Evaluation of a Prediction Model for the Development of Atrial Fibrillation in a Repository of Electronic Medical Records

Matthew J. Kolek, MD, MSci; Amy J. Graves, SM, MPH; Meng Xu, MS; Aihua Bian, MPH; Pedro Luis Teixeira, MS; M. Benjamin Shoemaker, MD, MSci; Babar Parvez, MD, MSci; Hua Xu, PhD; Susan R. Heckbert, MD, PhD; Patrick T. Ellison, MD, PhD; Emelia J. Benjamin, MD, ScM; Alvaro Alonso, MD, PhD; Joshua C. Denny, MD; Karel G. Moons, PhD; Ayumi K. Shintani, PhD, MPH; Frank E. Harrell Jr, PhD; Dan M. Roden, MD; Dawood Darbar, MD

**IMPORTANCE** Atrial fibrillation (AF) contributes to substantial morbidity, mortality, and health care expenditures. Accurate prediction of incident AF would enhance AF management and potentially improve patient outcomes.

**OBJECTIVE** To validate the AF risk prediction model originally developed by the Cohorts for Heart and Aging Research in Genomic Epidemiology–Atrial Fibrillation (CHARGE-AF) investigators using a large repository of electronic medical records (EMRs).

**DESIGN, SETTING, AND PARTICIPANTS** In this prediction model study, deidentified EMRs of 33,494 individuals 40 years or older who were white or African American and had no history of AF were reviewed and analyzed. The participants were followed up in the internal medicine outpatient clinics at Vanderbilt University Medical Center for incident AF from December 31, 2005, until December 31, 2010. Adjusting for differences in baseline hazard, the CHARGE-AF Cox proportional hazards model regression coefficients were applied to the EMR cohort. A simple version of the model with no echocardiographic variables was also evaluated. Data were analyzed from October 31, 2013, to January 31, 2014.

**MAIN OUTCOMES AND MEASURES** Incident AF. Predictors in the model included age, race, height, weight, systolic and diastolic blood pressure, treatment for hypertension, smoking status, type 2 diabetes, heart failure, history of myocardial infarction, left ventricular hypertrophy, and PR interval.

**RESULTS** Among the 33,494 participants, the median age was 57 (interquartile range, 49-67) years; 57% of patients were women, 43% were men, 85.7% were white, and 14.3% were African American. During the mean (SD) follow-up of 4.8 (0.9) years, 2,455 individuals (7.3%) developed AF. Both models had poor calibration in the EMR cohort, with underprediction of AF among low-risk individuals and overprediction of AF among high-risk individuals (10th and 90th percentiles for predicted probability of incident AF, 0.005 and 0.179, respectively). The full CHARGE-AF model had a C index of 0.708 (95% CI, 0.699-0.718) in our cohort. The simple model had similar discrimination (C index, 0.709; 95% CI, 0.699-0.718; P = .70 for difference between models).

**CONCLUSIONS AND RELEVANCE** Despite reasonable discrimination, the CHARGE-AF models showed poor calibration in this EMR cohort. This study highlights the difficulties of applying a risk model derived from prospective cohort studies to an EMR cohort and suggests that these AF risk prediction models be used with caution in the EMR setting. Future risk models may need to be developed and validated within EMR cohorts.
Atrial fibrillation (AF), the most common sustained cardiac arrhythmia, is becoming increasingly prevalent in the Western world.1,2 The number of patients with AF in the United States is projected roughly to double by the year 2050, to an estimated 12 to 16 million persons.2,3 Atrial fibrillation is associated with significant morbidity,4,5 mortality,6-9 decreased quality of life,10 and increased health care expenditures.11,12 Developing strategies for the prediction and prevention of AF in high-risk individuals remains an underexplored and important area of research.13

In 2012, the Cohorts for Heart and Aging Research in Genomic Epidemiology-Atrial Fibrillation (CHARGE-AF) investigators developed and validated a risk model for prediction of incident AF.14 The model was developed using pooled data from prospective cohort studies, including the Atherosclerosis Risk in Communities Study,15 Cardiovascular Health Study,16 and Framingham Heart Study,17 and was validated in the Age Gene/Environment Susceptibility–Reykjavik study and Rotterdam Study cohorts. The model is especially well suited for primary care settings because it does not require laboratory or echocardiographic (ECG) variables.

