AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

Detecting moderate or complex congenital heart defects in adults from an electronic health records system

Alpha Oumar Diallo,[1] Asha Krishnaswamy,[2] Stuart K Shapira,[2] Matthew E Oster,[1,2,3,4] Mary G George,[5] Jenna C Adams,[4] Elizabeth R Walker,[4] Paul Weiss,[6] Mohammed K Ali,[1,7,8] and Wendy Book[4]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA, [2]National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA, USA, [3]Sibley Heart Center Cardiology, Children's Healthcare of Atlanta, Atlanta, GA, USA, [4]Emory University School of Medicine, Atlanta, GA, USA, [5]National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA, [6]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA, [7]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA and [8]Department of Family and Preventive Medicine, School of Medicine, Emory University, Atlanta, GA, USA

Correspondence to Alpha Oumar Diallo, MPH, Centers for Disease Control and Prevention, 1600 Clifton Road NE, MS C-25, Atlanta, GA 30333-4027, USA; adiallo@cdc.gov; 4dialloo@gmail.com

ABSTRACT

Background: The prevalence of moderate or complex (moderate-complex) congenital heart defects (CHDs) among adults is increasing due to improved survival, but many patients experience lapses in specialty care or their CHDs are undocumented in the medical system. There is, to date, no efficient approach to identify this population.

Objective: To develop and assess the performance of a risk score to identify adults aged 20-60 years with undocumented specific moderate-complex CHDs from electronic health records (EHR).

Methods: We used a case-control study (596 adults with specific moderate-complex CHDs and 2384 controls). We extracted age, race/ethnicity, electrocardiogram (EKG), and blood tests from routine outpatient visits (1/2009 through 12/2012). We used multivariable logistic regression models and a split-sample (4: 1 ratio) approach to develop and internally validate the risk score, respectively. We generated receiver operating characteristic (ROC) c-statistics and Brier scores to assess the ability of models to predict the presence of specific moderate-complex CHDs.

Results: Out of six models, the non-blood biomarker model that included age, sex, and EKG parameters offered a high ROC c-statistic of 0.96 [95% confidence interval: 0.95, 0.97] and low Brier score (0.05) relative to the other models. The adult moderate-complex congenital heart defect risk score demonstrated good accuracy with 96.4% sensitivity and 80.0% specificity at a threshold score of 10.

Conclusions: A simple risk score based on age, sex, and EKG parameters offers early proof of concept and may help accurately identify adults with specific moderate-complex CHDs from routine EHR systems who may benefit from specialty care.

Key words: risk score, electronic health record, congenital heart defects, adult, electrocardiogram, case-control study

INTRODUCTION

Congenital heart defects (CHDs), the most prevalent of birth defects, occur in approximately 1 newborn per 100 live births, comprising about 40 000 newborns per year in the United States.[1] As a quarter of these cases are critical (requiring surgery or life-saving procedures in the first year of life) and have high mortality rates, CHDs have long been viewed as primarily a pediatric disease. However, significant advances in surgical and medical care in the past 5 decades have greatly extended the lifespan of patients with CHDs.[2] As a result, the vast majority of patients with both critical (69%) and non-critical (95%) CHDs are now living to adulthood.[3] This new and aging adult population, already accounting for more than half of the estimated 2-3 million CHD survivors,[4] is projected to further increase in the future,[5] giving rise to unforeseen clinical and public health challenges. Thus, more accurate and consistent identification of adults with undocumented CHDs is imperative and is the overarching goal of our study.

Adults with CHDs face an array of cardiovascular and non-cardiovascular complications, including heart failure and renal dysfunction, resulting in a higher rate of hospitalizations compared to the general population, with most cases originating from emergency departments and requiring cardiac surgeries.[6–8] Therefore, clinical guidelines from the American College of Cardiology/American Heart Association Task Force on Practice Guidelines recommend specialized care for adults with moderate or complex (moderate-complex) CHDs, who make up approximately half of the adult population with CHDs.[9] A population-based study demonstrated the importance of these guidelines, with a resulting reduction in mortality.[10] However, cross-sectional studies conducted in Canada and the Netherlands estimated that a high proportion of young adults aged 18-22 years (47-60%) do not receive the recommended follow-up care.[11,12] Potential contributors are the "fixed for life" syndrome [a 1960s misnomer coined by the medical community that gave patients an assurance of being surgically fixed for life[13]] distance from specialized care centers, male gender, cost of care, cardiology visits outside university settings, and lack of awareness or education of the issues among patients, families, and care providers.[11,12,14]

While guidelines to address lifelong care for adults with specific moderate-complex CHDs are needed, the first challenge is to locate such individuals. The "fixed for life" syndrome and other factors make this a particular challenge as many adults with CHDs experience lapses in specialty care or their CHDs are undocumented in the medical system. However, a high proportion of young adults continue to receive primary care, and there is increasing utilization of electronic health records (EHR) systems (electronic medical records) that contain electrocardiogram (EKG) and other commonly collected data. These factors present a unique and efficient opportunity for potential identification of adults with moderate-complex CHDs– ie those most likely to benefit from referral to specialty care.[15,16]

The objective of this study was to develop a risk score algorithm based on routine EHR data for identification of adults with specific moderate-complex CHDs. We validated the risk score internally with hopes that it can be used by physicians to identify adults with undocumented specific moderate-complex CHDs who may need specialty care.

