ACHD Risk Score: A tool for identifying adults with moderate or complex congenital heart defects using Electronic Health Records data from the Emory Healthcare Data Warehouse, Atlanta, GA

By

Alpha Oumar Diallo
Master in Public Health

Global Epidemiology

____
Mohammed Ali, MBChB, MSC, MBA
Committee Chair

____
Asha Krishnaswamy, MSc, B.EE
Committee Member

ACHD Risk Score: A tool for identifying adults with moderate or complex congenital heart defects using Electronic Health Records data from the Emory Healthcare Data Warehouse, Atlanta, GA

By

Alpha Oumar Diallo

B.A.
Carleton College, Northfield, MN
2012

Thesis Committee Chair: Mohammed Ali, MBChB, MSC, MBA

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology 2015

Abstract

ACHD Risk Score: A tool for identifying adults with moderate or complex congenital heart defects using Electronic Health Records data from the Emory Healthcare Data Warehouse, Atlanta, GA

By Alpha Oumar Diallo

Background: It is hypothesized that adults with moderate to complex congenital heart defects (CHD) are increasing, but many patients experience lapses in specialist care. There is, to date, no validated procedure to identify this population.

Objectives: To develop and validate a risk score to identify adults aged 20-60 years old with moderate to complex CHD from routine provider and health system electronic health records (EHR).

Methods: We used a case-control design (596 adults with physician-diagnosed moderate to complex CHDs receiving care at Emory's adult CHD clinic and 2,384 controls [persons without ICD-9 codes for CHD] receiving care at other Emory facilities]). We extracted data regarding age, race/ethnicity, EKG, and laboratory tests from routine outpatient visits between January 2009 and December 2012 from Emory Healthcare's EHR Data Warehouse. We used multivariable logistic regression models and a split-sample (4:1 ratio) approach to develop and validate the risk score, respectively. We generated receiver operating characteristic (ROC) curves to assess the ability of models to predict adult moderate to complex CHD.

Results: Three models (laboratory, non-laboratory, and simplified) were produced and validated internally. The non-laboratory algorithm (ACHD model) based on age, sex, and electrocardiogram markers was chosen. Validation studies of the ACHD model showed a ROC c-statistic of 0.97 [95% Confidence interval (CI): 0.95, 0.99]. The ACHD Risk Score, developed using the ACHD model, also demonstrated good accuracy with 93.69% sensitivity and positive predictive value of 69.80% at a score threshold of 11.

Conclusion: A simple non-laboratory risk score based on age, sex, and EKG marker may help accurately identify adults with moderate to complex CHD from routine EHR systems. External validation studies within large longitudinal clinical cohorts are required to assess wider performance of this tool.

ACHD Risk Score: A tool for identifying adults with moderate or complex congenital heart defects using Electronic Health Records data from the Emory Healthcare Data Warehouse, Atlanta, GA

By

Alpha Oumar Diallo

B.S.
Carleton College, Northfield, MN
2012

Thesis Committee Chair: Mohammed Ali, MBChB, MSC, MBA

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2015

Acknowledgements

I am grateful to Dr. Mohamed Ali, MBChB, MSC, MBA, my committee chair, for his guidance and genuine support but most importantly for allowing to be a late addition to his cohort of advisees. His flexibility encouraged me to be accountable for my work, which allowed me to grow both professionally and personally. I am also thankful to Asha Krishnaswamy, MSc, B.EE, my CDC field advisors, for her expertise and passion for this project, which fueled me to power on. She has taught me a great deal about believing in one's vision. I am most grateful to my advisors for their belief in me and providing me the perfect level of support that was conducive to my professional and personal growth.

I would like to acknowledge Stuart K. Shapira, MD, PhD, Wendy Book, MD, David Kleinbaum, PhD, Paul Weiss, MS and Adam S. Vaughan, MS for their technical assistance in congenital heart defects and statistical methodology.

I would like to thank my fellow colleagues, Adam Benson, Eeshwar Chandrasekar, Sonali Gupta, Neetu Hariharan, Molly Steel, Wen Yang and Joana Yu, who have provided valuable feedback and support through all the stages of this thesis project.

Finally, I would like to thank my family for their invaluable support and motivation, especially my younger siblings: 1) Bintou Diallo, whose achievements have been incredibly inspiring; 2) Diaraye, whose humor and love is always refreshing to go with her focused and unrelenting perseverance; and 3) Algassimou, whose brilliance continues to amaze me.

Table of Contents

Introduction:

Congenital heart defects (CHD) are structural abnormalities of the heart, aorta, or other large blood vessels that arise during fetal development, and cumulatively, represent the most prevalent birth defect in the United States (1). An estimated 2 million infants, adolescents, and adults live with CHD in the United States (2). Advances in medical and surgical care have led to increased life expectancy among these individuals. As a result, there are now slightly more adults than children and adolescents living with CHDs, and this figure will continue to increase over time (1).

Increasing age among those with CHDs presents new health challenges and increased risk of a multitude of late cardiovascular complications including heart failure, arrhythmia, and thrombosis (3), and non-cardiac complications such as renal dysfunction and restrictive lung disease (4). Consequently, adults with CHDs use substantially more health care resources than does the general population. For instance, the number of hospital admissions in the US among adults with CHDs doubled between 1998 and 2005 from 35,992 to 72,656, with most admissions originating from emergency departments and involving cardiac surgeries, compared to only a 13% increase in the general population (5).

Preventative care and specialty management of CHDs can increase quality of life and prevent complications and reduce mortality and healthcare costs among those affected (6). Guidelines recommend that adults with moderate or complex CHDs receive periodic (every 6-24 months) follow-up in adult CHD centers (7, 8). However, studies conducted in Canada and the Netherlands, where patients have access to care through universal healthcare systems, estimated that a high proportion of young adults (47-60%) between 18-22 years old, do not receive continuity of care (9, 10). One potential reason for this trend can be due to not requiring cardiology follow-up if the patients have certain mild CHDs. Another Canadian study reported that 53% of young adults with complex CHDs were not even identified and under cardiology care

(11). Proximity to specialized centers, being male, cardiology visits outside the university setting, and cost also contribute to adults' failure to receive continuity of care. However, these patients continue to receive primary care, suggesting that access may be a less important barrier than awareness of these patients' conditions among the medical community (9, 12). Failure to receive continuity of specialized care, especially among those with complex CHDs, increases the risk of delayed recognition of complications and poor outcomes (13, 14).