Novel risk models should be validated (ie, evaluated in new settings) before they are incorporated into routine care.20 Electronic medical records (EMRs) are becoming ubiquitous in clinical practice, and 1 potential use for EMR repositories in etiologic research is to validate existing risk prediction models. In addition, risk models are unlikely to be widely used unless they can be incorporated into EMR systems. We therefore evaluated the CHARGE-AF risk model for incident AF in a large, de-identified EMR repository.

Methods

Study Population
The study population was selected from a deidentified version of the Vanderbilt University Medical Center EMR (hereinafter referred to as the Vanderbilt EMR). This resource, termed the Synthetic Derivative,21 consists of deidentified medical records of Vanderbilt University Medical Center, Nashville, Tennessee, inpatients and outpatients; as of December 31, 2015, it contained nearly 2.6 million individuals. The Synthetic Derivative consists of the deidentified version of the Vanderbilt EMR that has been judged by the Vanderbilt University institutional review board as falling under the designation of nonhuman subjects under the Common Rule (45 CFR Part 46); therefore, this study and other Synthetic Derivative research were deemed exempt by the Vanderbilt University institutional review board.

To ensure that they had adequate follow-up, individuals met criteria for a medical home model, in which they were followed up in a Vanderbilt University internal medicine clinic with at least 3 visits documented within a 24-month period.22 Other criteria for entry into the study included being 40 years or older, being self-identified as white or African American race, and no known history of AF as of December 31, 2005. Individuals were excluded if they had billing codes for AF or there was mention of AF in ECG impressions or structured problem lists as determined by natural language processing.23 Individuals were also excluded from the study if they had International Classification of Diseases, Ninth Revision (ICD-9), or Current Procedural Terminology (CPT) codes for heart transplant at the beginning of follow-up. Data studied included all inpatient and outpatient ICD-9 and CPT codes, ECGs, and problem lists, and manual review included all inpatient and outpatient records.

Outcome
The follow-up period for incident AF was from December 31, 2005, until December 31, 2010. Ascertainment of incident AF was accomplished by using a validated algorithm that incorporates natural language processing and billing codes, as previously described.21 This automated algorithm was optimized through multiple reiterations with sensitivity analyses with different cutoffs and manual review of medical records until a positive predictive value for AF of greater than 95% was achieved. Cases were defined by natural language processing of cardiologist-interpreted ECG impressions, 4 or more occurrences of ICD-9 codes for AF, or AF instances recorded in the problem list. To be classified as free of AF, an individual record could not contain mention of AF in the ECG impressions or structured problem lists or ICD-9 codes for AF or atrial flutter. The 4 ICD-9 codes used in the automated AF algorithm are the most commonly used codes for AF or atrial flutter. Individuals with 1 to 3 ICD-9 codes for AF or atrial flutter were excluded from the study cohort. In effect, these steps resulted in a more sensitive method for excluding AF at baseline and a more specific method for ascertaining incident AF during follow-up. We assessed the accuracy of the AF algorithm in the present study by manually reviewing the full EMR in a random set (enriched for incident AF by selecting roughly equal proportions of individuals with and without AF) of 200 incident AF cases and control individuals block randomized to case-control ordering, as is common for validating EMR phenotypes.24

Predictors
CHARGE-AF model predictors were ascertained from records available from January 1 to December 31, 2005. Sex, race, age, weight, height, body mass index, and systolic and diastolic
blood pressure were directly extracted from structured fields in the Synthetic Derivative. History of myocardial infarction, heart failure, and type 2 diabetes were determined by using ICD-9 codes, incorporating laboratory values and medication records. Treatment for hypertension was assessed using a previously validated algorithm incorporating medication records in the Synthetic Derivative. This algorithm was previously shown to have sensitivity and positive predictive values of 88% and 93%, respectively. We obtained PR interval and left ventricular hypertrophy data from outpatient ECG reports. Current smoking status was determined by using an existing algorithm with a reported positive predictive value of 93% in the Vanderbilt EMR.