METHODS

Data sources and study design
We extracted EHR data from Emory Healthcare, the largest multi-specialty healthcare provider in the state of Georgia. We employed a model validation study design to identify characteristics and factors that distinguish people with and without specific moderate-complex CHDs.

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 (ie holdout) groups, respectively. The split-sample approach, especially the validation group, lends itself to the reliability and performance evaluation of the "best" model in an independent sample derived from the same population where new data are unavailable.[17]

The Institutional Review Board of Emory University (Atlanta, GA, USA) approved this study protocol.

Study population
Study participants included adult outpatients, aged 20-60 years, cared for in at least one of Emory Healthcare's facilities, and who had at least one EKG between January 2009 and December 2012. Cases were defined as patients receiving care at the Emory Adult Congenital Heart Center (ACHC) who had select moderate-complex CHDs based on International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes shown in Supplementary Table S1. We selected these commonly occurring moderate-complex CHDs to demonstrate a proof of concept as the test case. A moderate CHD included at least one of the following: common truncus, stenosis of pulmonary valve, or tetralogy of Fallot. A complex CHD included at least one of the following: transposition of great arteries, tricuspid atresia and stenosis (congenital type), hypoplastic left heart syndrome, or common ventricle. Since the use of ICD-9-CM codes to define the cases limited the ability to distinguish CHD severity of each case, every case with a particular ICD-9-CM code was included in just one severity category (eg all cases with stenosis of the pulmonary valve were included in the moderate CHD category although some can be severe, and mild cases are likely to have resolved in childhood). We excluded adult cases of mild CHDs (Supplementary Table S1)—specifically, isolated atrial septal defect, isolated ventricular septal defect, and patent ductus arteriosus. Patients with these CHDs have high likelihood of undergoing resolution in childhood, and per current guidelines, they tend not to require lifelong cardiac care. Additionally, mild CHDs are highly misclassified due to diagnostic and data entry errors associated with such conditions.[18]

Cases with moderate-complex ICD-9-CM codes comprised 44% of patients with CHD diagnoses cared for in the Emory ACHC. Since Emory ACHC physicians verify and enter their own diagnosis and billing codes, only patients with specific moderate-complex CHDs followed in the Emory ACHC were included as cases in order to increase diagnostic accuracy. Accuracy of ICD-9-CM CHD codes used outside the Emory ACHC could not be verified. Therefore, Emory Healthcare patients with any ICD-9-CM CHD code, who had not attended the Emory ACHC, were excluded from the derivation and validation cohorts.