Given the high costs associated with hospitalizations and surgery and the high proportion of adults with moderate and complex CHDs not receiving specialized care, it is in the financial interests of insurance companies and other providers in the United States to identify adults with moderate and complex CHD and get them the care as early possible to prevent costly cardiac interventions, especially since patients can no longer be denied coverage or charged higher premiums due to a preexisting conditions like CHDs under the 2010 Patient Protection and Affordable Care Act (15).

The high proportion of young adults continuing to receive primary care and the increasing utilization of electronic health records (EHR) systems among health providers present a unique and efficient opportunity to facilitate identification of adults with moderate or complex CHDs – i.e. those most likely to benefit from diagnosis and referral to specialist care (16, 17).

Identifying adults with moderate or complex CHD from large cohorts would enable health providers to make specialized care more accessible, encourage individuals not receiving specialized care to seek cardiology follow-up, and identify primary care physicians caring for these individuals and educate them on how to best utilize specialized care resources to improve the health of their patients. However, there is no validated procedure for using data from EHR systems to identify adults with moderate or complex CHD. To address this deficit, we derived and internally validated a risk score based on common indicators found in patient records: demographic, electrocardiogram (e.g., PR interval, left ventricular hypertrophy, and heart rate), and laboratory parameters (e.g., B-type Natriuretic Peptides and blood hemoglobin levels).

Methods:

Data Sources and Study Design

We used patient administrative and clinical records from Emory Healthcare's Data Warehouse. Emory represents the largest multispecialty health care provider in the state of Georgia. We used a case-control study design to identify characteristics and factors that distinguish people with and without moderate-to-complex CHD. To develop and validate a risk score, we used a split-sample validation approach: study participants were randomly split into two groups at a ratio of 4 to 1 between the model development and validation (i.e. holdout) groups, respectively. The split sample approach, and the validation group especially, allows for the evaluation of the "best" model's reliability and performance in an independent sample derived from the same population when new data is unavailable (18).

The Institutional Review Board of Emory University (Atlanta, GA, USA) approved this study protocol and the Institutional Review Board of the Centers for Disease Control and Prevention was also notified.

Study Population

Study participants were adult outpatients, 20-60 years old, cared for in at least one of Emory Healthcare's facilities that had electrocardiogram (EKG) test results between January 2009 and December 2012. Cases were defined as patients with physician diagnosed moderate-to-complex CHDs based on International Classification of Diseases Ninth Revision and Clinical Modifications (ICD-9-CM codes 745.0-746.7) and receiving care at the Emory Adult Congenital Heart Clinic (EACHC). Control subjects were patients seen at any other Emory Healthcare facilities who did not have any ICD codes representing mild, moderate, or complex CHD diagnosis. Figure 1 shows a flow chart of patients who met inclusion/exclusion criteria for the study population. Of the 33,660 patients with valid EKG test results, 1,362 were seen at EACHC and 32,298 were cared for at other Emory Healthcare facilities. Among the EACHC patients, 766 were excluded for having ICD codes (n = 582), ICD codes corresponding to mild CHD (n = 167), and missing age values (n = 17). Among the non-EACHC patients, 485 were excluded due to having missing age data (n = 350) and ICD codes representing CHD diagnosis (n = 135; mild CHD = 84; moderate CHD = 41; complex CHD = 10); we randomly selected 2,384 of the remaining 31,813 non-EACHC patients to represent the controls. The final analysis data set contained 2,980 patients, which was composed of 596 cases and 2,384 controls.

Data Collection

We extracted data from Emory Healthcare's electronic health record (EHR) system regarding age, race/ethnicity, and EKG and laboratory test results during routine outpatient visits during January 2009 and December 2012.

Study variables

The primary outcome variable was clinically-diagnosed moderate-to-complex CHD as defined by the American College of Cardiology (8). Moderate CHD was defined as having any of the following: common truncus, stenosis of pulmonary valve, and/or Tetralogy of Fallot. Complex CHD was defined as having any of the following: transposition of great arteries, tricuspid artesia and stenosis (congenital type), hypoplastic left heart syndrome, and/or a common ventricle. We excluded adult patients with mild CHDs (i.e. isolated arterial septal defect and isolated ventricular septal defects) as non-cases because these defects have high likelihood of undergoing resolution in childhood, and they tend to not require lifelong cardiac care (see Figure 2). Additionally, mild defects are highly misclassified due to diagnostic and data entry errors associated with these conditions (19). Where patients had two or more diagnoses, the most complex diagnosis classification was assigned as the patient's primary diagnosis.

Initially, we examined factors (out of age, sex, race/ethnicity EKG parameters, B-type Natriuretic Peptide (BNP), hemoglobin, echocardiograms and chest X-ray) related to the outcome. Age, sex, and race/ethnicity were self-reported items in the EHR. EKG data were automatically coded by and stored in the MUSE® Cardiology Information System (GE Healthcare, Wauwatosa, WI, USA) to provide a sense of the heart's electrical conduction and size of the atria and ventricles. This data was directly linked with the administrative data in the EHR and created de-identified patient records before extracting data. However, we excluded race/ethnicity and hemoglobin, due to the large amount of missing data, and echocardiogram and chest X-rays because they were non-structured data. Age, sex, EKG, and BNP –the only remaining blood biomarker– were included in the analysis. Please see Table 2 for the list of predictors (exposure variables) and their classifications as examined in our models.

Statistical analysis:

We performed analyses using SAS 9.4 (Gary, NC, 2014) and VassarStats Clinical Calculator 1 (Poughkeepsie, NY, 2015). We randomly assigned four-fifths (n = 2,384) of the study participants into the model derivation cohort and reserved one-fifth (n = 596) for internal validation. To ensure even distribution of demographic and clinical characteristics between the derivation and validation cohorts, we estimated and compared frequencies (Chi-square or Fisher's Exact Tests for categorical variables, and student's t-test for continuous variables).

We used bivariate logistic regression to identify predictors related to moderate-to-complex CHD and retained exposures associated with the outcome that were statistically significant at $p < 0.05$. Collinearity diagnostics were performed and effect modifiers were not considered to simplify and facilitate the implementation of the algorithm and score (20).

In a full multivariable unconditional logistic regression model, we examined exposures that remained independently associated with the outcome using backward stepwise selection (variables remaining were significant at $p < 0.05$). In another model, we also evaluated whether excluding the laboratory predictor and statistically significant predictors with small coefficient estimates was associated with any major change (an increased or decreased predictive value).

