A Comparison of Patient History- and EKG-based Cardiac Risk Scores

Andrew C. Miller, Ph.D.[1], **Ziad Obermeyer, M.D., M.Phil.**[2], **Sendhil Mullainathan, Ph.D.**[3]
[1]**Columbia University, New York, NY, USA**
[2]**University of California, Berkeley, CA, USA**
[3]**University of Chicago, Chicago, IL, USA**

Abstract

Patient-specific risk scores are used to identify individuals at elevated risk for cardiovascular disease. Typically, risk scores are based on patient habits and medical history — age, sex, race, smoking behavior, and prior vital signs and diagnoses. We explore an alternative source of information, a patient's raw electrocardiogram recording, and develop a score of patient risk for various outcomes. We compare models that predict adverse cardiac outcomes following an emergency department visit, and show that a learned representation (e.g. deep neural network) of raw EKG waveforms can improve prediction over traditional risk factors. Further, we show that a simple model based on segmented heart beats performs as well or better than a complex convolutional network recently shown to reliably automate arrhythmia detection in EKGs. We analyze a large cohort of emergency department patients and show evidence that EKG-derived scores can be more robust to patient heterogeneity.

1 Introduction

Heart disease annually claims the lives of over 600,000 people in the United States and over 17 million people worldwide[1;2]. Cardiovascular disease is varied. Coronary heart disease can progress to heart attack or stroke. Cardiac arrythmias, such as atrial fibrillation, are strongly associated with stroke, ventricular fibrillation, and sudden cardiac death — patients with atrial fibrillation have a $2.5\times$ increased risk of sudden cardiac death[3]. Early detection of cardiovascular disease is critical, and accurate characterization of patient risk can be used to improve care.

Risk models for a variety of cardiovascular diseases are typically based on patient attributes, behaviors, and medical history. Common predictors include age, gender, blood pressure, cholesterol level, smoking habits, and history of heart disease or hypertension. Both atrial fibrillation and general cardiovascular disease risk scores have been developed based on these common patient characteristics[4;5;6].

In this work, we study the use of raw electrocardiogram (EKG) waveform data to assess the risk of future cardiac disease. Risk scores derived from an EKG measurement do not depend on patient history and can be more generally applicable to new patients, patients with missing health information, or patients from clinically underserved populations. We compare risk scores based on traditional factors (e.g. age, gender, history of heart disease, hypertension, high cholesterol, etc.) to those derived purely on EKG waveforms, as well as the combination of the two sources.

We develop predictive risk scores on a dataset of 48,777 emergency department (ED) encounters from 33,806 unique patients and validate models on a separate set of 12,715 ED encounters from a separate set of 9,658 patients. These risk scores predict a variety of outcomes: development of atrial fibrillation, stroke within six months, a major adverse cardiac event (MACE) within six months (i.e. heart attack or revascularization), and the result of a troponin lab — a common blood test — at multiple time horizons. Reliable risk assessment of these outcomes can lead to a more efficient allocation of time and resources and better decision-making.

We show that the EKG waveform provides different (and often complementary) information about patient risk. For some outcomes (e.g. MACE and troponin labs) the raw EKG waveform was both more accurate and more robust to heterogeneity in the underlying population. For others (e.g. atrial fibrillation and stroke) EKG waveform data do not outperform simple patient history-based risk scores in the general population, but behave quite differently (and more robustly) in different patient subpopulations.

We also compare the performance of two deep EKG-based predictors: (i) a convolutional residual neural network that has recently been shown to automate arrhythmia detection in EKGs[7], and (ii) a simple multi-layer perceptron model applied to segmented beats that we develop in this work. We find that the simple model performs as well as or better than the complicated residual network in the prediction problems we study.

2 Background

Cardiac risk factors Assessing risk of heart disease is a well-studied area. For atrial fibrillation, risk models based on a small number of variables have been shown to be predictive in long time horizons. A model based on age, race, height, weight, blood pressure, current smoking, antihypertensive medication, diabetes, and a history of myocardial infarction and heart failure was shown to have reasonable predictive power (validation C-statistic/AUC of .70)[4]. Notably, it was also shown that simple features derived from electrocardiograms — PR interval and left ventricular hypertrophy — did not improve predictive performance[4]. However, it is believed that long term heart rhythm monitoring could improve early detection of asymptomatic atrial fibrillation[5;8].

For general cardiovascular risk, the Framingham risk score is based on a small number of patient predictors[9]. In heterogeneous populations, however, it can struggle to achieve good predictive performance. Another study found relatively low predictive capacity of the Framingham-based risk scores in an 8-year followup study[10], with C-statistic/AUCs in the range of .577 to .583.

Additionally, we are motivated to develop EKG-based risk assessments because they do not explicitly rely on patterns of health-seeking behavior. As argued in [11], predictive models based on patterns of health care consumption could inadvertently increase the allocation of resources to those who access the health care system the most. EHR-based predictors may, in part, predict future consumption of health care, and decisions based on these risk scores may disproportionately benefit those who can frequently seek treatment. Thus, it is valuable to explore alternative sources of information about a patient's health to build predictive models.