Control subjects were patients seen at any other Emory Healthcare facility, other than the ACHC, who did not have an ICD-9-CM code for any mild, moderate, or complex CHD diagnosis. Figure 1 shows a flow chart of cases and controls that met inclusion/exclusion criteria for the study population. Of the 33 660 patients with EKG test results, 1362 received care at the Emory ACHC and 32 298 were cared for at other Emory Healthcare facilities. Emory ACHC patients were excluded for missing ICD-9-CM code
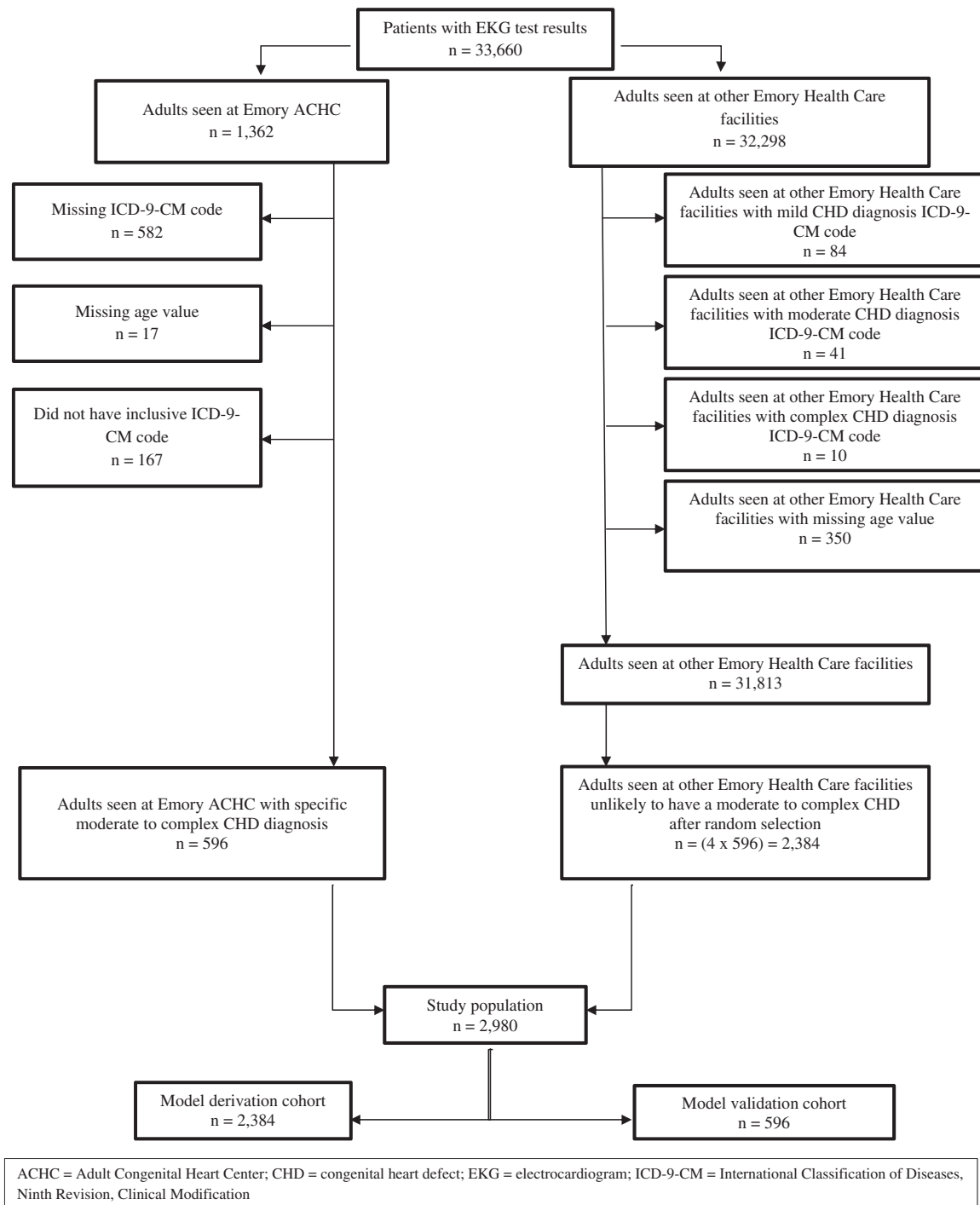
Figure 1. Flow chart of patients who met inclusion/exclusion criteria for the study population.

(n = 582) or age values (n = 17) and for not having an inclusive ICD-9-CM code (n = 167). Among the Emory Healthcare patients not attending the Emory ACHC, 485 were excluded, due to either having an ICD-9-CM code for a CHD (mild CHD = 84, moderate CHD = 41, and complex CHD = 10) or missing age values (n = 350). We randomly selected controls from the remaining 31 813 non-ACHC Emory Healthcare patients in a 1: 4 case: control ratio. The final analysis dataset contained 2980 patients, which was composed of 596 cases and 2384 controls. To validate the selection process for controls and to ensure that there was not notable mis-

classification bias, we performed detailed chart review of those in the derivation cohort who were identified by the final model as potentially false positive.

Data collection and study variables

Data extracted from Emory Healthcare's EHR system included age and race/ethnicity, as well as EKG data and blood biomarker test results for cardiac insufficiency (B-type natriuretic peptide (BNP) concentration) and anemia (hemoglobin concentration) performed

during routine outpatient visits between January 2009 and December 2012.

Statistical analysis and model development

We performed analyses using SAS 9.4 (Cary, NC, 2014) and Vassar-Stats Clinical Calculator 1 (Poughkeepsie, NY, 2015). We randomly assigned four-fifths (n = 2384) of the study participants into the model derivation cohort and reserved one-fifth (n = 596) for internal validation. To ensure an 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 first used bivariate logistic regression to examine relationships between exposures and the presence of any of the specific moderate-complex CHDs and identified those that showed a statistically significant association at $p < .05$. Collinearity diagnostics were performed and effect modifiers were not considered to simplify and facilitate the implementation of the algorithm and score.[19]

In a full multivariable unconditional logistic regression model, referred to as the full model, we examined exposures that remained independently associated with the outcome using backward stepwise selection. In another model, referred to as the non-blood biomarker model, we evaluated whether excluding the blood biomarker (ie BNP) and the statistically significant predictors with small coefficient estimates was associated with any major change (an increased or decreased predictive value). Similarly, we excluded various combinations of EKG parameters and BNP, which we referred to as simplified A, B and C models.

To identify the "best" model—the one that best distinguished cases with specific moderate-complex CHDs from controls—we calculated two measures of model performance: receiver operating characteristic (ROC) c-statistics and the Brier score. The c-statistic, which varies between 0.5-1.0, shows how well the models discriminated between those with and without the outcome; higher values indicate better discrimination. The Brier score, used to assess both calibration and goodness of fit, estimates the expected squared difference between the outcome and predicted probabilities; the Brier score ranges from 0 to 1 with lower scores indicating higher accuracy of predicted probability. In cases where it was difficult to discern the difference between models using the above criteria, we used the Bayesian Information Criterion (BIC), a likelihood-based statistic that penalizes models with higher numbers of variables; lower BIC values indicate better fit.

