Development of a Prediction Model for Incident Atrial Fibrillation using Machine Learning Applied to Harmonized Electronic Health Record Data

Premanand Tiwari, MS; Katie Colborn, PhD; Derek E. Smith, PhD; Fuyong Xing, PhD; Debashis Ghosh, PhD; Michael A. Rosenberg, MD

Abstract

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, whose early detection could lead to significant improvements in outcomes through appropriate prescription of anticoagulation. Although a variety of methods exist for screening for AF, there is general agreement that a targeted approach would be preferred. Implicit within this approach is the need for an efficient method for identification of patients at risk. In this investigation, we examined the strengths and weaknesses of an approach based on application of machine-learning algorithms to electronic health record (EHR) data that has been harmonized to the Observational Medical Outcomes Partnership (OMOP) common data model. We examined data from a total of 2.3M individuals, of whom 1.16% developed incident AF over designated 6-month time intervals. We examined and compared several approaches for data reduction, sample balancing (re-sampling) and predictive modeling using cross-validation for hyperparameter selection, and out-of-sample testing for validation. Although no approach provided outstanding classification accuracy, we found that the optimal approach for prediction of 6-month incident AF used a random forest classifier, raw features (no data reduction), and synthetic minority oversampling technique (SMOTE) resampling (F_1 statistic 0.12, AUC 0.65). This model performed better than a predictive model based only on known AF risk factors, and highlighted the importance of using resampling methods to optimize ML approaches to imbalanced data as exists in EHRs. Further studies using EHR data in other medical systems are needed to validate the clinical applicability of these findings.

Introduction

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, and its prevalence is increasing[1]; ~5.1M individuals had AF in 2010, and an increase of 9.3-12.1M is anticipated by 2030[2]. Importantly, the increased risk of mortality with AF is almost entirely due to an increased risk of thromboembolic stroke[3, 4]. This risk could be reduced if a moderate or high-risk patient with AF is started on oral anticoagulation[5-17]. A major challenge in the management of patients with AF is that often stroke is the first presentation of AF[18], indicating that simply waiting for a patient to develop AF may not be the optimal approach to limit the risk of stroke. On the other hand, population-wide screening for AF is not currently recommended[19-21], although some suggest that targeted screening may be useful[21]. A model that could predict risk of AF over a 6-month period could be applied to target screening to identify a patient with AF prior to the next clinic visit.

The promise of electronic health record (EHR) data has included the potential to leverage 'big data' analytical approaches to predict clinical outcomes within a real-world context. However, despite widespread adoption of EHRs as mandated under the Affordable Care Act,[22] there are limited examples of practical application of EHR data to predict a meaningful clinical outcome[23-27]. In addition to technical limitations of working with data at the scale of the EHR, there are also challenges in performance of external validation across healthcare systems[28-30]. Nonetheless, with increasing availability of cloud-computing[25] platforms and data storage[31, 32], as well as scalable computational models that can be developed and potentially shared across healthcare systems, opportunities to apply EHR data to clinical decision making are emerging.

A great deal of enthusiasm has accompanied applications of deep learning[33] and artificial intelligence to outperform humans in image recognition[34, 35], text recognition[36, 37], and games[38], such as checkers[39] or Go[40]. However, within the healthcare setting, the 'black box' characteristic of machine learning (ML) has caused hesitancy in application. In certain situations, ML approaches, such as support vector machines[41] or random forests[42], have been found to produce greater predictive performance than standard regression models[43-45]. More recently, there has been increased recognition that deep-learning models[33, 46], composed of multiple hidden layers of a neural network rather than a single layer, are better equipped to handle the large amount of data that exists in EHRs. However, in order to understand how these approaches can be applied to a clinical situation, such as prediction of incident AF, additional study is needed.

In this investigation, we developed and tested an ML model to predict 6-month incidence of AF using EHR data. We conducted a systematic examination of EHR data sampled from 2.3 million individuals, in whom we have harmonized 26,000 features, including diagnostic codes and

medications under the OMOP common data model. Among the characteristics we examined in this developmental process includes the appropriate use of data reduction techniques, data resampling to manage dataset imbalance, and identification of a classification algorithm based on training time and accuracy.

Methods

Study Population and Case Ascertainment

The UCHealth hospital system includes 3 large regional centers (North, Central, South) over the front range of Colorado that share a single Epic instance, which allows data from all centers to be pooled into a single data warehouse, a copy of which is located on the Google cloud platform. This warehouse of data was queried using Google BigQuery to create a dataset and conduct analyses directly on the Google cloud platform, where an array of machine-learning tools can be run on virtual machines. To create our study dataset, we applied a classification approach based on predicting risk of incident AF over a 6-month period, as this timeframe is also the standard follow-up time for most cardiovascular providers. We performed a SQL query on the UCHealth EHR for subjects with new diagnosis of AF obtained over a 6-month interval. To identify cases, we filtered out all patients with prevalent AF on first encounter, and then over 6-month intervals (from each encounter), assigned patients to a 'case' classification if they had AF diagnosed by ICD code (ICD-9 427.31 or ICD-10 I48.91) within that interval. Once a patient was designated a case, he/she was removed from the pool, and all non-case patients without AF were designated as 'controls'. Data was available in the EHR for the period from January 1, 2011 until October 1, 2018 for 2.3M subjects. This study protocol was approved for analysis of de-identified data (limited dataset with dates included) by the University of Colorado Institutional Review Board.

