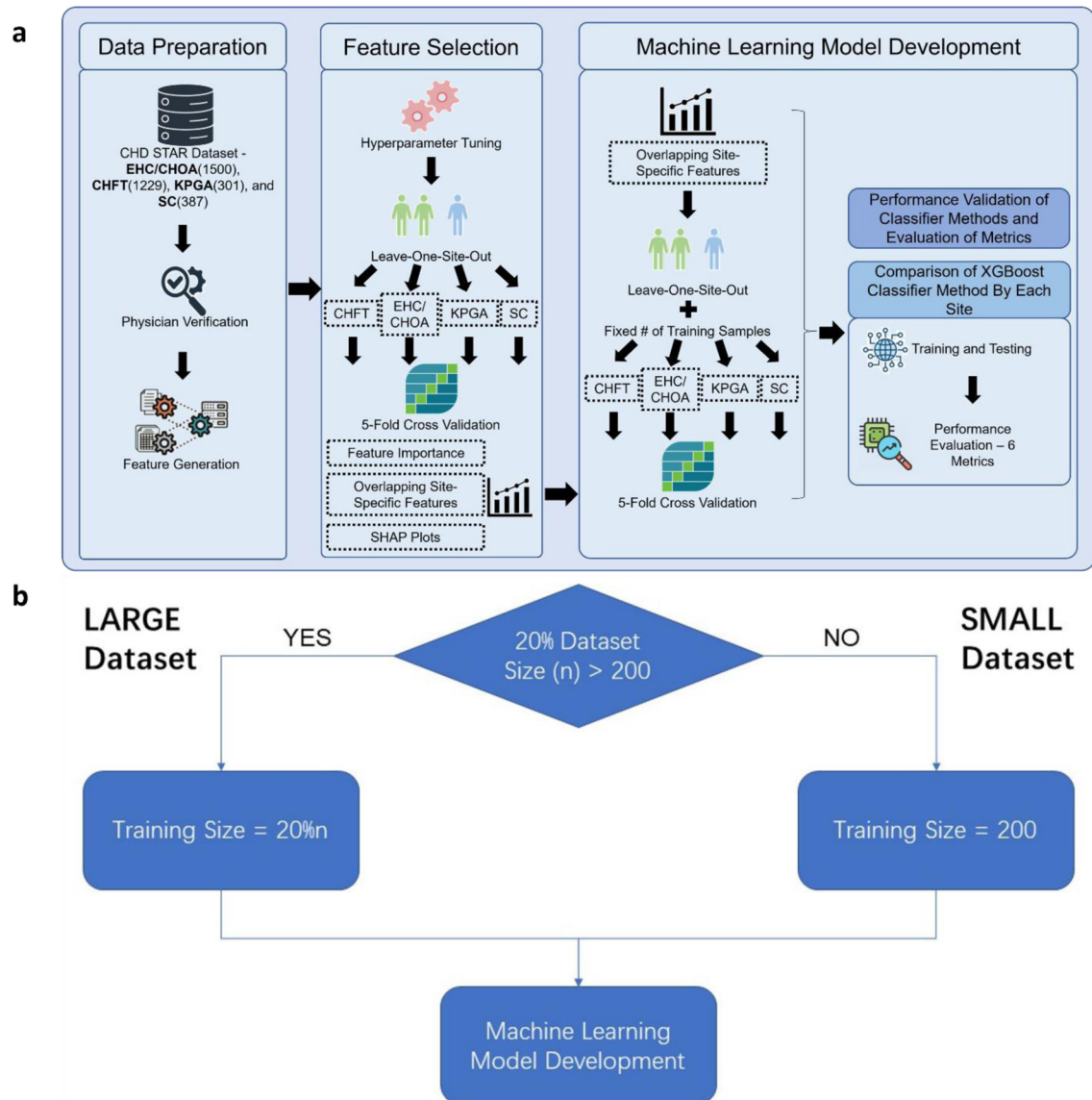


Figure 1. Workflow pipeline to develop a machine learning model and decide on the size of training/testing datasets for data from four datasets to improve the identification of people with CHD in clinical and administrative data