We then explored how the "best" model performed in the validation dataset (n = 596). To evaluate performance, we calculated sensitivity and specificity for both clinical and surveillance settings. The clinical setting refers to model parameters that would best identify potential high-risk individuals for further cardiology assessment and, if identified with one of the specific moderate-complex CHDs, provide the opportunity for those lost to follow-up to resume adult CHD specialty care; this setting prioritizes sensitivity over specificity. The surveillance setting refers to model parameters that would best minimize false positive cases for producing a sample of individuals with the specific moderate-complex CHDs, who are representative of the population in order to be able to evaluate basic epidemiology and healthcare utilization, as well as national prevalence and distributions from cross-sectional surveys that include EKG measures; this setting prioritizes specificity over sensitivity. We calculated true positives and true negatives for the clinical setting.

For the surveillance setting, we calculated positive predictive value (PPV) and negative predictive value (NPV) using an estimated population prevalence of 16.9 per 10 000 live births for the specific moderate-complex CHDs included in this study.[1] Decisions for the model were based on likelihood cutoff values from the ROC C-table for the presence of specific moderate-complex CHDs of 0.15 and 0.35, respectively, for the clinical and surveillance settings. The lower cutoff value was based on the need to have a model that is more sensitive than specific for the aims of the clinical setting; the larger cutoff is derived from the need for the model to be more specific than sensitive for surveillance purposes.

Development and determination of risk score threshold

After determining the final predictive model, a risk score was developed to simplify the computation of patients' total risk based on the Framingham risk score approach.[20] A score for each record in the validation dataset and the predictive probabilities of the scores were calculated using the risk score described above and $1/(1 + e^{(final\ model)})$. Predicted outcomes (having or not having one of the specific moderate-complex CHDs) were assigned to records based on whether the calculated risk scores were greater or equal to the tested threshold scores, which were 10-12. These predicted outcomes were then compared with the true/diagnosed outcome. The sensitivity and specificity of the tested threshold scores were calculated and compared using the VassarStats Clinical Calculator 1. The threshold for the risk score to accurately identify patients with potentially specific moderate-complex CHDs was determined by identifying the threshold of optimal permutation of sensitivity and specificity for a clinical setting, with the purpose of higher sensitivity to identify potential high-risk individuals for further cardiology assessment.

RESULTS

Descriptive characteristics of the model derivation cohorts are shown in Table 1. In the model derivation cohort, compared to controls, cases were more likely to be non-Hispanic white and were, on average, a decade younger: 59.8% of cases vs. 12.6% of controls were 20-34 years old; 40.2% of cases vs 87.4% of controls were 35-60 years old. Cases had a considerably higher QRS duration (131.1 vs. 89.8 msec) on EKG. Cases were more likely than controls to have enlarged right and/or left atria, right and left ventricular hypertrophy, right bundle branch block, left bundle branch block, and BNP greater than 100 pg/ml. Similar distributions of these characteristics and proportions of completeness (missing) for study variables were observed in the validation cohort, except that controls were as likely as cases to be non-Hispanic white (See Supplementary Tables S2 and S3, respectively).

After bivariate analysis, the full multivariable logistic regression model, which contained age, sex, QRS duration, QRS axis, right and/or left atrial enlargement, non-sinus rhythm, right ventricular hypertrophy, left ventricular hypertrophy, right bundle branch block, left bundle branch block, and BNP concentration, had a ROC c-statistic of 0.96 (95% confidence interval (CI): 0.95, 0.97), and a Brier score of 0.05 (Table 2). Removal of additional variables, including QRS axis and BNP concentration, did not improve or compromise the discriminative ability of the full model, as the ROC c-statistics for the non-blood biomarker and simplified A, B and C models (Supplementary Table S4) were similar to the full model at 0.96. The absolute differences in Brier scores among all models was ≤ 0.004. Because of the similarities in ROC c-statistics and Brier

Table 1. Detailed demographic and clinical characteristics of study population in model derivation cohort (n = 2384)

Characteristics	Cases (n = 485)	Controls (n = 1899)	P-value*
	Means (±SD) or N (%)		
Age (years)	34.00 (±10.52)	46.90 (±9.91)	<.0001
20-34	290 (59.79)	239 (12.59)	<.0001
35-49	139 (28.66)	761 (40.07)	<.0001
50-60	56 (11.55)	899 (47.34)	<.0001
Sex			
Females	260 (53.61)	981 (51.66)	.443
Race			<.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 American/Pacific Islander	1 (0.25)	7 (0.39)	
PR interval (msec)	160.10 (±51.48)	153.90 (±30.49)	.012
QRS duration (msec)	131.10 (±32.64)	89.83 (±14.99)	<.0001
QRS axis (degrees)	59.24 (±72.75)	38.49 (±37.24)	<.0001
Heart rate (bpm)	72.48 (±11.58)	70.62 (±13.08)	.0022
Atrial enlargement, right, left or biatrial	116 (23.92)	134 (7.06)	<.0001
Rhythm not sinus	131 (72.99)	61 (3.21)	<.0001
RVH	100 (20.62)	11 (0.58)	<.0001
LVH	68 (20.34)	86 (4.53)	<.0001
RBBB	310 (63.92)	85 (4.85)	<.0001
LBBB	73 (15.05)	38 (2.00)	<.0001
BNP			
BNP (pg/ml)	247.20 (±545.9)	435.7 (±722.60)	.003
BNP (>100 pg/ml)	110 (22.68)	64 (3.73)	<.0001