Clinical modeling Statistical machine learning techniques are used to find patterns in structured, high-dimensional signals. These tools are deployed in a variety of settings including prediction tasks, data exploration, and hypothesis generation. In medicine, machine learning tools have been used to predict disease presence[12], forecast patient outcomes[13;14], and characterize disease progression[15]. Commonly collected high-dimensional medical signals, such as echocardiograms and electrocardiograms potentially carry rich information unknown and unexploited by physicians—an opportunity for pattern recognition algorithms to be used for prediction and biomedical study.

Automatic diagnosis is a current focus of machine learning applied to EKGs. For example, recent work used deep convolutional residual networks to mimic (and in some cases exceed) the ability of cardiologists to label certain classes of arrhythmia present in a single lead EKG tracing[7]. In contrast, the focus of this work is to explore the use of EKGs to assess risk of future cardiac disease. We compare models using only raw EKG waveform data, simple EHR-derived features, and the conjunction of EKG waveform data and EHR-derived features to characterize patient risk with respect to certain adverse cardiac outcomes.

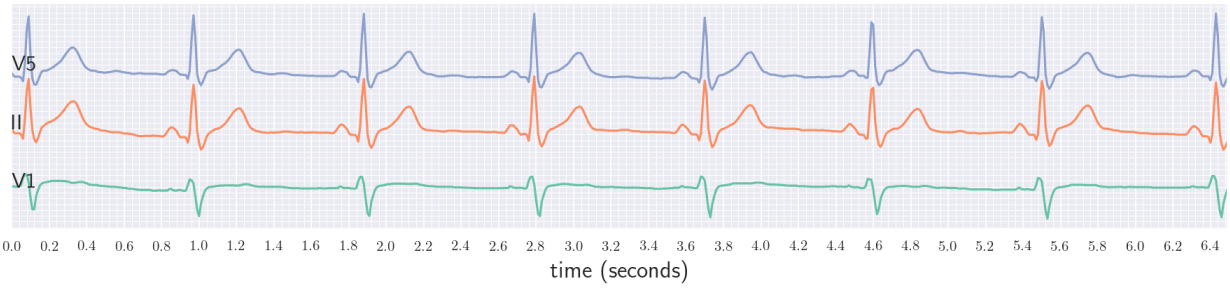
3 Data

Our data are derived from the electronic health records in the Partners health care system, associated with Brigham and Women's hospital (BWH). We focus on emergency department (ED) encounter level data; we assess risk of cardiac outcomes for a patient at the time of their ED visit. Our starting cohort includes every ED visit to BWH between 2010 and 2015 with an associated electrocardiogram recording.

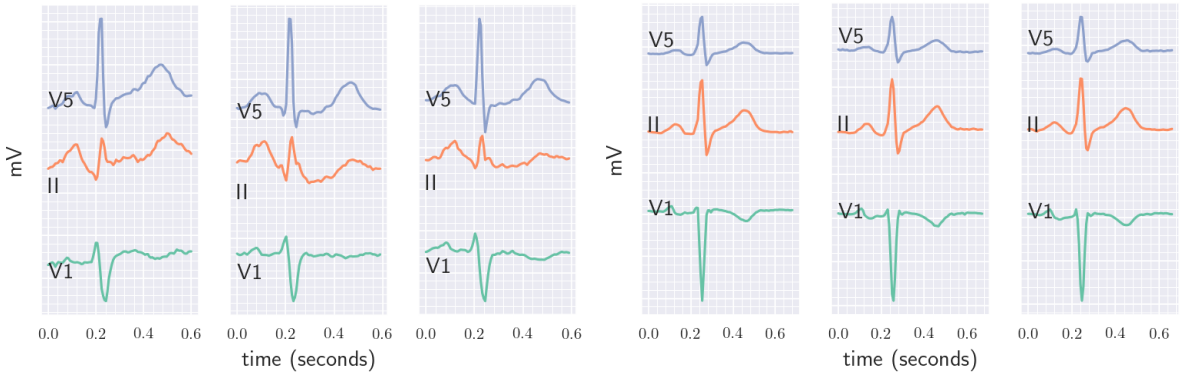
We develop and validate our models on patients with EKGs that exhibit normal sinus rhythm, as interpreted in the associated cardiology report. We exclude patients with EKGs that exhibit atrial fibrillation or atrial flutter. We also exclude EKGs that were noted by the cardiologist interpreter to be noisy or of poor measurement quality. Some ED encounters have multiple electrocardiogram recordings — when this is the case we use the first EKG taken during the encounter to develop and validate our predictive models.