Notes. a., This is the depiction of workflow pipeline implemented to develop a generalized machine learning model that improves the identification of people with CHD in clinical and administrative data from five sites, including feature generation, feature selection, model development and cross-validation procedures. b., This serves as an illustration of how we decided on the size (i.e. number of observations) of the training dataset, dependent on the number of observations in the overall dataset for each site. Site datasets ranged from 301 to 1497 samples. The number of samples reserved for training was either 20% of dataset or 200 samples, whichever resulted in the larger number for training. The selection of 200 as the minimum number of training samples ensures a sufficient number of samples for training while still retaining enough observations for testing and validation.

Acronym: **CHD**=Congenital Heart Defects; **EHC/CHOA**=Emory Healthcare/Children's Healthcare of Atlanta; **KPGA**=Kaiser Permanente of Georgia **SC**=South Carolina, SHAP=**S**hapley **A**dditive **E**xplanation values generated using XGBoost.

6 Performance Metrics:

- **AUROC** = area under the receiver operating characteristic curve
- **PPV** (positive predictive value) = $100 \times TP / (TP + FP)$
- **NPV** (negative predictive value) = $100 \times TN / (TN + FN)$
- **Sensitivity** = $100 \times TP / (TP + FN)$
- **Specificity** = $100 \times TN / (TN + FP)$
- **F1 score** = $2 \times (PPV \times Sensitivity) / (PPV + Sensitivity)$

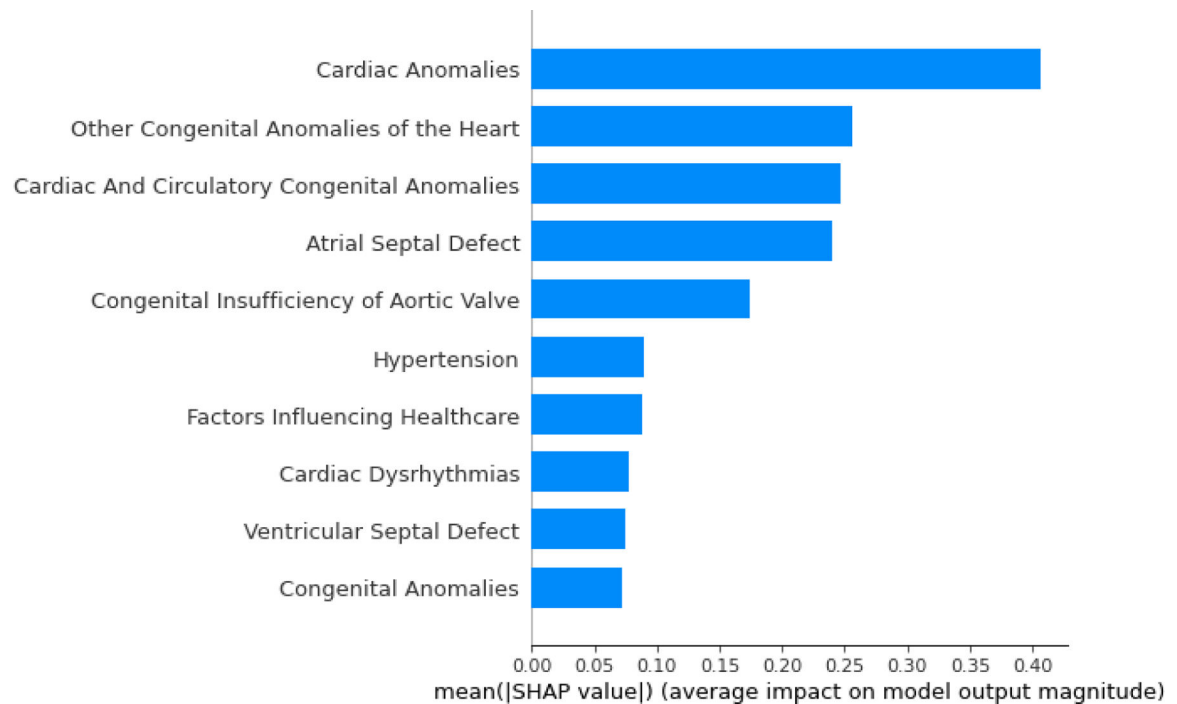


Figure 2. Bar Plot of Mean Absolute SHAP Values for Top 10 Most Impactful Features for Congenital Heart Defect (CHD) Identification Using the XGBoost Model