Common Data Model and Data splitting

To provide the opportunity for validation of findings in this study, we used a common data model for EHR data, based on the Observational Health Data Sciences and Informatics (OHDSI) collaboration, which uses the Observation Medical Outcomes Partnership common data model (OMOP-CDM)[47]. The OMOP CDM is a mapping of the raw EHR data to a harmonized dataset; for this investigation, we used this CDM with 26k variables (i.e., features) from the EHR, including diagnosis codes and medications. These values are time-stamped with the time of entry into the medical record, which is used to correlate with the timing of the outcome of interest. Features are encoded using one-hot encoding, and were collected cumulatively from the time of first encounter until diagnosis of AF (cases) or end of follow-up (controls). To reduce the time for computation, as well as preserve additional data for future validation studies within our medical system, we obtained a random sample of 412,291 subjects (407,550 controls and 4741 cases), which was then split into training (80%) and testing (20%) sets to compare the models developed in this

investigation. The training set underwent an additional split to create a validation set (10% of the training set) for comparing unsupervised stacked autoencoder models (see below). Figure 1 displays the data management scheme for this investigation.

Model Development

Hyperparameter tuning was performed using iterative random sampling of 10,000 records for manual grid search (neural networks), and 10-fold cross validation for automated grid search (for other machine learning approaches). See *Supplemental Methods* for details. Unsupervised analysis was performed first using principal component analysis to examine overall data structure. For dimensionality reduction, we employed stacked autoencoders[48] using fully connected neural networks of several architectures, with a goal to identify the lowest replication error (cross-entropy loss) within the validation set (10% of training set). We also conducted analyses using the full (non-reduced) feature set of 26k features for comparison.

We examined several strategies for resampling, including random oversampling, SMOTE[49], random undersampling, and cluster centroid. To identify the best resampling approach, we used random forest classifier, as pilot analyses using a smaller dataset suggested this approach might be superior to other ML approaches. We also compared with a model using no resampling (imbalanced).

Once we identified an optimal resampling approach, we compared several classification algorithms, including naïve Bayesian classification, random forest classification, boosted gradient classification, support vector machines, one-layer fully connected neural networks (shallow) and multiple layer fully connected neural networks (deep). Model comparison was based on area-under-curve and F_1 statistic[50-52]. Computation time includes all prior data sampling and algorithm performance. Once an optimal model and resampling approach were identified, we conducted sensitivity analysis using several alternative resampling and modeling approaches in combination to ensure that the combination (dimensionality reduction, resampling, and classification algorithm) identified was indeed optimal. Precision-recall and receiver-operator characteristic curves, as well as feature importance plots, were created for the optimal model for manual inspection.

Validation of Developed Model

The optimal model was then compared with an unregularized logistic regression model based on presence of known clinical predictors of AF, including hypertension (ICD-9 and 10–Hypertension: 401.x and I10, Obesity: 278.x and E66.9, Diabetes: 250.x and E 11.9, Coronary disease: 414.x and I25.1x, Heart failure: 428.x and I50.9, Valvular heart disease: 424.x and I08).

Computation and Analysis

All analyses were run on Google Cloud Platform, using 96 CPUs and 620 GB of RAM. Scripts were composed in Python (version 3) and were run on Jupyter Notebook with Tensorflow platform on the Google Cloud Platform. Machine learning packages included *scikit-learn* and *keras.* Confidence intervals were calculated using Wald method[53, 54], although almost all were within the rounding error of the estimates due to the large testing sample size (N = 82,458), and are not displayed. See *Supplemental Methods* for additional details.

Results

Across the entire UCHealth population of 2.3M, we identified over 26k patients with 6-month incident AF (Table 1). Although essential hypertension was the most common data element in both groups, patients with 6-month incident AF had more standard cardiovascular diagnoses compared to the larger population who were never diagnosed with AF. From this population, a random sample selected for model development included 407,550 controls and 4,741 cases, which were split into a training set (80%) of 326,040 controls and 3,793 cases and a testing set (20%) of 81,510 controls and 948 cases (Figure 1).

To explore the data structure of the EHR data prior to modeling, we performed principal component analysis (PCA), where we noted a large amount of overlap between components in patients with and without 6-month incident AF within the first two principal components (Figure 2A). We noted that only a small amount of overall variability was explained by the first few components, and that there was not a clear plateau present over the first 500 components that were analyzed (Figure 2B). We then created several architectures of stacked autoencoder (SAE) neural networks, using regularization techniques such as drop-out and several activation functions. We found that a deep neural network [10000, 2000, 500, 2000, 10000] with three encoding and three decoding layers, dropout, and sigmoid activation function, resulted in the lowest reconstruction (validation) error (Table 2).