3.1 Outcomes

Our models predict multiple outcomes: future diagnosis of atrial fibrillation, stroke within a six month window, a major adverse cardiac event within a six month window, and troponin lab values at multiple time horizons (we report values for within 30 days). These outcomes are described in more detail below.



(a) Example 10-second EKG, consisting of leads V1, II, and V5.



(b) Patient a

(c) Patient b

Figure 1. Top: full EKG data from the three long leads, V1, V5, and II. Bottom: segmented beats from two patients, using the method from [16]. Each initial EKG has been detrended by subtracting out low frequency signals using a Gaussian filter, and scaled to be within the range [-1, 1]. Each beat has been resampled to a grid of length 100.

Future atrial fibrillation Atrial fibrillation is a common cardiac arrhythmia that has been known to co-occur with more serious cardiac outcomes, including stroke[17] and sudden death[3]. We construct atrial fibrillation-related outcomes from ICD-9 diagnosis records, joining codes 427.31 (atrial fibrillation) and 427.32 (atrial flutter). A patient diagnosed with atrial fibrillation or flutter at least three days after the start date of their ED encounter is associated with outcome $Y^{(afib)} = 1$. Note that our cohort does not include EKGs that exhibit atrial fibrillation or flutter, according to the accompanying cardiology notes.

Stroke-6m ECG abnormalities have been shown to correlate with near-term stroke[18]. We construct a stroke outcome variable within a six month window after the ED encounter from a set of ICD-9 diagnosis codes. We combine the following categories to form the majority of our stroke observations: cerebral artery occlusion (434.91), acute, but ill-defined, cerebrovascular disease (436), intracerebral hemorrhage (431), subarachnoid hemorrhage (430), subdural hemorrhage (432.1), cerebral embolism with cerebral infarction (434.11), unspecified intracranial hemorrhage (432.9), cerebral thrombosis with cerebral infarction (434.01), and others that include "cerebral infarction" in their description (e.g. 433.11, 434.90, 434.10, 434.00, 433.01, 433.31, and 433.21). A patient with any of these diagnoses within sixth months of the ED visit has the outcome $Y^{(stroke)} = 1$.

MACE-6m We also predict the occurrence of a major adverse cardiac event (MACE) within six months following the ED encounter. These events include myocardial infarction, cardiac arrest, receiving a stent, or a coronary artery bypass grafting. A patient with any of these outcomes within sixth months has value $Y^{(mace)} = 1$. Importantly, we report predictive accuracy on a set of patients who were not tested for cardiovascular disease in their emergency department encounter — i.e. did not receive a stress test or catheterization.

Troponin labs Troponin lab tests measure the concentration of a family of proteins (troponin I and T) that are found in skeletal and cardiac muscle fibers. In healthy patients, troponin concentration is small. However, damage to heart

muscle cells causes troponin to be released into the blood, increasing concentration. Troponin lab tests are used to determine if a patient is currently or about to suffer a heart attack. We use high troponin lab observations as an additional surrogate for an adverse cardiac outcome. Based on the guidelines in [19], we define high to be a value of .04 ng/mL or greater. We construct the outcome of troponin laboratory tests at multiple time horizons — 3, 7, 30, and 180 days after the emergency department visit.[1] For each patient, we take the maximum measured troponin value in the window starting at the ED date and ending at each of the time horizons. We assign the outcome variable a 1 if the maximum value is equal to or greater than .04 ng/mL, and a 0 otherwise. Many values were missing; fewer than half of patients had a troponin lab done within 180 days of the ED visit. We treat these outcome values as missing at random when fitting and validating our models (i.e. they are ignored when fitting and analyzing models).

3.2 Patient-level Predictors

To predict the outcomes detailed in the previous section, we use features derived from patient health records and observed EKG tracings.

Patient demographics and simple history For each patient, we include age (standardized to mean 0, standard deviation 1 using the development set distribution), binary indicators for race (`race_black`, `race_white`, `race_hispanic`, `race_other`), and sex. Further, we include a binary indicator if a patient has any past diagnosis of diabetes, myocardial infarction, hypertension, or smoking, based on ICD-9 records (dating back to 1990).

We also include information about historical longitudinal vital signs, including pulse, temperature, height, weight, bmi, diastolic blood pressure, systolic blood pressure, hdl, ldl, and vldl cholesterol. Following [20] we include indicators if any of these values were ever observed

- to be non-zero (i.e. present in history)
- to be in the lowest decile
- to be in the highest decile
- within a normal range (between lowest and highest deciles)
- to be increasing (latest measurement is higher than first measurement)
- to be decreasing
- to be fluctuating (the standard deviation of the measurement is above the 75 percentile)

The demographic, diagnosis history, and vital sign statistics total to 88 features, and serve as our baseline predictors for each outcome. We refer to this set of predictors as the HISTORY features.