Notes. This bar plot shows the mean absolute SHAP values for the 10 most impactful features (variables) with the highest mean absolute SHAP values 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 of having or not having a CHD. Features with higher mean absolute SHAP values like having ICD-9-CM and ICD-10-CM diagnosis codes belonging to the CCS/CCSR categories for cardiac anomalies, other congenital anomalies of the heart, and cardiac and circulatory congenital anomalies are most influential in distinguishing between cases with CHD and cases without CHD. Other features that influence prediction include several CCS categories: atrial septal defect, congenital insufficiency of aortic valve, hypertension, factors influencing healthcare, cardiac dysrhythmias, ventricular septal defect, and congenital anomalies comorbidity groups.

Acronym: CHD=Congenital Heart Defects; SHAP=SHapley Additive exPlanation values.

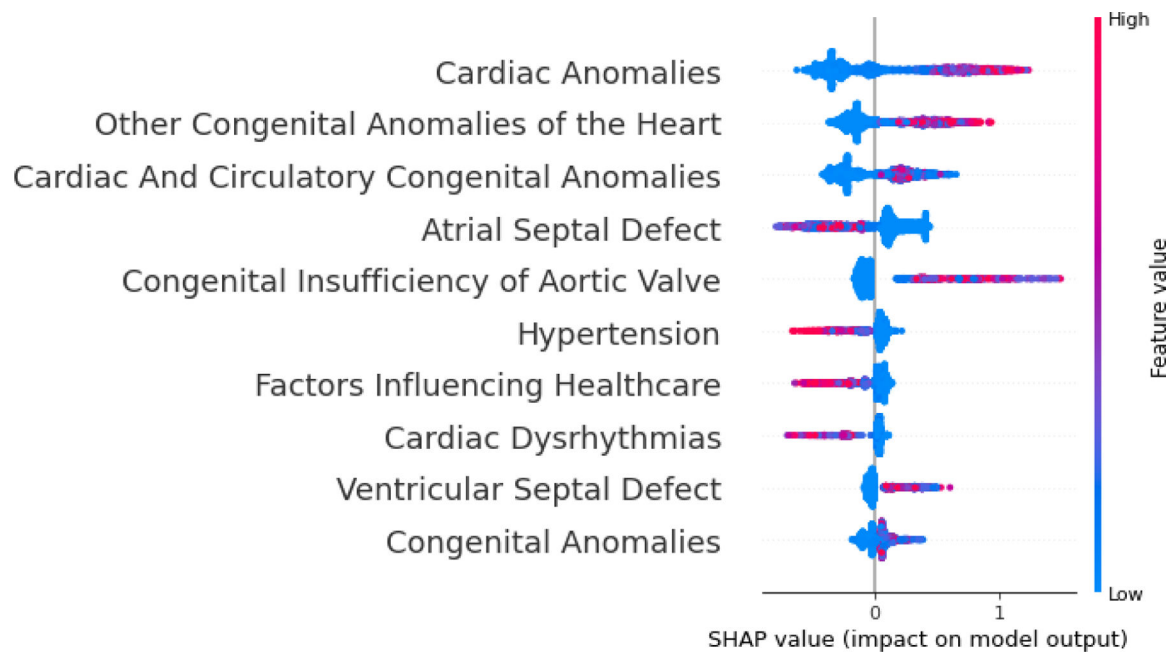


Figure 3. SHAP Summary Plot of Top 10 Common Relevant Features for Congenital Heart Defect (CHD) Prediction Using the XGBoost Model at Four Sites