To identify the "best" model —i.e., the one which best distinguishes patients with moderate to complex CHD from those without the outcome— we compared multivariate models using Receiver Operating Characteristic (ROC) *c*-statistics. We also evaluated how well the models predict the observed outcome in the data using the Hosmer-Lemeshow test (HL) for which a statistically significant value at $p < 0.05$ indicates a significant deviation between predicted and observed outcomes (20). In cases where it was difficult to discern the difference between models using the criteria above, we used the Bayes Information Criterion (BIC), a likelihood-based statistic that penalizes models with higher number of variables, to serve as the final criterion (21). Lower BIC values indicate better fit.

We conducted an internal validation to explore how the "best" model performs in the validation data set (n = 596). To evaluate performance, we used the Brier score (a global measure that calculates the sum of squared differences between the observed outcome and fitted probabilities and for which smaller values indicate better agreement between predicted and observed outcomes), the ROC *c*-statistics, and the HL test. To describe if and how the "best" model distinguishes patients with moderate to complex CHD from those without these conditions, we calculated diagnostic accuracy (sensitivity, specificity, positive predictive value [PPV] and negative predictive value [NPV]) of the model. The final model was accepted based on a combination of the measures described above and simplicity.

Risk score development

After determining the final predictive model, a risk score was developed to simplify the computation of patients' total risk based on methods outlined by Sullivan, Massaro & D'Agostino (2004). (22). To do this, continuous predictors were categorized (to simplify the implementation of the score in clinical settings) and reference values for each predictor were determined. Second, the difference between each category and the referent category was identified. Finally, the difference in coefficients for each category was multiplied by a constant before rounding the resulting number to the nearest integer. To use the score, the whole integer corresponding to the level of each exposure present is added up and examined against the cut-off for defining moderate to complex CHD.

Determination of Risk Score Cut-off

We used 1 / [1 + $e^{\wedge\text{-(final model)}}$] to calculate predictive probabilities of the final model and their corresponding sensitivities and specificities using the ROC curves and the VasserStats Clinical Calculator 1. The cut-off for the risk score to accurately identify cases of moderate to complex CHD was determined by identifying the threshold at which the optimal permutation of sensitivity and specificity was found.

Results:

Descriptive characteristics of the model derivation and internal validation cohorts and associated statistics are shown in Table 1a and Table 1b. Among the model derivation cohort, compared to control participants, adult CHD cases were, on average, a decade younger and had a considerably higher QRS duration (131.10 vs. 89.83 msec) on EKG. Cases were more likely than controls to be non-Hispanic white and have enlarged right and/or left atria, right and left ventricular hypertrophy, and right and left bundle branch blocks. A similar distribution of these characteristics was observed in the validation cohort except that controls were as likely to be white as cases.

Of the exposure variables examined in the bivariate analyses, age, sex, QRS duration, QRS axis, right and/or left atrial enlargement, non-sinus rhythm, right and left ventricular hypertrophy, right and left buddle branch block, and B-type natriuretic peptide (BNP), were all strongly associated with moderate to complex CHD (Table 2). The full multivariable logistic regression model, which included all variables listed above, had a ROC c-statistic of 0.96 [95% confidence interval (CI): 0.95, 0.97] (Table 3). Removal of additional variables, including QRS axis and BNP, the only cardiac biomarker evaluated, did not improve or compromise the discriminative ability of the model as the ROC c-statistics were similar across all three models produced (laboratory, non-laboratory, simplified) ranging from 0.960 to 0.964. Because of the similarities in ROC *c*-statistics, we also compared the Bayesian Information Criterion (BIC), for which lower values indicate better fit, across the models. In the model derivation data set, the BIC value for the non-laboratory (no-BNP) model (BIC = 931.81) was larger than that of the full model (BIC = 924.91) but lower than the simplified model (BIC = 937.11). Given the minimal differences in BIC values between non-laboratory and full models (deemed more complex because of the additional laboratory test), the non-laboratory (no-BNP) model was selected as the final algorithm.

Table 4 shows the summary statistics from internal validation regarding how the three models performed in the validation data set. The ROC *c*-statistics ranged from 0.97 to 0.98 across the three models. Although the full-model exhibited better global and calibration measures compared to the other models, the absolute difference was minimal and these better features were achieved at the expense of complexity. The performance of the simplified model and the non-laboratory were identical (Table 5). Both models had 94.59% sensitivity, 92.37% specificity, and 73.94% positive predictive value. As a result, the non-laboratory model was confirmed as the final algorithm (ACHD model) from the validation data. Table 6 and Figure 3 shows the estimate and ROC curve for the ACHD model in the validation cohort, respectively. The area under the curve, the probability that a random adult with moderate to complex CHD is correctly discriminated from adults without mild, moderate, or complex CHD, for the ACHD model was 0.97.

The ACHD Score, which was derived from the ACHD model, is presented in Table 7. Predictors and their categories are shown in the first two columns on the left. The third column contains point values corresponding to each category. Only one category can be chosen for each exposure and the total points for each indicator can be noted in the last column, "Points for each indicator." Adding the points of each predictor present gives each patient's final score. The differences in mean score between the derivation cohort, 8.88 (SD 6.06; min 0, max 38), and the validation cohort, 8.65 (SD 5.65; min 0, max 28), were not statistically significant (Student's t-test = 0.84, *P* = 0.40). In the validation cohort, the mean score was higher in cases than controls (17.71, SD 4.93 vs. 8.00, SD, 5.61, Student's t-test = 2.58, *P* = 0.01), in females than males (9.19, SD 5.62 vs. 8.00, SD, 5.61, Student's t-test = 2.58, *P* = 0.01), and in younger than older age groups (15.06, SD 5.40 vs. 5.00, SD 3.28).

From the validation cohort data set, the optimal threshold score, defined by a combination of slightly higher sensitivity than specificity, was 11, which was associated with a sensitivity of 93.69% and a specificity of 90.72% for the outcome (Table 8). Although lower thresholds were associated with higher sensitivity values, they did so at the expense of specificity.

Discussion:

We developed an empirically-derived risk score using routine clinical variables in electronic health records to differentiate between adults with and those unlikely to have moderate or complex congenital heart defects. The final algorithm (ACHD model) demonstrated good calibration and discriminatory power in a randomly-generated split-sample internal validation cohort. The final model correctly identified adult CHD patients with a 95% sensitivity and a 74% positive predictive value. The ACHD model was subsequently used to develop the ACHD Risk Score, which similarly demonstrated high sensitivity (94%) and positive predicted values (70%).