Electrocardiogram Waveform Data and Cardiology Remarks For each patient we consider the raw waveform values from EKG leads II, V1, and V5 (the long leads showing 10-second tracings), depicted in Figure 1a. Each EKG is subsampled to 100 Hz, detrended, and the voltages in each lead are scaled to lie in the range $[-1, 1]$. We use the raw EKG waveform data in two ways: the entire trace within a convolutional residual neural network [21] using the architecture described in [7]; and individually segmented beats within a simple multi-layer perceptron.

We also consider information from the text of the cardiology report associated with each EKG. EKG remarks are loosely structured text that indicate the presence (or absence) of certain physiologically meaningful features as interpreted by a physician. Using a set of regular expressions, we process the cardiology remark text into features that indicate the presence or absence of common EKG attributes. These include bundle branch block, aberrant bundle branch block, ventricular hypertrophy, ST depression, ST elevation, T wave presence, T wave inversion, J point elevation, J point repolarization, and prolonged QT. The cardiology report also includes continuously measured features of the EKG, including the PR interval, QRS duration, QT interval, QTc interval and the P-R-T axes. We use the remarks and interval information together as a set of features (the "remark" features), and compare their predictive ability to models that directly use the raw EKG waveform.

[1] In this write-up, we report predictive results for the 30 day outcome, but general patterns of comparison between models hold for the other time horizons.

Dataset Characteristics	Development	Test	Test (no history)
# EKGs	48777	16413	12715
Patient demographics			
# unique patients	33806	11270	9658
mean age (sd)	54.2 (18.1)	54.5 (18.3)	53.1 (18.5)
# female (%)	28482 (58.4 %)	9589 (58.4 %)	7278 (57.2 %)
Patient with history (%)			
mi	3940 (8.1 %)	1318 (8.0 %)	0 (0.0 %)
diabetes	5009 (10.3 %)	1671 (10.2 %)	0 (0.0 %)
hypertension	8619 (17.7 %)	2883 (17.6 %)	0 (0.0 %)
stroke	1989 (4.1 %)	677 (4.1 %)	0 (0.0 %)
smoking	5585 (11.5 %)	1897 (11.6 %)	321 (2.5 %)
Outcomes: total positive (%)			
future-afib	3759 (7.7 %)	1344 (8.2 %)	451 (3.5 %)
stroke-6m	937 (1.9 %)	308 (1.9 %)	196 (1.5 %)
mace-6m	2134 (4.4 %)	716 (4.4 %)	487 (3.8 %)
troponin: (# labs observed, % positive)			
trop-7d	1910 (17113, 11.2 %)	585 (5796, 10.1 %)	388 (4317, 9.0 %)
trop-30d	2181 (17988, 12.1 %)	666 (6095, 10.9 %)	433 (4504, 9.6 %)
trop-180d	2854 (20074, 14.2 %)	888 (6836, 13.0 %)	551 (5023, 11.0 %)

Table 1. Summary of our data, split by development, test, and no history cohorts.

3.3 Sample Summary

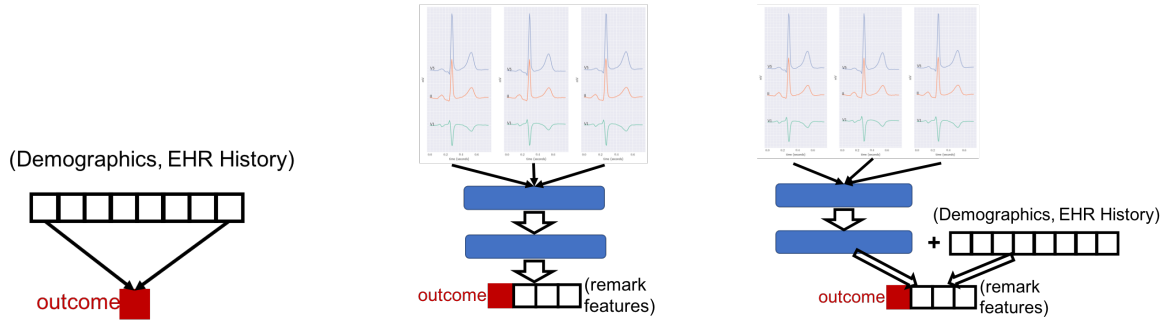
Table 1 summarizes the cohort examined in this study. We construct our cohort such that every encounter has an associated EKG (and all EKGs have associated cardiology remarks); no raw EKG waveforms were treated as missing. Some EKG report entries were missing — for example the QT, QT corrected were sometimes absent from the final cardiology report. We use a simple imputation strategy, substituting in the sample average. Further, every ED encounter has associated patient demographic and current age information. Our HISTORY features are binary indicators — if a diagnosis is present in the EHR records (dating back to 1990 for some patients) the feature is set to one; all others are set to zero. ED encounters with missing outcomes (in the troponin labs) are ignored in the model building and evaluation phases.