**Statistical Analysis**

Baseline characteristics present at the beginning of the observation period are presented in terms of median (interquartile range) for continuous variables and frequencies with percentage for categorical variables. The number of nonmissing values for each variable is also given. Single imputation of missing values was performed using predictive mean matching. To assess the validity of using single rather than multiple imputation, we conducted 5 separate imputations. These resulted in almost identical C indices (eTable 1 in the Supplement).

To evaluate the CHARGE-AF model in the Vanderbilt Synthetic Derivative cohort, we used the Cox proportional hazards regression model derived in the CHARGE-AF study (Table 1). As in the original CHARGE-AF study, we compared discrimination using the full model and a simple model that did not incorporate ECG variables. Comparisons used the rcorr.cens function in the R Hmisc package (R Foundation for Statistical Computing), which provides a test of whether one model gives predictions that are more concordant than the other in a way that preserves pairing of 2 predictions in the pair of patients under consideration.

We adjusted for differences in baseline hazard between the original CHARGE-AF data set and our cohort by constructing the linear predictor for each observation using the original coefficients from the CHARGE-AF model and the covariate values in the Vanderbilt Synthetic Derivative cohort data, with replacement of the mean covariate values to reflect the Vanderbilt data. We then examined the calibration and discrimination of the CHARGE-AF model in our cohort. To assess calibration, the observed risk for developing AF during the study period was plotted against the predicted risk. The generated curve was compared against a hypothetical ideal curve with a slope of 1 and intercept of 0. For discrimination, which measures the ability of the model to distinguish individuals who will develop AF from those who will not, we calculated the continuous-time C index with censoring. We also constructed a Kaplan-Meier curve showing cumulative incidence of AF in our cohort. Statistical analyses were conducted from October 31, 2013, to January 31, 2014, using R software (version 3.0.2; R Foundation for Statistical Computing). A 2-sided $P \leq .05$ was considered statistically significant. The detailed statistical code is included in eTable 1 in the Supplement.

**Results**

Based on the prospectively defined criteria, 33,494 individuals were included in the analysis. Baseline characteristics for the study cohort are presented in Table 2. Median age was 57 (interquartile range, 49-67) years; 57% were women, 43% were men, 85.7% were white, and 14.3% were African American. During the mean (SD) follow-up of 4.8 (0.9) years, 2455 individuals (7.3%) developed AF. A Kaplan-Meier curve for the cumulative incidence of AF is shown in Figure 1.

The calibration curve plotting the predicted probability of AF-free survival and the observed AF-free survival in our cohort indicated that the model had poor fit (Figure 2). We found underprediction of AF for individuals with lower (<0.15) predicted probabilities. This group included most of the individuals in the cohort. We also found overprediction of AF for higher probabilities ($\geq$0.15). The 10th and 90th percentiles for predicted probability of incident AF were 0.005 and 0.179, respectively. Across this range, the maximum calibration error was 0.0526 (5.3%).

The full CHARGE-AF model had a C index of 0.708 (95% CI, 0.699-0.718) in our cohort. The simple model, which did not include ECG predictors (PR interval and left ventricular hypertrophy), had discrimination similar to that of the full model (C Index, 0.709; 95% CI, 0.699-0.718; $P = .70$ for difference between the models).

Because the primary study outcome, incident AF during the 5-year follow-up period, was ascertained by an automated algorithm, we conducted a manual review of indi-
To assess the accuracy of the algorithm, we selected 200 records at random after enrichment of the sample for incident AF. A manual medical record review revealed that 88 individuals had incident AF during the 5-year follow-up period (eTable 2 in the Supplement). The sensitivity of the AF algorithm was 96.5% (95% CI, 90.1%-98.8%); specificity, 94.8% (95% CI, 89.1%-97.6%); positive predictive value, 93.2% (95% CI, 85.9%-96.8%); and negative predictive value, 97.3% (95% CI, 92.4%-99.1%). The 6 individuals who were not identified as having incident AF by the automated algorithm but who had AF on manual medical record review were each identified as having the arrhythmia based on mention of it in the clinic notes, although AF billing codes were not present. It was less clear why 3 individuals were identified as having incident AF by the algorithm but not confirmed on manual review, although incomplete medical record review and incorrect billing codes are possible explanations.