The ACHD Risk Score has several clinical and public health applications. Embedding an automated algorithm of this sort in EKG machines and EHR systems can provide automated screening during hospital or emergency room visits. It also has a high potential for identifying patients lost to cardiac follow-up from large patient cohorts whose data are stored on EHR systems. As more and more clinical settings utilize advanced EKG machines, this capability could be built in to flag potential high-risk individuals for further cardiology assessment. Identification may be an important event in encouraging these individuals to see cardiac specialists. Additionally, it can be used to identify primary care physicians and provide them with information about how to best utilize specialized care resources to improve their patients' health. Furthermore, from a public health standpoint, this tool can be employed to establish multicenter registries for surveillance across the country and overseas.

Using the model components may also help estimate national prevalence from cross-sectional surveys that include these measures and may have trouble defining CHDs based on self-report alone. EHR systems provide an added value over administrative data, which are regularly used to estimate disease incidence and prevalence, because they incorporate clinical data including laboratory results and facilitate the identification of patients (23-25). Administrative algorithms often rely on billing codes, specifically ICD codes, and include a criterion of two or more patient encounters with providers for the condition of interest, which may not always be available (19, 26, 27). On the other hand, a code derived from a risk score like the one we developed, which rely on both administrative and clinical data, can be applied to EHR systems to improve the accuracy of identifying patients (28). Although success of this algorithm depends on the quality of data stored in EHR systems, the completeness of the data utilized to develop this algorithm suggest that EKG data stored in EHR systems are viable option for identifying adults with moderate and complex CHDs (29-31). We chose a model where the variables are all likely to be available.

Several surgical risk scoring systems for the prediction of mortality, major adverse events, and prolonged lengths of stay among pediatrics with CHD demonstrated good predictive ability in the adult CHD population undergoing congenital heart surgery (32). To the best of our knowledge, our model is the first that predicts the likelihood of having moderate to complex congenital defects based on age, sex, and EKG markers obtained from an EHR system. Also, impressive performance of our tool suggests high potential for wider use.

The finding that BNP lacked added predictive value in the algorithm despite its statistical significance was consistent with the finding of other studies that indicated that increases in BNP values were associated with complex CHD, and should be used for specific clinical reasons, such as guiding therapy; however, researchers warn that BNP should be utilized with caution for screening purposes because it is a nonspecific biomarker of cardiac dysfunction (33, 34).

There are some limitations to our work. The risk tools we developed are not yet generalizable to populations outside Emory Healthcare settings as we could not validate the findings externally, particularly in populations with a higher proportion of racial minority groups. There were major differences in the distribution of racial groups within our group of cases than is usually found in other studies conducted in Metro Atlanta (35). The use of ICD codes to exclude controls is a potential source for disease misclassification resulting in an overestimation of the model, especially since providers who may not be familiar with CHD diagnoses are tasked with entering ICD codes in EHR systems (19).

In summary, the ACHD risk score we developed, which was composed of only age, sex and EKG markers, provided 94-95% sensitive and 91-92% specificity in the validation cohort. Although these tools require external validation in order to apply them in non-Emory Healthcare settings, they have potential to identify patients lost to cardiology follow-up whose data are stored in EHR systems, establish multicenter surveillance across the country and overseas, and flag potential high-risk individuals for further cardiology assessment when built into advanced EKG machines.

Work Cited

1. Centers for Disease Control Prevention. *Congenital heart defects: Data and Statistics*. 2014 [cited 2014 November 08, 2014]; Available from: http://www.cdc.gov/ncbddd/heartdefects/data.html.
2. Hoffman, J.I. and S. Kaplan, *The incidence of congenital heart disease.* J Am Coll Cardiol, 2002. **39**(12): p. 1890-900.
3. Verheugt, C.L., et al., *Mortality in adult congenital heart disease.* Eur Heart J, 2010. **31**(10): p. 1220-9.
4. Cohen, S.B., et al., *Extracardiac complications in adults with congenital heart disease.* Congenit Heart Dis, 2013. **8**(5): p. 370-80.
5. Opotowsky, A.R., O.K. Siddiqi, and G.D. Webb, *Trends in hospitalizations for adults with congenital heart disease in the U.S.* J Am Coll Cardiol, 2009. **54**(5): p. 460-7.
6. Dearani, J.A., et al., *Caring for adults with congenital cardiac disease: successes and challenges for 2007 and beyond.* Cardiol Young, 2007. **17 Suppl 2**: p. 87-96.
7. Warnes, C.A., et al., *ACC/AHA 2008 Guidelines for the Management of Adults with Congenital Heart Disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing committee to develop guidelines on the management of adults with congenital heart disease).* Circulation, 2008. **118**(23): p. e714-833.
8. Warnes, C.A., et al., *Task force 1: the changing profile of congenital heart disease in adult life.* J Am Coll Cardiol, 2001. **37**(5): p. 1170-5.
9. Mackie, A.S., et al., *Children and adults with congenital heart disease lost to follow-up: who and when?* Circulation, 2009. **120**(4): p. 302-9.
10. Winter, M.M., B.J. Mulder, and E.T. van der Velde, *Letter by Winter et al regarding article, "Children and adults with congenital heart disease lost to follow-up: who and when?".* Circulation, 2010. **121**(12): p. e252; author reply e253.
11. Reid, G.J., et al., *Prevalence and correlates of successful transfer from pediatric to adult health care among a cohort of young adults with complex congenital heart defects.* Pediatrics, 2004. **113**(3 Pt 1): p. e197-205.
12. Gilboa, S.M., et al., *Mortality resulting from congenital heart disease among children and adults in the United States, 1999 to 2006.* Circulation, 2010. **122**(22): p. 2254-63.
13. Gurvitz, M.Z., et al., *Changes in hospitalization patterns among patients with congenital heart disease during the transition from adolescence to adulthood.* J Am Coll Cardiol, 2007. **49**(8): p. 875-82.
14. Mackie, A.S., et al., *Health care resource utilization in adults with congenital heart disease.* Am J Cardiol, 2007. **99**(6): p. 839-43.
15. Vonder Muhll, I., G. Cumming, and M.A. Gatzoulis, *Risky business: insuring adults with congenital heart disease.* Eur Heart J, 2003. **24**(17): p. 1595-600.
16. Blumenthal, D. and M. Tavenner, *The "meaningful use" regulation for electronic health records.* N Engl J Med, 2010. **363**(6): p. 501-4.
17. Hsiao, C.J. and E. Hing, *Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2012.* NCHS Data Brief, 2012(111): p. 1-8.
18. Kleinbaum, D.G., et al., *Applied regression analysis and other multivariate methods*. 4th ed. 2008: Thomson Books/Cole.
19. Broberg, C., et al., *Accuracy of administrative data for detection and categorization of adult congenital heart disease patients from an electronic medical record.* Pediatr Cardiol, 2015. **36**(4): p. 719-25.