Abbreviations: BNP: B-type natriuretic peptide; bpm: beats per minute; LBBB: left bundle branch block; LVH: left ventricular hypertrophy; msec: milliseconds; pg/ml: picogram per milliliter; RBBB: right bundle branch block; RVH: right ventricular hypertrophy.

*All tests were chi-square for categorical variables and Student's t-tests for continuous variables at 0.05 significance level.

Table 2. Multivariable models using logistic regression backwards stepwise approach in the model derivation cohort (n = 2384)

Models	ROC c-statistic (95% CI)	Brier Score	BIC[a]
Non-blood biomarker model	0.96 (0.95, 0.97)	0.05	929.4
Simplified model **A**	0.96 (0.95, 0.97)	0.05	931.8
Simplified model **B**	0.96 (0.95, 0.97)	0.05	941.9
Simplified model **C**	0.96 (0.95, 0.97)	0.05	937.1
Full model	0.96 (0.95, 0.97)	0.05	924.9

Abbreviations: BIC: Bayesian Information Criterion; ROC: receiver operating characteristic.

[a]Lower BIC and Brier score values indicate better fit. Higher values of c-statistic indicate better discrimination.

scores, we also compared the BIC. The BIC value for the non-blood biomarker model (BIC = 929.4) was larger than that of the full model (BIC = 924.9) but lower than all other simplified models. The non-blood biomarker model was selected as the final algorithm, because it was the simplest model that maintained high performance based on the combination of a high c-statistic and low Brier score and BIC value relative to the full and simplified models. Additionally, since BNP is not consistently obtained on patients, this model was selected, because it did not contain a blood biomarker variable.

Table 3 summarizes the performance of the models in the validation dataset for the clinical and surveillance settings. When assessed

for the clinical setting, the non-blood biomarker model had a high sensitivity (94.6%), specificity (92.4%), and percent correctly classified (92.8%); the other models performed similarly. The performance statistics of the non-blood biomarker model when examined for the surveillance setting were also very high (sensitivity: 87.4%; specificity: 97.7%), as were the other models.

Retrospective chart reviews of false positive controls showed that none had the moderate-complex CHDs included in the study. Six controls were identified as having structural CHDs: 5 were mild (1 atrial septal defect, 2 bicuspid aortic valve, and 2 patent foramen ovale) and 1 was a moderate-complex CHD not included in the case group (total anomalous pulmonary venous return). The remaining positive controls had either conduction defects (eg right and/or left bundle branch block, complete heart block), rhythm defects (eg arrhythmias, tachycardias, bradycardias), cardiomyopathies (hypertrophic, dilated, ischemic), pericarditis, or no cardiac pathology.

The adult moderate-complex congenital heart defect (ACHD) risk score, which was derived from the non-blood biomarker model, is presented in Table 4. 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 recorded in the last column, "Points for each indicator." Adding the points of each predictor present provides each patient's final score.

An optimal threshold score of 10 was selected using the validation cohort dataset. This threshold was defined by a combination of

Table 3. Performance characteristics of the non-blood biomarker, simplified, and full models in the identification of adults with specific moderate or complex CHDs for the clinical (A) and surveillance (B) setting, validation cohort (n = 596)

A[a]	Sensitivity (%)	Specificity (%)	TP (%)	TN (%)	CC (%)
Non-blood biomarker model	94.6	92.4	73.9	98.7	92.8
Simplified model **A**	94.6	92.0	72.9	98.7	92.5
Simplified model **B**	94.6	92.6	74.5	98.7	93.0
Simplified model **C**	94.6	92.4	73.9	98.7	92.8
Full model	93.7	92.8	74.8	98.7	93.0
B[a]	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**[b]	**NPV (%)**[b]	**CC (%)**
Non-blood biomarker model	87.4	97.7	6.0	100.0	95.8
Simplified model A	87.4	97.7	6.0	100.0	95.8
Simplified model B	87.4	97.5	5.6	100.0	95.6
Simplified model C	86.5	97.5	5.5	100.0	95.7
Full model	87.4	97.7	6.0	100.0	95.8

Abbreviations: CC: correctly classified; CHDs: congenital heart defects; NPV: negative predictive value; PPV: positive predictive value; TN: true negative; TP: true positive.

[a]ROC C-table cutoff value for the clinical (A) and surveillance (B) settings were 0.15 and 0.35, respectively.