Because increasing age is the strongest predictor of AF in the CHARGE-AF model, we evaluated whether the model predicts incident AF better than knowledge of age alone. We generated a Cox proportional hazards regression model for the development of AF at 5 years in our cohort with age as the only dependent variable, resulting in a C index of 0.684 (95% CI, 0.674-0.694). The C indices for the Vanderbilt University age-only model and the externally validated CHARGE-AF model could not be compared directly because the models are not nested. We also generated a scatterplot of age and cumulative predicted probability of AF at 5 years (Figure 3). This model showed a broad distribution of AF risk across the spectrum of ages, indicating that age alone is an imprecise predictor of AF risk.

**Discussion**

We evaluated the CHARGE-AF full and simple risk models in a large cohort of individuals within the Vanderbilt EMR repository (Synthetic Derivative). The full model had a C index of 0.708 (95% CI, 0.699-0.718). The simple CHARGE-AF model, which did not include ECG variables as predictors, had similar discrimination, with a C index of 0.709 (95% CI, 0.699-0.718). However, calibration for both models was poor in our cohort, indicating a failure
of validation. Our study represents a novel use of an EMR repository to evaluate an existing AF risk model and illustrates the limitations of applying a model developed in prospective cohort studies to a real-world EMR context.

Our findings illustrate several potential uses for EMR repositories in biomedical research. First, EMR repositories could serve as an inexpensive and efficient complement to community cohort studies for the development of prediction models. Second, an EMR repository could be used as an independent cohort to externally evaluate an existing model, as we did here. In fact, risk models are unlikely to be widely used unless they can be incorporated into EMR systems. Finally, given that EMRs are integrated into clinical practice, prediction models could be incorporated into these systems to prospectively identify individuals at high risk for AF or other diseases, with the ultimate goal of developing individualized preventive strategies. Specifically, improved knowledge about individual AF risk might enable aggressive risk factor modification, more intensive screening, diligent evaluation at the first sign of symptoms, and modification of stroke risk.

The CHARGE-AF model has recently been tested in additional community cohorts. When applied to the Multi-Ethnic Study of Atherosclerosis cohort,33 the simple CHARGE-AF model had good discrimination (C index, 0.779; 95% CI, 0.744-0.814) but suboptimal calibration, with overprediction of AF in higher-risk individuals. When applied to more than 24,000 participants in the European Prospective Investigation of Cancer Norfolk cohort,34 the CHARGE-AF simple model again had good discrimination (C index, 0.81; 95% CI, 0.75-0.85) but also poor calibration. These studies, along with our current findings, illustrate the difficulty of applying a risk model to diverse populations, particularly in an EMR setting.

Large, prospective community cohort studies, such as the Atherosclerosis Risk in Communities Study, the Cardiovascular Health Study, Framingham Heart Study, Age Gene/Environment Susceptibility–Reykjavik study, and Rotterdam Study, have been instrumental in identifying risk factors for common diseases.35-39 In the case of coronary heart disease, the discoveries of these studies have been translated into strategies for primary and secondary prevention that have had important effects on cardiovascular morbidity and mortality.40-45 Electronic medical record repository studies might emerge as an important complement to prospective cohort studies. Although EMR repositories might have important shortcomings, including inadequate disease classifications and missing data, they also offer several advantageous that could be leveraged for etiologic research. Because data in EMR repositories are collected during routine clinical care, the cost of these studies is small relative to that of prospective cohort studies. Notably, the National Heart, Lung, and Blood Institute recently announced an initiative to use large EMR studies to enhance the clinical utility and reproducibility of clinical research while reducing costs.46 We propose that automated algorithms could be deployed within EMR repositories to prospectively identify and flag individuals at high risk for AF or other common diseases. These data could then be used to guide primary prevention strategies. Peterson et al47 pursued a similar strategy to identify and genotype individuals at risk for receiving medications that have pharmacogenetic variations in efficacy. Although no specific treatment for the primary prevention of AF has been established,33 angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers have been associated with a decreased incidence of AF in post hoc analyses of randomized clinical trials and in retrospective cohort studies.48,49