20. Kleinbaum, D.G. and M. Klein, *Logistic Regression: A Self-learning Text*. Third ed. 2010, New York: Springer Publishers.
21. Ridker, P.M., et al., *Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score.* Jama, 2007. **297**(6): p. 611-9.
22. Sullivan, L.M., J.M. Massaro, and R.B. D'Agostino, Sr., *Presentation of multivariate data for clinical use: The Framingham Study risk score functions.* Stat Med, 2004. **23**(10): p. 1631-60.
23. Bobo, W.V., et al., *An electronic health record driven algorithm to identify incident antidepressant medication users.* J Am Med Inform Assoc, 2014. **21**(5): p. 785-91.
24. Deshpande, A.D., M. Schootman, and A. Mayer, *Development of a claims-based algorithm to identify colorectal cancer recurrence.* Ann Epidemiol, 2015. **25**(4): p. 297-300.
25. Hayrinen, K., K. Saranto, and P. Nykanen, *Definition, structure, content, use and impacts of electronic health records: a review of the research literature.* Int J Med Inform, 2008. **77**(5): p. 291-304.
26. Benchimol, E.I., et al., *Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data.* Gut, 2009. **58**(11): p. 1490-7.
27. Benchimol, E.I., et al., *Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada.* J Clin Epidemiol, 2014. **67**(8): p. 887-96.
28. Tang, P.C., et al., *Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures.* J Am Med Inform Assoc, 2007. **14**(1): p. 10-5.
29. Byrd, J.B., et al., *Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: The VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program.* American Heart Journal, 2013. **165**(3): p. 434-440.
30. Staroselsky, M., et al., *Improving electronic health record (EHR) accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input.* Journal of General Internal Medicine, 2005. **20**: p. 100-100.
31. Widdifield, J., et al., *Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario, Canada.* Mult Scler, 2014.
32. Kogon, B. and M. Oster, Assessing surgical risk for adults with congenital heart disease: are pediatric scoring systems appropriate? J Thorac Cardiovasc Surg, 2014. **147**(2): p. 666-71.
33. Eindhoven, J.A., et al., *The usefulness of brain natriuretic peptide in complex congenital heart disease: a systematic review.* J Am Coll Cardiol, 2012. **60**(21): p. 2140-9.
34. Vuolteenaho, O., M. Ala-Kopsala, and H. Ruskoaho, *BNP as a biomarker in heart disease.* Adv Clin Chem, 2005. **40**: p. 1-36.
35. Bjornard, K., et al., *Patterns in the prevalence of congenital heart defects, metropolitan Atlanta, 1978 to 2005.* Birth Defects Res A Clin Mol Teratol, 2013. **97**(2): p. 87-94.

Tables

Table 1a Detailed demographic and clinical characteristics of study population in Model Derivation Cohort (n = 2,384)

Characteristics	**Cases (n= 485)**	**Controls (n= 1,899)**	**P-value***
	Means (±SD) or N (%)		
Age (years)	34.00 (±10.52)	46.90 (±9.91)	<0.0001
20-34	290 (59.79)	239 (12.59)	<0.0001
35-49	139 (28.66)	761 (40.07)	<0.0001
50-60	56 (11.55)	899 (47.34)	<0.0001
Sex			
Females	260 (53.61)	981 (51.66)	0.443
Race			<0.0001
Non-hispanic white	322 (78.82)	845 (54.31)	
Non-hispanic black	77 (18.87)	643 (42.98)	
Asian	4 (0.98)	30 (1.93)	
Hispanic	4 (0.98)	31 (1.99)	
Native Indian/Pacific Islander	1 (0.25)	7 (0.39)	
PR Interval (msec)	160.10 (±51.48)	153.90 (±30.49)	0.012
QRS Duration (msec)	131.10 (±32.64)	89.83 (±14.99)	<0.0001
QRS axis (degrees)	59.24(±72.75)	38.49 (±37.24)	<0.0001
Heart Rate (bpm)	72.48 (±11.58)	70.62 (±13.08)	0.002
Atrial Enlargement, Right, Left and Biatrial	116 (23.92)	134 (7.06)	<0.0001
Rhythm not sinus	131 (72.99)	61 (3.21)	<0.0001
RVH	100 (20.62)	11 (0.58)	<0.0001
LVH	68 (20.34)	86 (4.53)	<0.0001
RBBB	310 (63.92)	85 (4.85)	<0.0001
LBBB	73 (15.05)	38 (2.00)	<0.0001
BNP			
BNP (pg/ml)	247.20 (±545.9)	435.7 (±722.60)	0.003
BNP (>100 pg/ml)	110 (22.68)	64 (3.73)	<0.0001

Abbreviations: BPM, beats per minute; MSEC, milliseconds; RVH, right ventricular hypertrophy; LVH, left ventricular hypertrophy; RBBB, right bundle branch block; LBBB, left bundle branch block; BNP, B-type Natriuretic Peptide; pg/ml, picogram per milli-liter

*All tests were chi-square for categorical variables and student t-tests for continuous variable at 0.05 significance level.