[b]PPV and NPV were calculated using an estimated population prevalence of 16.9 per 10 000 for the specific moderate or complex CHDs and a standard population of 10 000.

Table 4. Adult moderate or complex congenital heart defect (ACHD) score tool

Indicator	Categories	Points for Each Category	Points for Each Indicator
Demographics			
Age	20-29	7	Points:
	30-39	5	
	40-49	3	
	50-60	0	
Sex	Male	0	Points:
	Female	2	
Electrocardiogram			
QRS duration			Points:
Against	<80	0	
Neutral	80-119	3	
Support	120-149	6	
Likely	≥150	10	
Atrial enlargement right,	Absent	0	Points:
left, or biatrial	Present	3	
Rhythm not sinus	Absent	0	
	Present	4	
Right ventricular hypertrophy	Absent	0	
	Present	3	
Left ventricular hypertrophy	Absent	0	Points:
	Present	2	
Right bundle branch block	Absent	0	Points:
	Present	4	
Left bundle branch block	Absent	0	Points:
	Present	2	
			Total= [a]

[a]Threshold score is 10.

slightly higher sensitivity than specificity in order to identify potential high-risk individuals for further cardiology specialty care assessment. This optimal threshold score was associated with a sensitivity of 96.4%, specificity of 80.0%, and percent correctly classified of 83.1% for the outcome (Supplementary Table S5).

CONCLUSION

We developed an empirically-derived risk score using specific clinical variables—age, sex, EKG parameters—found in EHRs to differentiate between adults with and those unlikely to have specific moderate-complex CHDs. The final algorithm (non-blood biomarker model) demonstrated good calibration and discriminatory power in a randomly-generated split-sample internal validation cohort. The final model correctly identified adult patients with specific moderate-complex CHDs with a 95% sensitivity and a 92% specificity for the clinical setting and 87% sensitivity and 98% specificity for the surveillance setting. The non-blood biomarker model was subsequently used to develop the ACHD risk score tool for use by clinicians to identify individuals with characteristics and EKG signs suggesting the presence of one of the specific moderate-complex CHDs; the risk score tool similarly demonstrated high sensitivity (96%) and specificity (80%).

The primary premise for our model is that EKG abnormalities have long been reported as associated with post-surgical repair of CHDs, particularly with conduction issues, such as elongated QRS duration and bundle branch block. EKG anomalies, such as tall R-waves that could result from right ventricular pressure overload and wide QRS duration of over 180 msec predicting sudden death in patients with tetralogy of Fallot,[21] are examples of how EKG parameters have been guiding clinicians in identifying and monitoring patients with CHDs. The high sensitivity and specificity results of our model further demonstrate the accuracy of EKG readings in identifying patients with specific moderate-complex CHDs.

The ACHD risk score tool, developed from the algorithm for the clinical setting, has helpful clinical applications. Most contemporary EKG machines provide an automated "read" with flags to alert providers to potential specific conditions. Examples of flags include "pulmonary disease pattern" or "acute myocardial infarction." As many adults with specific moderate-complex CHDs are seen in emergency rooms and primary care offices, the addition of an EKG flag to alert providers to potential "probable CHD" would create opportunities to improve care. These would include flagging potential high-risk individuals for further cardiology assessment, bringing some of those lost to follow-up back into adult CHD specialty care,

and avoiding misclassification with other cardiac pathology, for example, myocardial infarction, thus preventing unnecessary testing.

Like other initiatives that use EHR data to establish multicenter research cohorts (eg the Electronic Medical Records and Genomics [eMERGE Network]),[22] the algorithm that uses the surveillance setting to distinguish adults with and without the specific commonly occurring moderate-complex CHDs could be employed to establish multicenter surveillance of adults with these CHDs who might be receiving care outside of cardiac care specialty clinics. Identifying individuals through the use of this algorithm could lead to improved understanding of health services utilization, morbidity, mortality, and healthcare disparities in this population, more so than what can be gleaned from a potentially-biased sample of patients attending a cardiac care specialty clinic. Such population-based samples could additionally aid in studying the impact of secondary prevention on clinical outcomes and quality of life in this expanding population. Our algorithm might also aid the estimation of basic epidemiologic measures of patients with moderate-complex CHDs, such as national prevalence and distributions from cross-sectional surveys that include EKG measures (eg National Health and Nutrition Examination Survey).