Development and Testing Split We split our cohort by patients into a development set (75% of patients) and a testing set (25% of patients) — no patients overlap in these two groups. We report predictive performance on the test data set in this writeup. Additionally, we consider a "no history" subset of patients within the test group who have no recorded history of myocardial infarction, diabetes, stroke, hypertension, or atrial fibrillation — a cohort where predictive accuracy cannot be based on past occurrences. We report the predictive performance of EKG-based and EHR-based algorithms on these patients with no previous diagnoses in these categories. When reporting future afib predictions, we exclude test patients with any history of atrial fibrillation or flutter. Furthermore, for the MACE outcome, we report accuracy only on patients who were left untested in their emergency department visit (for over 10 days).

4 Methods

We compare an array of predictive models that use raw EKG waveform data, electronic health record historical data, or a combination of the two. For each set of features below, we build a model to predict a binary outcome.

History Baseline (HISTORY) We construct a predictive model following the approach in [4] that includes race, height, weight, systolic and diastolic blood pressure, ldl, hdl, and vldl cholesterol, diabetes, antihypertensive medication, smoking history, and past diagnosis of heart disease or cardiac arrest. We construct statistical summaries of height and weight (including body mass index), systolic and diastolic blood pressure measurements, and historical cholesterol levels, following the "patient demographic and simple history" description above. The HISTORY model is a L2-regularized



(a) HISTORY (or REMARK) model **(b)** BEATNET **(c)** BEATNET+HISTORY

Figure 2. Compared models. Not depicted is the residual network based on [7].

logistic regression model, tuned using a fixed validation set (15% of the total dataset), depicted in Figure 2a.

EKG: Cardiology Remark Features (REMARK) We also include a model based on the cardiology report into our comparison. These predictors include the following automatically generated features: PR interval, QRS duration, QT interval, QTc interval and the P-R-T axes. We also include indicator features based on cardiologist derived remarks, that indicate the presence/absence of bundle branch block, aberrant bundle branch block, ventricular hypertrophy, ST depression, ST elevation, T wave presence, T wave inversion, J point elevation, J point repolarization, and prolonged QT. These features were derived via regular expression matching on the cardiology report text.

Electrocardiogram: Residual Network (RESNET) We directly incorporate information from the raw EKG waveform using a convolutional residual neural network[21] modeled on the architecture described in [7] for arrhythmia detection. We augment their single-lead model to the multi-lead setting. In our experiments, we fix the number of repeated residual blocks to 8 (as opposed to the 15 used in [7]); we observe no difference in performance.

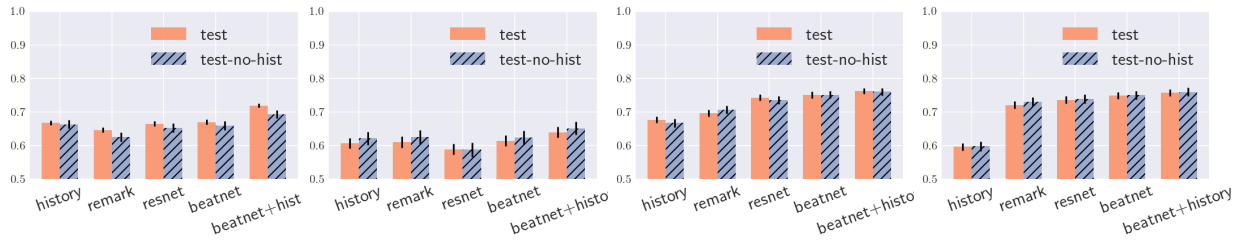
Electrocardiogram: Beat-Segmented Model (BEATNET) We also build models based on a beat-by-beat segmentation of the 10 second EKG signal. We construct these EKG-derived features by first detecting the R peak in the full EKG recording using the method detailed in [16]. We then extract each beat, leaving 2/3 of the cycle after the R peak, and 1/3 of the cycle before. We use the same beat extraction location for each of the three leads (II, V1, and V5). Each extracted beat is resampled to a fixed grid of length 100 (using linear interpolation), creating a beat observation of size 3×100 samples. We predict each outcome using a simple multi-layer perceptron with two hidden layers each of size 500 activation units, with a rectified linear unit as the non-linearity applied to each activation. We train with dropout applied to each layer of hidden activations with probability $p = .5$. We aggregate beats back into a single example by taking the average value across the beats from the same EKG. We leave more complex temporal modeling of within EKG beat-to-beat variability for future work. The BEATNET model is depicted in Figure 2b.

Hybrid Models (BEATNET+HISTORY) We also compare features sets that include combinations of baseline, remark, and raw waveform features. We report results on the beatnet model plus patient history predictors. We use a wide-and-deep model architecture[22] that concatenates the last layer's representation to the patient history features before outputting the outcome predicted probability value. This architecture is depicted in Figure 2c.

For the RESNET, BEATNET, and BEATNET+HISTORY models, we append cardiology remark features to the outcome variable, as we find that this helps regularize our predictor. Note that we do not need cardiology remark features to make predictions, they are only used within training. This multi-task outcome is depicted in Figure 2b and 2c.