Table 1b Detailed demographic and clinical characteristics of study population in Model Validation Cohort (n = 596)

Characteristics	Cases (n = 111)	Controls (n = 485)	P-value*
	Means (±SD) or N (%)		
Age (years)	34.06 (±10.82)	47.00 (±9.61)	<0.0001
20-34	63 (56.76)	59 (12.16)	<0.0001
35-49	31 (27.93)	201 (41.44)	0.008
50-60	17(15.32)	225 (46.39)	<0.0001
Sex			
Females	70 (57.85)	264 (52.91)	0.328
Race			0.113
Non-hispanic white	79 (80.61)	302 (68.79)	
Non-hispanic black	16 (16.33)	123 (28.02)	
Asian	1 (1.02)	8 (1.82)	
Hispanic	2 (2.04)	4 (0.91)	
Native Indian/Pacific Islander	0(0.00)	2 (0.46)	
PR Interval (msec)	156.6 (±51.42)	155.50 (±28.77)	0.833
QRS Duration (msec)	131.1 (±31.24)	88.88 (±13.72)	<0.0001
QRS axis (degrees)	66.80(±75.61)	42.78 (±39.07)	0.0015
Heart Rate (bpm)	72.76 (±11.01)	70.95 (±13.73)	0.1388
Atrial Enlargement, Right, Left and Biatrial	32 (26.45)	183(36.67)	0.034
Rhythm not sinus	34 (28.0)	36 (7.21)	<0.0001
RVH	20 (16.53)	5 (1.00)	<0.0001
LVH	15 (12.4)	36 (7.21)	0.063
RBBB	79 (65.29)	106 (21.24)	<0.0001
LBBB	17 (14.05)	36 (7.21)	0.0158
BNP			
BNP (pg/ml)	374.1 (±765.00)	449.90 (±583.30)	0.630
BNP > 100 pg/ml	27 (24.32)	17 (3.51)	<0.0001

Abbreviations: BPM, beats per minute; MSEC, milliseconds; RVH, right ventricular hypertrophy; LVH, left ventricular hypertrophy; RBBB, right bundle branch block; LBBB, left bundle branch block; BNP, B-type Natriuretic Peptide; pg/ml, picogram per milli-liter

*All tests were chi-square for categorical variables and student t-tests for continuous variable at 0.05 significance level.

Table 2 Variables examined in multivariate logistic regression model using backwards stepwise approach and their resulting estimates, model derivation cohort (n = 2, 384)

Variables*	Definition	Estimate	Standard Error	Chi-Sq	*P*-value
Age	Included in model as continuous variable, split into tertiles at score development stage	-0.12	0.01	181.36	<0.0001
Sex	Male; Female	-0.87	0.19	20.13	<0.0001
QRS Duration	Included in model as continuous variable, split into the following quartiles at score development stage: Against (<80 msec); Neutral (80-120 msec); Likely (> 150 msec); Supports (> 120 - 150 msec)	0.05	0.00	118.01	<0.0001
QRS Axis	Included in model as continuous variable	0.01	0.00	14.60	0.0001
Atrial enlargement right, left, biatrial	Coded as present or absent by ECG Machine	1.47	0.26	31.55	<0.0001
Rhythm not sinus	Coded as present or absent by ECG Machine	1.84	0.30	38.65	<0.0001
Right ventricular hypertrophy	Coded as present or absent by ECG Machine	1.49	0.45	11.13	0.001
Left ventricular hypertrophy	Coded as present or absent by ECG Machine	0.75	0.33	5.26	0.022
Right buddle branch block	Coded as present or absent by ECG Machine	1.95	0.25	61.36	<0.0001
Left buddle branch block	Coded as present or absent by ECG Machine	0.84	0.40	4.26	0.039
B-type Natriuretic Peptide (>100 pg/mL)	Present if value greater than 100 pg/mL otherwise it was coded as absent	1.16	0.30	14.71	0.001

*Hear rate and PR interval were not statistically significant in bivariate logistic regression analysis and were, thus, excluded from multivariate logistic regression stage using backwards stepwise approach.

Table 3 Multivariable models using logistic regression backwards stepwise approach in the model derivation cohort (n = 2,384)

Models*	ROC c-statistic (95% CI)	BIC†	Brier Score†
Non-laboratory model-(ACHD model)	0.962 (0.952, 0.972)	931.81	0.049
Simplified model	0.960 (0.950, 0.971)	937.11	0.051
Full-model	0.964 (0.954, 0.973)	924.91	0.047

Abbreviation: ROC, Receiver Operation Curve; BIC, Bayesian information criterion

*Full-model contained variables found to be statistically significant variables during the multivariate logistic regression stage. The non-laboratory model excludes B-type Natriuretic Peptide (BNP) from the Full-model. The Simplified model excludes left bundle branch block, QRS axis & BNP from the Full-model.

†Lower BIC and Brier score values indicate better fit. Higher values of C-statistic indicate better discrimination.

Table 4 Summary Statistics Comparing the ACHD (non-laboratory), Simplified, and Full models to predict adult moderate or complex congenital heart defect, based on data from validation cohort (n = 596)

	ACHD Model {¥}	Simplified Model {¥}	Full Model {¥}
Global Measure			
Brier Score*	0.038	0.038	0.036
Discrimination			
C-statistic*	0.97 (0.95, 0.99)	0.97 (0.95, 0.99)	0.98 (0.96, 0.99)
Calibration			
Hosmer-Lemershow P value†	0.103	0.102	0.551

{¥} Full-model contained variables found to be statistically significant variables during the multivariate logistic regression stage. The ACHD Model excludes B-type Natriuretic Peptide (BNP) from the Full-model. The Simplified model excludes left bundle branch block, QRS axis & BNP from the Full-model.

*Lower values of Brier score values indicate better fit. Higher values of C-statistic indicate better discrimination.

† A significant value of Hosmer-Lemeshow statistic indicates a significant deviation between predicted and observed outcomes.

Table 5 Performance Characteristics of the ACHD (non-laboratory), Simplified, and Full models in the identification of adult moderate or complex congenital heart defect, validation cohort (n = 596)

Model	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
ACHD Model	94.59	92.37	73.94	98.68
Simplified Model	94.59	92.37	73.94	98.68
Full Model	93.69	92.78	74.82	98.67

Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table 6 Moderate or complex ACHD (non-laboratory) model, Model Derivation Cohort (n = 2,384)

Variables	Estimate	Standard Error	Chi-Sq	*P*-value
Age	-0.12	0.01	179.72	<.0001
Sex	-0.91	0.19	22.12	<.0001
QRS Duration	0.05	0.00	140.43	<.0001
QRS Axis	0.01	0.00	14.42	0.0001
Atrial Enlargement Right, Left, Biatrial	1.58	0.26	38.44	<.0001
Rhythm not sinus	2.02	0.29	47.90	<.0001
Right Ventricular Hypertrophy	1.54	0.44	12.47	0.0004
Left Ventricular Hypertrophy	0.83	0.32	6.66	0.010
Right Buddle Branch Block	1.84	0.25	56.41	<.0001
Left Buddle Branch Block	0.93	0.40	5.48	0.019

Best-fitting multivariate logistic regression model.