Given the high costs associated with hospitalizations and surgery, and the high proportion of adults with specific moderate-complex CHDs not receiving specialized care, it is in the financial interests of public health, insurance companies, and healthcare providers in the United States to identify adults with specific moderate-complex CHDs. This could ensure timely provision of care and education to prevent expensive cardiac interventions.[23]

Although administrative data are regularly used to estimate disease incidence and prevalence, EHR systems provide added value as they incorporate structured clinical data, including test results, which facilitate accurate identification of patients.[24–26] Administrative algorithms that rely on billing codes (specifically ICD codes) require a number of patient encounters with providers to ascertain the condition of interest, and the codes are not always available or accurate.[18,27,28] Indeed, in our study, we used the lack of ICD-9-CM codes to define controls without the specific moderate-complex CHDs, which may have misclassified a few cases as controls. However, we did not find any such misclassification in our retrospective chart review. Therefore, algorithms based on combined administrative and clinical data may be the most accurate and informative for identifying adults with the specific moderate-complex CHDs and tracking outcomes.[29] Although success of such algorithms depends on the quality of data stored in EHR systems, the combination of clinical and administrative data are likely the most viable option for identifying adults with both high sensitivity and specificity.[30–32]

As the population of adults with specific moderate-complex CHDs ages, the number of patients in the above 40 age-group will also increase, leading to dilution of our model's use of age for categorizing patients with and without specific moderate-complex CHDs, leaving sex and EKG variables as the sole identification parameters. Therefore, incorporating a single test (ie EKG) into the identification algorithm becomes logical since such an abbreviated algorithm has the potential to provide an easy screening modality in clinical settings that lack the ability to identify adults with specific moderate-complex CHDs.

Other scoring systems have been developed for assessing outcomes for patients with CHDs. For example, several surgical risk scoring systems for the prediction of mortality, major adverse events, and prolonged lengths of stay among pediatric-aged patients with CHDs have demonstrated good predictive ability among adults

with CHDs undergoing congenital heart surgery.[33] However, to the best of our knowledge, our model is the first proof of concept that one's likelihood of having one of the specific moderate-complex CHDs can be identified based on age, sex, and EKG parameters obtained from an EHR system. Moreover, the performance statistics of our algorithm suggest high potential for wider use. In addition to flagging adult patients with specific moderate-complex CHDs during routine visits when programmed in EKG machines and, potentially, recommend them specialized care, this algorithm, which defines a set of data elements and logical expressions, can serve as the bases for the development of a computable phenotype.[34] This phenotype can be programmed into EHR systems, validated across research centers and deployed to establish multi-center registries to better measure prevalence and understand the health needs and outcomes of patients with specific moderate-complex CHDs to improve care and services.[22]

BNP, which is not obtained routinely, did not provide added predictive value in the algorithm, despite its statistical significance. This finding was consistent with conclusions from other studies that report increases in BNP concentration can be associated with complex CHDs but are best used for specific clinical reasons, such as guiding therapy.[35,36]

There are some limitations to our work. We selected a limited number of CHDs compared to the full spectrum recorded in the Emory EHR system. However, our selection is based on commonly occurring CHDs in order to demonstrate proof of concept. The non-blood biomarker algorithm and the ACHD risk tool that were developed are not yet generalizable to populations outside the Emory Healthcare setting as external data validation must follow.[37] The use of ICD-9-CM codes to exclude controls who might have a CHD 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-9-CM codes in EHR systems.[18] Furthermore, longitudinal studies that utilize clinical characterization of adults with and without specific moderate-complex CHDs will be important to verify the accuracy of the ACHD risk score.

The algorithm that was developed to distinguish adults with and without specific moderate-complex CHDs—composed of only age, sex, and EKG markers—provided within a validation cohort either a 95% sensitivity and 92% specificity, or an 87% sensitivity and 98% specificity, for use in a clinical or surveillance setting, respectively. Although the EHR-based algorithm requires further external validation before incorporation into EKG machines, it has the potential to identify patients lost to cardiology follow-up, establish multicenter population-based surveillance and epidemiologic assessment of adults with specific moderate-complex CHDs, and flag potential high-risk individuals for further cardiology assessment.

ABBREVIATIONS

ACHC: Emory Adult Congenital Heart Center
BNP: B-type natriuretic peptide
BIC: Bayesian Information Criterion
CHD: congenital heart defect
EKG: electrocardiogram
EHR: electronic health records
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
NPV: negative predictive Value
PPV: positive predictive value
ROC: receiver operating characteristic

CDS 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 (CDC).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

COMPETING INTERESTS

None.