Notes. This figure shows the SHAP summary plot of the ten common relevant features across the pooled data from all four sites using the XGBoost model for CHD prediction. The color bars represent raw SHAP values for each feature. Each dot on the plot represents the SHAP value of that feature for one of the individuals in the dataset. The blue to red color range represents the value of the feature for each person. Features more indicative of true CHD, including ICD-9-CM and ICD-10-CM codes belonging to CCS categories for cardiac anomalies, other congenital anomalies of the heart, congenital insufficiency of aortic valve, and ventricular septal defect; and ICD-10-CM codes belonging to CCSR categories for cardiac and circulatory congenital anomalies are denoted by SHAP values whose bars have red portions to the right of '0', whereas features more indicative of no CHD, including ICD-9-CM and ICD-10-CM codes belonging to CCS categories for atrial septal defect, hypertension, factors influencing healthcare, and cardiac dysrhythmias comorbidity groups, are denoted by SHAP values whose bars have red portions to the left of '0'. Each feature's numerical value signifies the frequency of a particular CCS/CCSR code per patient. For example, if a case has a value of 0 for the cardiac anomalies CCS code, it signifies that the patient experienced 0 encounters with ICD9/10 codes related to cardiac anomalies.

Acronyms: CHD=Congenital Heart Defects; SHAP= SHapley Additive exPlanation values; SC=South Carolina.

XGBoost AUROC at 4 Different Sites

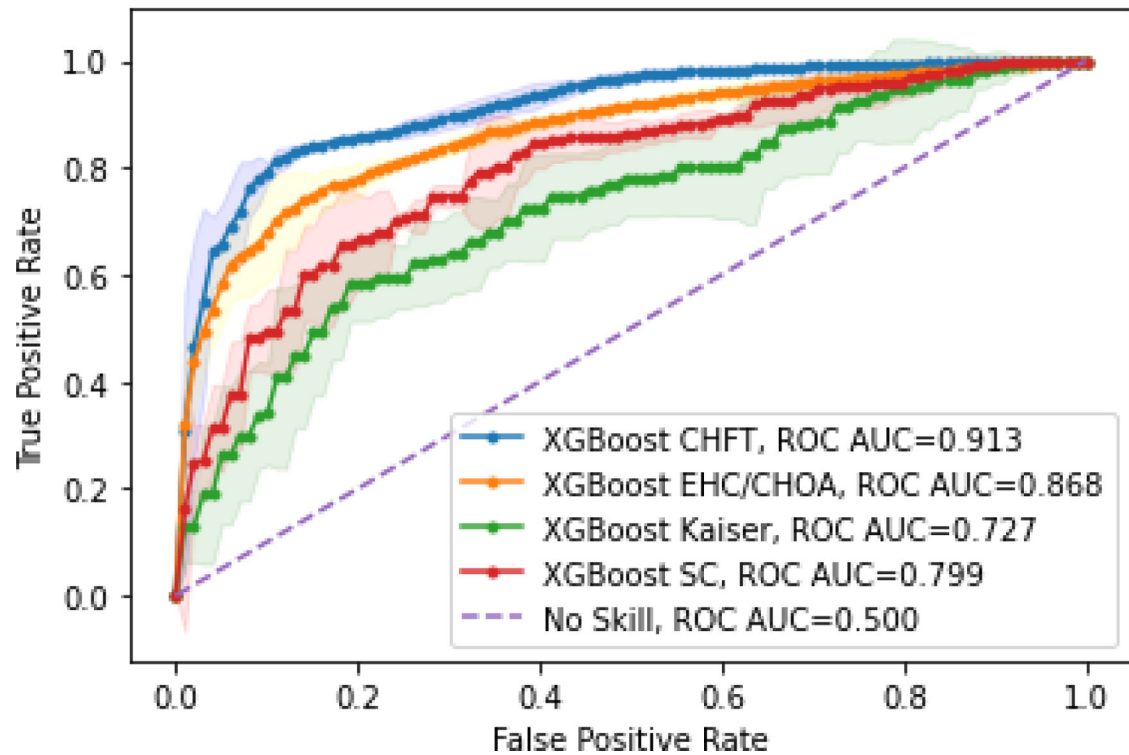


Figure 4. Receiver Operating Curve (ROC) Analyses with Area Under the Receiver Operating Curve (AUROC) Values Tested at Four Different Sites

Notes. This figure shows the ROC curves of machine learning models at four different sites for CHD prediction. XGBoost model at CHFT site has the highest AUROC, 0.913, compared to other models.