Our study has several important limitations. One of these relates to data collection and ascertainment. Although all patient variables were entered into the Vanderbilt EMR prospectively, the nature of the EMR with data entered by multiple users might lead to more inaccuracies when compared with carefully curated prospective cohorts, such as those studied by the CHARGE-AF consortium. Despite a 12-month run-in period from January 1 until December 31, 2005, when individuals with incident AF were excluded from the final study cohort, we did not conduct rigorous screening (eg, ambulatory ECG monitoring) to exclude baseline AF. In addition, patients with AF may not seek medical attention for more than 12 months, and a longer run-in period may be needed. Predictor and outcome variables were extracted from the EMR repository using automated algorithms. Although many of these variables (eg, age, sex, race, height, weight, body mass index, blood pressure, and PR interval) are structured data fields within the repository, other predictors, such as type 2 diabetes and heart failure, depend on billing codes, laboratory or note data, and medication records, potentially resulting in important inaccuracies. The automated algorithm for assignment of incident AF relied primarily on natural language processing of ECG impressions, problem lists, and clinic notes but also used billing codes. We conducted a manual review of a sample of medical records that demonstrated good performance of the algorithm for ascertainment of AF status, with high sensitivity and specificity.
Our AF algorithm with a high diagnostic accuracy that exceeds those used by the CHARGE-AF cohorts may explain in part the poor calibration of the risk prediction score in an EMR setting.

Because individuals in our study were not prospectively enrolled and followed up for incident AF, our analysis is prone to indication bias, wherein individuals who developed AF may have had more clinical encounters than those who did not. Because individuals were not followed up at prespecified intervals, our results might be influenced by loss of follow-up. In addition, the classification of incident AF might have been inaccurate if individuals sought care outside Vanderbilt University Medical Center such that their AF diagnoses were not captured in the Vanderbilt EMR.

Although the CHARGE-AF models had satisfactory discrimination in our cohort, calibration was poor. One potential reason for the poor calibration is important differences in the characteristics of the CHARGE-AF discovery cohorts and our cohort. A direct comparison of baseline characteristics between our EMR cohort and the CHARGE-AF cohorts is difficult because 5 separate community cohorts were used to formulate the CHARGE-AF model. However, overall, individuals in our cohort tended to be younger, heavier, more likely to be smokers, and more likely to use antihypertensive therapy and generally tended to be sicker than those enrolled in the CHARGE-AF cohorts. Because age is the most important predictor of AF, the inclusion of younger individuals in our study might account, at least in part, for the poor calibration of the CHARGE-AF model in our EMR cohort. We chose the age cutoff for inclusion in our study based on previous findings that the incidence of AF begins to increase rapidly after 40 years of age. Our goal was to be able to apply the CHARGE-AF model in a primary care setting, and we postulate that it is particularly important to identify relatively young patients at risk for AF because they might benefit most from preventative measures, although postulation remains to be proven in prospective studies. Additional potential causes for failure of the models to accurately predict the development of AF include loss to follow-up (including death), differences in how predictors were defined, and more structured surveillance for incident AF in the CHARGE-AF cohorts (greater sensitivity). Owing to the nature of our study, we were unable to include comprehensive data on death (ie, using the Social Security Death Index) other than what was available in the EMR.

Conclusions

We evaluated the CHARGE-AF full and simple risk models in a large cohort of individuals without a history of AF at baseline in the Vanderbilt EMR repository. The models performed poorly in our EMR cohort, illustrating the difficulty of applying risk models developed within prospective cohort studies to a real-world EMR context. Risk models for the development of AF or other complex disorders are unlikely to be widely used in clinical care unless they can be incorporated into EMR systems. Risk models, therefore, should be derived from and validated in different EMR cohorts, with the goal of prospectively and automatically identifying individuals at high risk for AF and implementing personalized strategies for primary prevention.

REFERENCES