Model Fitting We train all models by minimizing a regularized cross entropy loss over model parameters θ and regularization parameter λ

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \ell(Y_n, \hat{Y}_n) + \lambda \cdot R(\theta)\,, \tag{1}$$



(a) future afib **(b)** stroke-6m **(c)** mace **(d)** troponin-30d

Figure 3. Predictive AUCs, all held out patients vs. patients with no history of relevant disease.

where

$$\hat{Y}_n = 1/(1+\exp(-f_\theta(X_n))) \ , \quad \ell(Y_n, \hat{Y}_n) = Y_n \cdot \log \hat{Y}_n + (1-Y_n) \cdot \log(1-\hat{Y}_n) \, . \tag{2}$$

Here $n = 1 \ldots, N$ indexes ED encounters, and $R(\theta) = ||\theta||_2^2$ is L2 regularization. We tune λ using a dedicated validation set made up of 20% of the development dataset (9,635 of the 48,777 ED encounters). For the REMARKand HISTORYmodels X_n are features and $f_\theta(X_n) = \theta_0 + \theta^\intercal X_n$ is a simple logistic regression model, and we explore a grid of λ values, keeping the one with the lowest validation loss.

For the neural net models, X_n is a multi-channel EKG (or beat segment), and $f_\theta(X_n)$ is a composition of convolutional, non-linear, and fully connected maps from the high dimensional X_n space to the out come prediction logit-probability space. In these models, we fix λ but tune performance by early stopping, keeping the model parameters θ corresponding to the best validation loss. For all neural-network based models (e.g. RESNET, BEATNET, and the hybrid models) we train using the `adam` optimizer[23] with step size .001 (reduced by a factor of two after every 15 epochs), and minibatch size 256. We train for a total of 80 epochs and return the model with the best validation loss.[2]

5 Results

For each model, we report performance on the held out set of patients (and various subpopulations of those patients) measured by the area under the ROC curve (AUC). For each model type, we apply the same predictive model to each test set subpopulation. Figure 3 depicts the AUC (and $\pm$ one standard error) for the HISTORY, REMARK, RESNET, BEATNET, and BEATNET+HISTORY models on the full set of test patients and the "no-history" set.

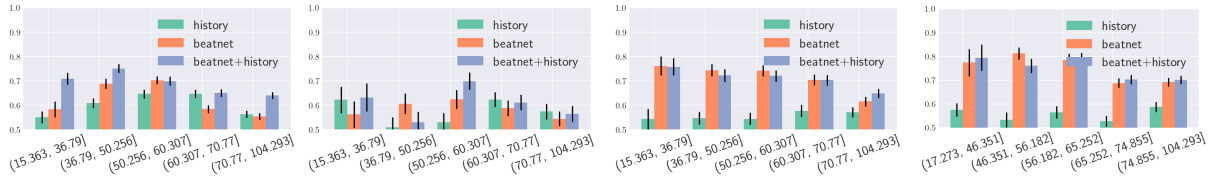
For prediction of atrial fibrillation and stroke within six months, the BEATNET and RESNET perform similarly to the historical features, with future afib AUCs in the range .65 to .68, and future stroke in the range .58 to .62. For both of these outcomes, including the HISTORY features alongside the raw EKG beat waveform improved prediction. For of MACE and troponin within 30 days, the EKG-based predictors substantially outperformed the HISTORY-based predictors.

Furthermore, all predictors performed similarly on the subsample of patients with no prior history of heart disease. For the HISTORY-based predictor, patient demographics and longitudinal vitals entered more heavily in the prediction. This robustness could be partially explained by our choice of cohort — only patients with EKGs that exhibit normal sinus rhythm were included.

We also found that our BEATNET model performed as well or better than the RESNET model[7] trained on the entire EKG. We note that the BEATNET predictor is quite a bit simpler — the beat segmented MLP has 408,014 parameters, whereas the RESNET has 1,215,182 parameters ($\approx 3\times$). One could imagine a further reduction using a convolutional beat specific neural network or a sequence-based beat model — developments we leave for future work. Further, the restricted structure of the BEATNET model suggests that the information in the EKG that is predictive of future outcomes is within the cardiac cycle itself, and not in rhythmic variation from cycle to cycle (as detectable by the RESNET model).

The following subsections discuss predictive performance on a variety of patient subpopulations.

[2]Model implementations are available at `https://github.com/andymiller/ekg-risk-models`.



(a) future afib **(b)** stroke-6m **(c)** mace **(d)** troponin-30d

Figure 4. Predictive AUCs, stratified by age.