Acronyms: ROC=Receiver Operating Curve; AUROC=Area Under the Receiver Operating Curve.

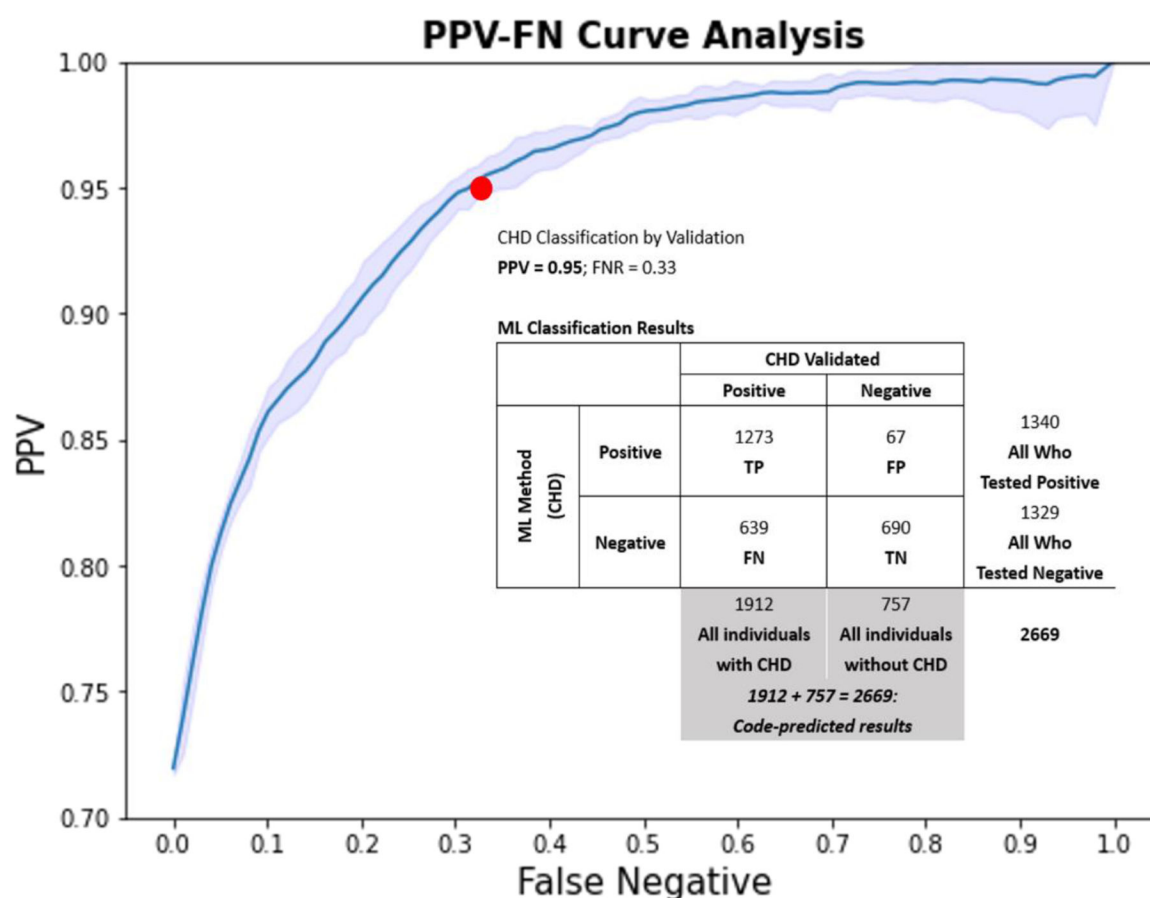


Figure 5. Positive Predictive Value (PPV)-False Negative (FN) Rate Analyses for the XGBoost Model for Four Sites

Notes. This figure shows the average PPV-FN rate curve for CHD prediction for pooled data across four sites (CHFT-EHC/CHOA-KPGA-SC). The threshold of PPV=0.95 and FN rate=0.33 was chosen as the operating point, as seen by the red dot. Beyond the operating point, additional increases in PPV are associated with exponential increases in FN rates. The dark blue line shows the mean of the 5-fold cross validation performances, while the light blue shaded area represents the range of values within 2 SD of the mean. The operating point is selected from one of the five folds which had the median false negative rate, thus it does not appear on the dark blue line.