REFERENCES

1. Reller MD, Strickland MJ, Riehle-Colarusso T, *et al*. Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *J Pediatr* 2008; 153 (6): 807–13.
2. Warnes CA, Liberthson R, Danielson GK, *et al*. Task force 1: the changing profile of congenital heart disease in adult life. *J Am Coll Cardiol* 2001; 37 (5): 1170–5.
3. Oster ME, Lee KA, Honein MA, *et al*. Temporal trends in survival among infants with critical congenital heart defects. *Pediatrics* 2013; 131 (5): e1502–8.
4. Gilboa SM, Devine OJ, Kucik JE, *et al*. Congenital heart defects in the United States: estimating the magnitude of the affected population in 2010. *Circulation* 2016; 134 (2): 101–9.
5. Khairy P, Ionescu-Ittu R, Mackie AS, *et al*. Changing mortality in congenital heart disease. *J Am Coll Cardiol* 2010; 56 (14): 1149–57.
6. Verheugt CL, Uiterwaal CS, van der Velde ET, *et al*. Mortality in adult congenital heart disease. *Eur Heart J* 2010; 31 (10): 1220–9.
7. Cohen SB, Ginde S, Bartz PJ, *et al*. Extracardiac complications in adults with congenital heart disease. *Congenit Heart Dis* 2013; 8: 370–80.
8. Opotowsky AR, Siddiqi OK, Webb GD. Trends in hospitalizations for adults with congenital heart disease in the U.S. *J Am Coll Cardiol* 2009; 54 (5): 460–7.
9. Warnes CA, Williams RG, Bashore TM, *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). Developed in Collaboration With the American Society of Echocardiography, Heart Rhythm Society, International Society for Adult Congenital Heart Disease, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *J Am Coll Cardiol* 2008; 52: e143–263.
10. Mylotte D, Pilote L, Ionescu-Ittu R, *et al*. Specialized adult congenital heart disease care: the impact of policy on mortality. *Circulation* 2014; 129 (18): 1804–12.
11. Mackie AS, Ionescu-Ittu R, Therrien J, *et al*. Children and adults with congenital heart disease lost to follow-up: who and when? *Circulation* 2009; 120 (4): 302–9.
12. Winter MM, Mulder BJ, van der Velde ET. 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): e252; author reply e253.
13. Warnes CA. The adult with congenital heart disease: born to be bad? *J Am Coll Cardiol* 2005; 46 (1): 1–8.
14. Reid GJ, Irvine MJ, McCrindle BW, *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): e197–205.
15. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med* 2010; 363 (6): 501–4.
16. Hsiao CJ, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2012. *NCHS Data Brief* 2012; 111: 1–8.
17. Kleinbaum DG, Kupper LL, Nizam A, *et al*. *Applied Regression Analysis and Other Multivariate Methods*. 4th ed. Belmont, CA, USA: Thomson Books/Cole; 2008.
18. Broberg C, McLarry J, Mitchell J, *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): 719–25.
19. Kleinbaum DG, Klein M. *Logistic Regression: A Self-Learning Text*. 3rd ed. New York, NY: Springer Publishers; 2010.
20. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med* 2004; 23 (10): 1631–60.
21. Gatzoulis MA, Balaji S, Webber SA, *et al*. Risk factors for arrhythmia and sudden cardiac death late after repair of tetralogy of Fallot: a multicentre study. *Lancet* 2000; 356 (9234): 975–81.
22. Kullo IJ, Haddad R, Prows CA, *et al*. Return of results in the genomic medicine projects of the eMERGE network. *Front Genet* 2014; 5: 50.
23. Ministeri M, Alonso-Gonzalez R, Swan L, Dimopoulos K. Common long-term complications of adult congenital heart disease: avoid falling in a H.E.A.P. *Expert Rev Cardiovasc Ther* 2016; 14 (4): 445–62.
24. Bobo WV, Pathak J, Kremers HM, *et al*. An electronic health record driven algorithm to identify incident antidepressant medication users. *J Am Med Inform Assoc* 2014; 21 (5): 785–91.
25. Deshpande AD, Schootman M, Mayer A. Development of a claims-based algorithm to identify colorectal cancer recurrence. *Ann Epidemiol* 2015; 25 (4): 297–300.
26. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008; 77 (5): 291–304.
27. Benchimol EI, Guttmann A, Griffiths AM, *et al*. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 2009; 58 (11): 1490–7.
28. Benchimol EI, Guttmann A, Mack DR, *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): 887–96.
29. Tang PC, Ralston M, Arrigotti MF, Qureshi L, *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): 10–5.

30. Byrd JB, Vigen R, Plomondon ME, *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. *Am Heart J* 2013; 165 (3): 434–40.

31. Staroselsky M, Volk L, Tsurikova R, *et al*. Improving electronic health record (EHR) accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input. *Int J Med Inform* 2006; 75 (10–11): 693–700.

32. Widdifield J, Ivers NM, Young 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* 2015; 21 (8): 1045–54.

33. Kogon B, Oster M. Assessing surgical risk for adults with congenital heart disease: are pediatric scoring systems appropriate? *J Thorac Cardiovasc Surg* 2014; 147 (2): 666–71.

34. Richesson R, Smerek M, Rusincovitch S, *et al*. Electronic health records-based phenotyping. 2014. http://rethinkingclinicaltrials.org/resources/ehr-phenotyping/ Accessed August 5, 2018.

35. Eindhoven JA, van den Bosch AE, Jansen PR, *et al*. The usefulness of brain natriuretic peptide in complex congenital heart disease: a systematic review. *J Am Coll Cardiol* 2012; 60 (21): 2140–9.

36. Vuolteenaho O, Ala-Kopsala M, Ruskoaho H. BNP as a biomarker in heart disease. *Adv Clin Chem* 2005; 40: 1–36.

37. Bjornard K, Riehle-Colarusso T, Gilboa SM, *et al*. Patterns in the prevalence of congenital heart defects, metropolitan Atlanta, 1978 to 2005. *Birth Defects Res A Clin Mol Teratol* 2013; 97: 87–94.