EKG-based models are robust to age As expected, age is an influential variable for the patient HISTORY-based risk scores. When we look within a single age group, we see a sharp reduction in the predictive accuracy of the HISTORY model, as depicted in all panels of Figure 4. This reduction, however, does not happen as dramatically in the EKG-based models. For instance, the HISTORY model predicts MACE outcomes with out-of-sample AUC of about .675, while the BEATNET predicts with AUC of .750. When we stratify by age the HISTORY-based predictive AUCs all drop to the range of .54-.57 (Figure 4c). However, the BEATNET scores stay high in most quintiles — for the youngest four quintiles we observe AUCs within .7 and .77. In the oldest quantile, we observe a drop off to .6.[3]

When the BEATNET predictor performs similarly to the HISTORY model, we still see a robustness to within-age-bucket comparisons. Figure 4a shows us that BEATNET outperforms the HISTORY model within the lowest three quintiles. We also note that the combination of EKG beat waveforms and patient history features creates a predictor that performs the best, comparatively, within each age quintile.

Race and Sex In three of our four outcomes, we find that EKG-based predictors exhibit parity among different racial and sex groups. Focusing on MACE, we find that the HISTORY model has a significantly different predictive accuracy between black and white patients, where the black subpopulation exhibits more accurate predictions (.72 vs. .65), seen in Figure 5. However, this gap is significantly reduced when considering only the EKG-based risk scores (see Figure 5c). Notably, the gap reappears in the BEATNET+HISTORY model that combines EKG and HISTORY features. For troponin values within 30 days, the HISTORY model performs significantly worse for the black subpopulation (AUC .54 vs .6.). Again, using EKG-based features erases this discrepancy (Figure 5d). When predicting high troponin values within 30 days, we notice a significant difference between men and women using the HISTORY-based risk score (Figure 6d. This difference disappears when considering EKG features (BEATNET and RESNET). It also disappears when considering the union of EKG and HISTORY features.

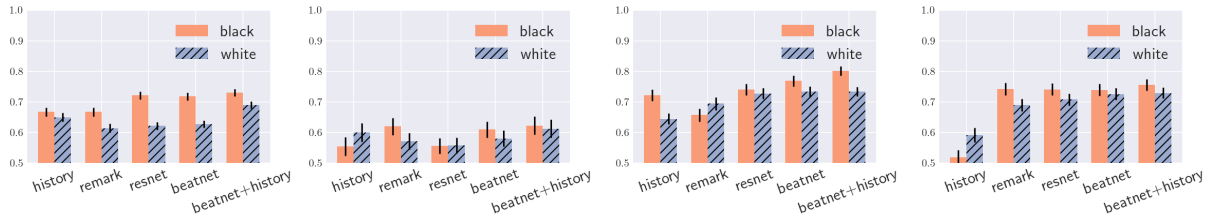
Curiously, the opposite pattern holds for atrial fibrillation prediction. We notice that HISTORY-based risk scores perform equally well between black patients and white patients (with AUC .67-.68). However, the EKG-based scores exhibit a large gap — the black cohort has AUC .71, whereas the white cohort has AUC .62. Overall, we find that our patient HISTORY model performs similarly to the predictive model in [4].

Neural Network Models Detect New Features We find that the raw waveform data adds more predictive information above the cardiology report features for the future afib and MACE outcomes. For MACE prediction, we find that the BEATNET significantly improved over remark features by (.75 to .70 AUROC), suggesting that there are subtle-yet-predictive morphological features present in the raw EKG beat data that are either not noticed or not looked for in the cardiology interpretation. Further, for the future atrial fibrillation outcome, we find that adding raw EKG information to the baseline history features improves predictive performance among the "no history" test cohort.

6 Discussion and Conclusion

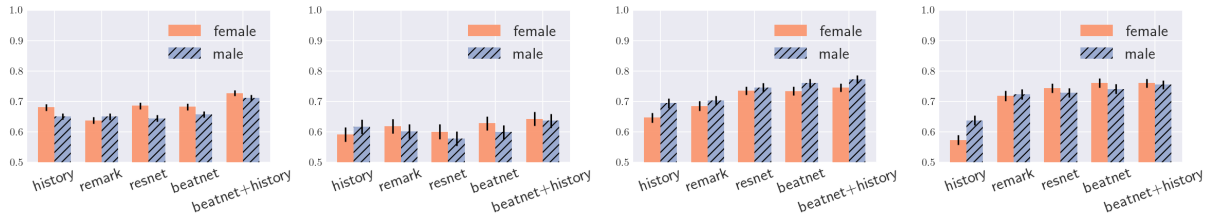
We show that EKG-based risk scores for cardiovascular disease can be competitive with and sometimes outperform patient history-based risk scores. Furthermore, we show that EKG-based risk scores can be far more robust to patient heterogeneity, as it measures the current physiological state of the patient and not a set of historical patient attributes. These results imply that patient history-based risk scores can be less portable across different patient populations,

[3] Interestingly, the EKG-only models tended to predict more accurately in the younger populations, and struggle in older populations. The effect is particularly pronounced in the MACE and troponin predictions, seen in Figure 4.