Acronyms: **PPV**=Positive Predictive Value; **FN**=False Negative; **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** = $100 \times \text{TP}/(\text{TP}+\text{FP})$ $100 \times 1273/(1273+67) = 95\%$
- **NPV** = $100 \times \text{TN}/(\text{TN}+\text{FN})$ $100 \times 690/(690+639) = 52\%$
- **Accuracy** = $100 \times (\text{TP}+\text{TN})/N$ $100 \times (1273+690)/2669 = 74\%$
- **Sensitivity** = $100 \times \text{TP}/(\text{TP}+\text{FN})$ $100 \times 1273/(1273+639) = 67\%$
- **Specificity** = $100 \times \text{TN}/(\text{TN}+\text{FP})$ $100 \times 690/(690+67) = 91\%$

○ **CHD prevalence** = $100 \times (TP + FN) / N$ 100 X (1273+639)/2669 = 72%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Descriptive characteristics for cases with at least 1 CHD ICD-9-CM or ICD-10-CM code^{*} with an encounter between 2010–2019 in 4 datasets

CHARACTERISTICS	CHFT	Combined EHC + CHOA Validation Set	KPGA	SC Dataset	Overall	$\chi^{2†}$
	N= 1213	N=1497	N=301	N=323	N= 3334	
Sex [‡]						
Females	629 (51.9%)	758 (50.7%)	150 (49.8%)	165 (51.1%)	1702 (51.1%)	P = 0.9
Males	584 (48.1%)	738 (49.3)	151 (50.2%)	158 (48.9%)	1631 (48.9%)	
Race						
Black	260 (21.4%)	413 (27.6%)	106 (35.2%)	95 (29.4%)	874 (26.2%)	<0.001
White	842 (69.4%)	910 (60.8%)	129 (42.9%)	201 (62.2%)	2082 (62.5%)	
Other [§]	23 (1.9%)	76 (5.2%)	29 (9.6%)	17 (5.3%)	145 (4.4%)	
Unknown	88 (7.3%)	96 (6.4%)	37 (12.3%)	10 (3.1%)	231 (6.9%)	
Ethnicity						
Hispanic or Latino	27 (2.2%)	116 (7.7%)	14 (4.6%)	10 (3.1%)	167 (5.0%)	<0.001
Non-Hispanic or Latino	727 (59.9%)	1202 (80.3%)	0 (0%)	0 (0%)	1929 (57.9%)	
Unknown	459 (37.9%)	179 (12.0%)	287 (95.4%)	313 (96.9%)	1238 (37.1%)	
Age Group (at first encounter)						
0–10 years	0 (0%)	685 (45.8%)	0 (0%)	126 (39.0%)	813 (25.1%)	<0.001
11–19 years	24 (2.2%)	74 (4.9%)	51 (16.9%)	50 (15.5%)	199 (6.2%)	
20–40 years	815 (72.8%)	349 (23.3%)	101 (33.6%)	71 (22.0%)	1336 (41.2%)	
41+ years	280 (25.0%)	389 (26.0%)	147 (48.8%)	76 (23.5%)	892 (27.5%)	
CHD Native Anatomy Severity Group						
Severe	585 (48.2%)	249 (16.6%)	22 (7.3%)	58 (18.0%)	855 (25.6%)	<0.001
Non-Severe	434 (35.8%)	771 (51.5%)	138 (45.8%)	130 (40.2%)	1532 (46.0%)	
No CHD	194 (16.0%)	477 (31.9%)	141 (46.9%)	135 (41.8%)	947 (28.4%)	
PPV of ICD codes for CHD						
Overall codes	84.0% (1019/1213)	68.1% (1020/1497)	53.2% (160/301)	58.2% (188/323)	71.6% (2387/3334)	<0.001
Severe codes	98.8% (585/592)	95.8% (249/260)	71.0% (22/31)	98.3% (58/59)	97.0% (914/942)	<0.001
Non-severe codes [¶]	69.9% (434/621)	62.3% (771/1237)	51.1% (138/270)	49.3% (130/264)	61.6% (1473/2392)	<0.001