Table 7 Moderate or complex adult congenital heart defect (ACHD) risk score

Variables		Categories	Points for each category	Points for each indicator
Demographics				
Age		20-29	7	
		30-39	5	
		40-49	3	
		50-60	0	Points:
Sex		Male	0	
		Female	2	Points:
Electrocardiogram				
QRS Duration				
	Against	<80	0	
	Neutral	80-119	3	
	Support	120-149	6	
	Likely	≥150	10	Points:
Atrial enlargement right, left, biatrial		Absent	0	
		Present	3	Points:
Rhythm not sinus		Absent	0	
		Present	4	Points:
Right ventricular hypertrophy		Absent	0	
		Present	3	
Left ventricular hypertrophy		Absent	0	
		Present	2	Points:
Right buddle branch block		Absent	0	
		Present	4	Points:
Left buddle branch block		Absent	0	
		Present	2	Points:
				Total= *

*Threshold scores and their assocated sensitivity and positive predicted values: total points = **≥10**, Sen (96 %), PPV (52%); **≥11**, Sen (94 %), PPV (70%); **≥12**, Sen (88 %), PPV (75%)

Table 8 Performance characteristics of ACHD score around threshold score (11)*

Threshold score	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
10	96.40	80.00	52.45	98.98
11	93.69	90.72	69.80	98.43
12	88.29	93.61	75.97	97.22

Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value

*ACHD score is based on the non-laboratory model.

Figures and Appendix

Figure 1 Flow chart of patients who met inclusion/exclusion criteria for the study population

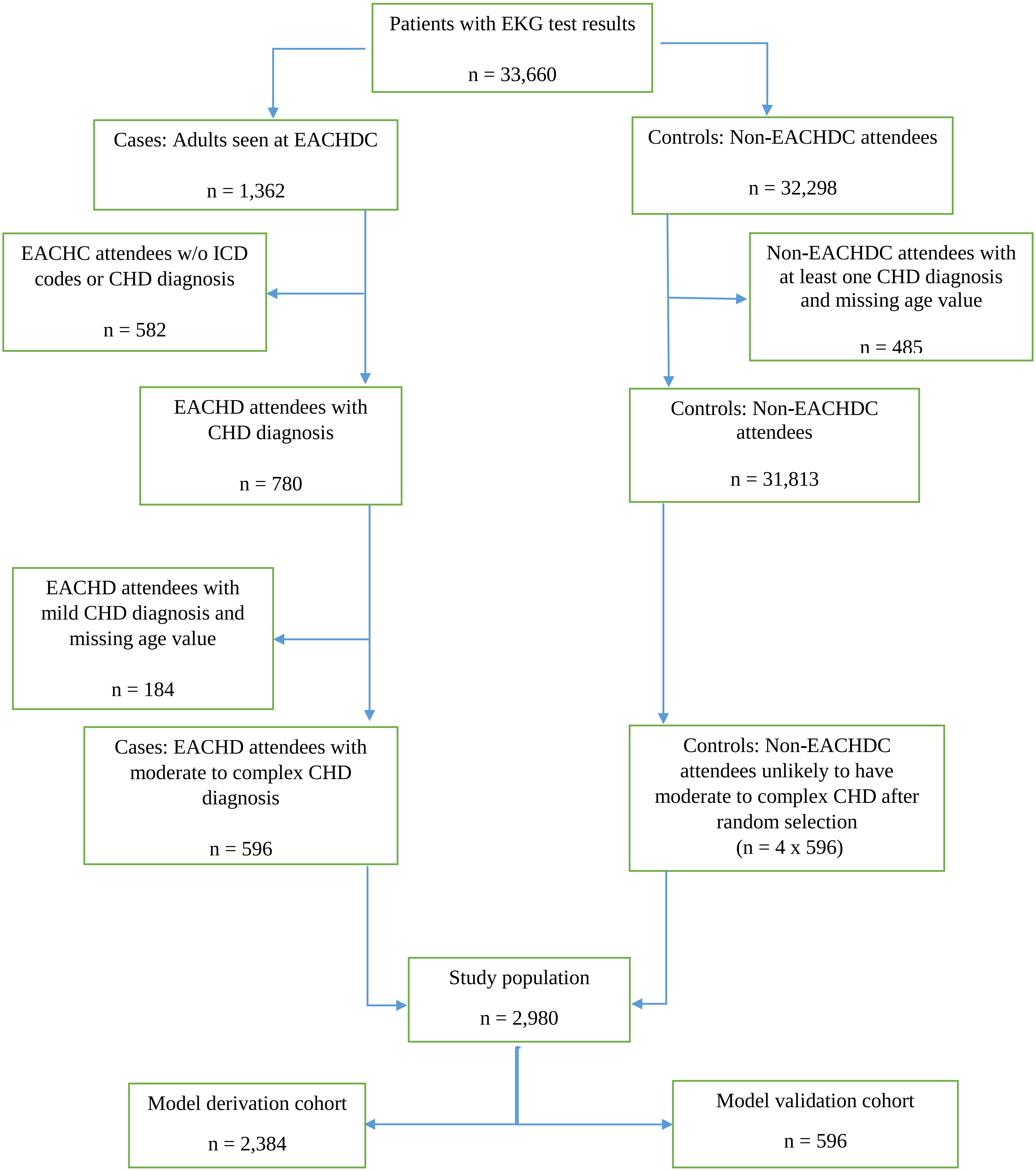
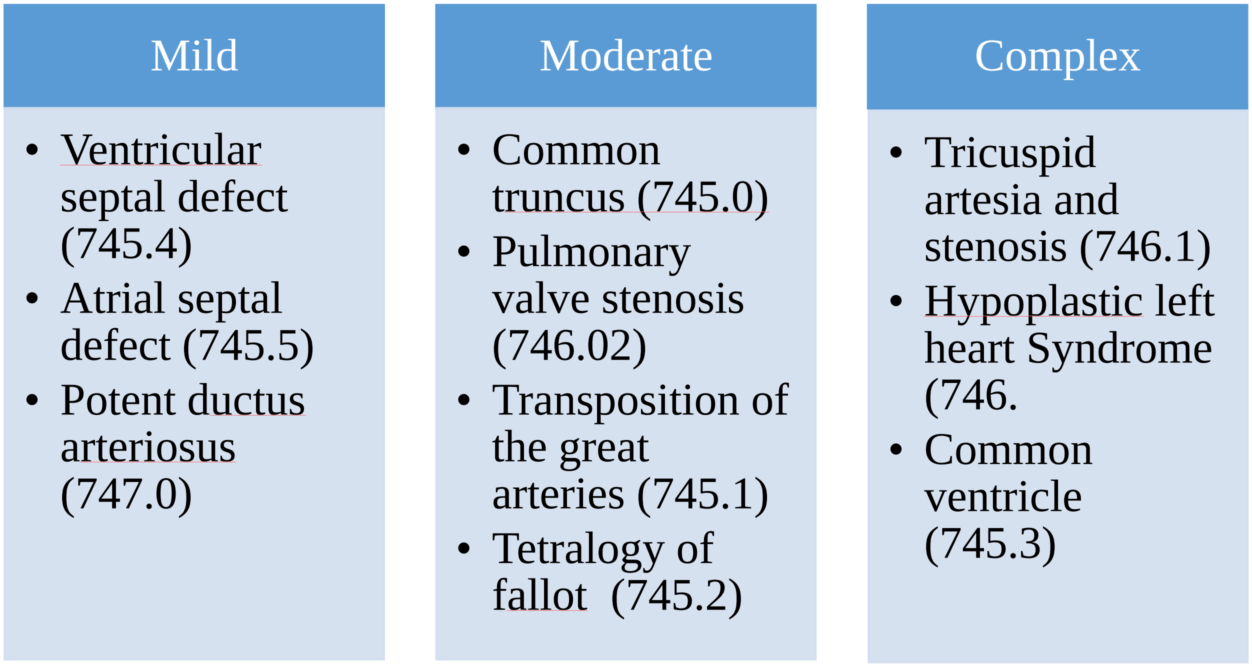