(a) future afib **(b)** stroke-6m **(c)** mace **(d)** troponin-30d

Figure 5. Predictive AUCs for each model, stratified by race



(a) future afib **(b)** stroke-6m **(c)** mace **(d)** troponin-30d

Figure 6. Predictive AUCs for each model, stratified by sex.

whereas EKG-based risk scores (or their combination) can be more broadly applicable.

One shortcoming of this work is that our validation and development sets were derived from the same underlying patient population. We investigate robustness to patient heterogeneity, but ideally we would draw validation examples from a completely different hospital system. Furthermore, some historical features (e.g. smoking) are not reliably documented in health records.

Another shortcoming of the neural network approach is that models are black box, making it difficult to interpret which features of the EKG are relied upon when making a prediction. Recent methods for interpreting complex discriminative models could alleviate this problem[24]. However, ideally one would learn a decomposition of the EKG signal that enables a cardiologist to interpret the predictive features — for example a separation of rhythmic and characteristic wave variation (e.g. P, QRS, T waves) and how they interact to produce a prognostic risk score. A potential avenue toward this approach is to build upon a generative model of the full EKG[25] that separates variation in cardiac rhythm and cycle morphology.

Another area of interest for clinical predictive modeling is studying multi-task predictors[26], compared to the models developed in this work. Does simultaneous prediction help identify predictive features and reduce model complexity? Ideally, we would combine all sources of patient information — history, EKG, and multiple outcomes — into a single model that captures patterns of missingness and overlap in predictive features in a way that is reliable and portable across different clinical settings. The wide-and-deep (and potentially multi-task) framework is a sensible starting point, but a more nuanced model of patient attribute and EKG feature interaction should be developed and validated. This is an ongoing model development and validation challenge for data science in health care.

References

1. Center for Disease Control. Heart Disease Facts; 2017. `https://www.cdc.gov/heartdisease/facts.htm`.
2. World Health Organization. Cardiovascular Disease;. `http://www.who.int/cardiovascular_diseases/en/`.
3. Eisen A, Ruff CT, Braunwald E, Nordio F, Corbalán R, Dalby A, et al. Sudden cardiac death in patients with atrial fibrillation: insights from the ENGAGE AF-TIMI 48 Trial. Journal of the American Heart Association. 2016;5(7):e003735.
4. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. Journal of

the American Heart Association. 2013;2(2):e000102.
5. Alonso A, Norby FL. Predicting atrial fibrillation and its complications. Circulation Journal. 2016;80(5):1061–1066.
6. Conroy R, Pyörälä K, Fitzgerald Ae, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European heart journal. 2003;24(11):987–1003.
7. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:170701836. 2017;.
8. Dewland TA, Vittinghoff E, Mandyam MC, Heckbert SR, Siscovick DS, Stein PK, et al. Atrial ectopy as a predictor of incident atrial fibrillation: a cohort study. Annals of internal medicine. 2013;159(11):721–728.
9. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837–1847.
10. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. PLoS One. 2012;7(3):e34287.
11. Mullainathan S, Obermeyer Z. Does machine learning automate moral hazard and error? American Economic Review. 2017;107(5):476–80.
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama. 2016;316(22):2402–2410.
13. Oermann EK, Rubinsteyn A, Ding D, Mascitelli J, Starke RM, Bederson JB, et al. Using a machine learning approach to predict outcomes after radiosurgery for cerebral arteriovenous malformations. Scientific reports. 2016;6:21161.
14. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. Prediction policy problems. American Economic Review. 2015;105(5):491–95.
15. Futoma J, Sendak M, Cameron B, Heller K. Predicting disease progression with a model for multivariate longitudinal clinical data. In: Machine Learning for Healthcare Conference; 2016. p. 42–54.
16. Christov II. Real time electrocardiogram QRS detection using combined adaptive threshold. Biomedical engineering online. 2004;3(1):28.
17. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke. 1991;22(8):983–988.
18. Christensen H, Christensen AF, Boysen G. Abnormalities on ECG and telemetry predict stroke outcome at 3 months. Journal of the neurological sciences. 2005;234(1-2):99–103.
19. Newby LK, Jesse RL, Babb JD, Christenson RH, De Fer TM, Diamond GA, et al. ACCF 2012 expert consensus document on practical clinical considerations in the interpretation of troponin elevations: a report of the American College of Cardiology Foundation task force on Clinical Expert Consensus Documents. Journal of the American College of Cardiology. 2012;60(23):2427–2463.
20. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data. 2015;3(4):277–287.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
22. Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, et al. Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM; 2016. p. 7–10.
23. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.
24. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 1135–1144.
25. Sayadi O, Shamsollahi MB, Clifford GD. Synthetic ECG generation and Bayesian filtering using a Gaussian wave-based dynamical model. Physiological measurement. 2010;31(10):1309.
26. Harutyunyan H, Khachatrian H, Kale DC, Galstyan A. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:170307771. 2017;.