Abbreviations: CHD= Congenital Heart Disease, ICD = International Classification of Diseases, CHFT=Center for Heart Failure Therapies, EHC=Emory Healthcare, CHOA= Children's Healthcare of Atlanta, SC=South Carolina Department of Public Health, KPGA= Kaiser Permanente of Georgia, PPV= Positive Predictive Value, defined as having any CHD when a CHD ICD-9 or ICD-10 code for a CHD was present.

^{*} 87 codes in ICD-9-CM code group 745.xx – 747.xx and ICD-10-CM code group Q20.x – Q26.x (Table S1).

[†] p-values calculated using Pearson's Chi-squared test comparing four data sources.

[‡] 1 case was missing sex in CHFT and the validation dataset.

[§] "Other race" includes American Indian/Alaskan Native, Asian, Native Hawaiian or Other Pacific Islander, and multi-racial.

// 94 CHFT cases with invalid encounter dates were excluded from age group calculations.

// Cases having both severe and non-severe categorized as severe.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.
Performance Validation of XGBoost Machine Learning Models Across Four Datasets *

Dataset	AUROC	PPV	NPV	Sensitivity	Specificity	F1-score
CHFT	0.91 [0.91,0.92]	0.92 [0.92,0.95]	0.63 [0.51,0.73]	0.94 [0.87,0.97]	0.58 [0.53,0.73]	0.93 [0.91,0.94]
EHC/CHOA	0.87 [0.86,0.88]	0.82 [0.81,0.82]	0.72 [0.71,0.73]	0.89 [0.89,0.90]	0.57 [0.56,0.59]	0.85 [0.85,0.86]
KPGA	0.73 [0.69,0.77]	0.67 [0.63,0.71]	0.65 [0.62,0.69]	0.74 [0.69,0.78]	0.57 [0.49,0.67]	0.70 [0.67,0.72]
SC	0.80 [0.79,0.81]	0.75 [0.74,0.78]	0.74 [0.67,0.78]	0.85 [0.79,0.88]	0.63 [0.60,0.65]	0.80 [0.77,0.82]
entire dataset [†]	0.88 [0.87,0.88]	0.84 [0.84,0.85]	0.72 [0.70,0.73]	0.91 [0.90,0.92]	0.56 [0.55,0.59]	0.87 [0.87,0.88]
Entire dataset optimized for ppv [‡]	0.88 [0.88,0.89]	0.95 [0.95,0.95]	0.54 [0.52,0.55]	0.69 [0.67,0.71]	0.91 [0.91,0.91]	0.80 [0.87,0.88]

* Validation EHC/CHOA (1500 cases), Center for Heart Failure therapies (CHFT), South Carolina (SC), Kaiser Permanente of Georgia (KPGA).

[†] Machine learning performance for the entire dataset using the default thresholds at 0.5.

[‡] Machine learning performance for the entire dataset prioritizing PPV by setting the thresholds such that PPV is set to 95%.

Note. Four machine learning models conducted with 5-fold cross-validation across four sites; median and 95% confidence intervals (95%CI) displayed for six performance metrics.

Acronyms: AUROC=Area Under the Receiver Operating Characteristics; **PPV**=Positive Predictive Value; **NPV** = Negative Predictive Value; **CI**=Confidence Interval, CHFT =Center for Heart Failure Therapies, EHC= Emory Healthcare, CHOA= Children's Healthcare of Atlanta, KPGA= Kaiser Permanente of Georgia, SC= South Carolina Department of Public Health.