Patients with EKG test results
n = 33,660
Cases: Adults seen at EACHDC
n = 1,362
Controls: Non-EACHDC attendees
n = 32,298
EACHC attendees w/o ICD codes or CHD diagnosis
n = 582
Non-EACHDC attendees with at least one CHD diagnosis and missing age value
n = 485
EACHD attendees with CHD diagnosis
n = 780
Controls: Non-EACHDC attendees
n = 31,813
EACHD attendees with mild CHD diagnosis and missing age value
n = 184
Cases: EACHD attendees with moderate to complex CHD diagnosis
n = 596
Controls: Non-EACHDC attendees unlikely to have moderate to complex CHD after random selection
(n = 4 x 596)
Study population
n = 2,980
Model derivation cohort
n = 2,384
Model validation cohort
n = 596

Figure 2 Types of congenital heart defects and corresponding ICD-9 codes.



Mild

- Ventricular septal defect (745.4)
- Atrial septal defect (745.5)
- Potent ductus arteriosus (747.0)

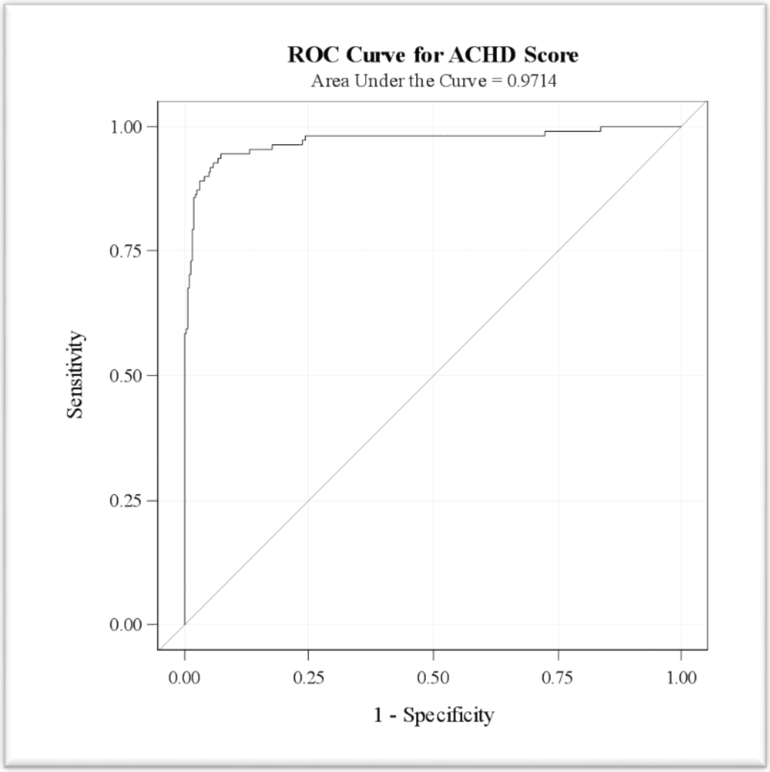
Moderate

- Common truncus (745.0)
- Pulmonary valve stenosis (746.02)
- Transposition of the great arteries (745.1)
- Tetralogy of fallot (745.2)

Complex

- Tricuspid artesia and stenosis (746.1)
- Hypoplastic left heart Syndrome (746.
- Common ventricle (745.3)

Figure 3 ROC curve of ACHD Model



Appendix 1 Development of moderate or complex adult congenital heart defect (ACHD) risk score*

Variables	Categories	Reference value (Wij)	βi	βi (Wij – WiREF)	Pointsij = βi (Wij – WiREF) x 2
Demographics					
Age			-0.1215		
	20-29	24.5		3.70575	7
	30-39	34.5		2.49075	5
	40-49	44.5		1.27575	3
	50-60	55=W1REF		0	0
Sex			0.9069		
	Male	0=W2REF		0	
	Female	1		0.9069	2
Electrocardiogram					
QRS Duration			0.0531		
Against	<80	74		0	0
Neutral	80-119	100		1.3806	3
Support	120-149	134.5		3.21255	6
Likely	≥150	168.4		5.01264	10
Atrial enlargement right, left, biatrial			1.5836		
	Absent	0=W4REF			
	Present	1		1.5836	3
Rhythm not sinus			2.0158		
	Absent	0=W5REF			
	Present	1		2.0158	4
Right ventricular hypertrophy			1.5438		
	Absent	0=W6REF			
	Present	1		1.5438	3
Left ventricular hypertrophy			0.8257		
	Absent	0=W7REF			
	Present	1		0.8257	2
Right buddle branch block			1.8429		
	Absent	0=W8REF			
	Present	1		1.8429	4
Left buddle branch block			0.9336		
	Absent	0=W9REF			
	Present	1		0.9336	2
				Maximum Possible	37

*Continuous predictors were categorized (to simplify the implementation of the score in clinical settings) and reference values for each variable were determined. Second, the difference between each category and the referent category was identified. Finally, the difference in coefficients for each category was multiplied by a constant (2) before rounding the resulting number to the nearest integer.