We then examined the role of undersampling and oversampling methods to identify the optimal approach to manage the imbalance between cases and controls that we identified in this dataset. Using a random forest classification algorithm with three-layer sigmoid/dropout SAE (see above), we found that SMOTE not only provided the shortest training times, but it also resulted in the best classification F_1 score (Table 3), compared with other methods.

Using the SMOTE resampling strategy combined with three-layer SAE (dropout + sigmoid activation), we then examined several classification algorithms to identify a potential 'overall' best model. Several models, specifically support vector machines, did not converge after over 24

hours of processing and were not included. Among the approaches examined, we found that random forest classification was superior to other methods that included regularized regression, boosted gradient descent, shallow, and deep neural networks (Table 4).

To ensure that our modeling combination did not favor a particular combination of dimensionality reduction, resampling, and classification algorithm, we then performed a sensitivity analysis using various combinations of each. We found that a model that used the raw features as input (no dimensionality reduction), SMOTE and random forest classifier (F1 0.12, AUC 0.65) performed better than other combinations, including raw features, SMOTE and 2-layer neural network (F1 0.05, AUC 0.53), and was overall the best predictive model identified from our EHR. This model had a specificity of 94.7%, sensitivity of 35.74%, negative predictive value of 99.3% and positive predictive value of 6.93% at a probability (decision) cutoff of 0.5. The confusion matrix for the model is displayed in Supplemental Figure 1. As shown in Figure 3A, several additional probability cutoffs for classification before 0.4 provided over 95% sensitivity, at the expense of a significant decrease in specificity. Feature importance was examined for the optimal model, which found that none of the features most common across the population (Table 1) were of high importance in classification of 6-month AF (Supplemental Table 1 and Supplemental Figure 2), although several features predominant in incident AF patients had non-zero importance (42872402 Coronary arteriosclerosis in native artery: 1.61e-05; 254761 Cough: 4.11e-06; 201826 Type 2 diabetes mellitus: 1.96e-07). Calibration curves also revealed relatively poor classification across all probability thresholds (Supplemental Figure 3) for several models.

Finally, we found that the optimal model (raw features, SMOTE, random forest classifier) performed better than an unregularized logistic regression model based on known AF risk factors, which had an F_1 score of 0.073, AUC of 0.64 with SMOTE resampling prior to fitting (training), and an F_1 score of 0.00 and AUC of 0.5 without resampling prior to fitting.

Discussion

In this investigation of development of a machine-learning model using harmonized EHR data for predicting 6-month incident AF, we found that a random forest classifier created from raw feature inputs with SMOTE oversampling provided better classification than any other combinations of approaches tested, including deep learning models and known risk factors of AF. These results are significant because in addition to motivating future investigations to apply ML methods to EHR data to identify patients at risk for AF, they also incorporated harmonized data (OMOP-CDM), which means that the optimal model can not only be directly applied in our EHR, but to data from the EHR of any other medical institution participating in OMOP/OHDSI. In clinical application, our model could thus be inserted directly back into the user interface to guide

targeted screening patients at risk of AF, including development of prospective follow-up studies to use the prediction for targeted screening for AF, including routine ECGs, implantable, or wearable devices.

However, there are several reasons for hesitancy before taking these results directly back to the bedside to guide clinical management, without additional investigation. First, the model that we identified was not extremely accurate, with an F_1 score well under 20%, and a sensitivity of 35.74% based on a cutoff probability of 0.5 for risk. Although we noted that the probability threshold can be lowered to improve sensitivity of classification, the drop in specificity, and number of false-positives, with such an approach would result in a large number of patients undergoing unfruitful screening. Second, we found that the features identified as important in the classification process were atypical, with many of unclear association with AF based on known pathophysiology[55, 56] and clinical risk factors[57, 58]. Although this finding may reflect the fact that none of the features was particularly strong in predicting risk of AF, as evident from the low feature-importance scores, it also suggests that the model may be overfitting our population (despite validation in a held-out testing set). Importantly, because we performed a harmonization step prior to the modeling process, the necessary external population validation to explore the possibility of overfitting is simply a matter applying this model directly to OMOP-CDM from an outside EHR. It is also important to keep our findings in context as there is a great deal unknown about clinical risk in development of AF, and one need look no further than to the field of genetic investigation, where whole-genome approaches have identified novel risk loci whose role in AF pathophysiology remains poorly understood[59, 60], but which provide far superior prediction over many candidate genes.

In the process of developing of a 6-month risk prediction model for AF, we made several important observations about the application of machine learning to EHR data. First, we found that dimensionality reduction was inferior to use of raw feature inputs for predicting incident AF. This finding likely reflects the sparsity of EHR data, which resulted in a significant loss of information even using dimensionality reduction methods with very low reconstruction loss. This finding implies that within our population, the strongest predictors of 6-month incident AF are relatively rare, and thus minimized with dimensionality reduction approaches. The relatively obscure features identified in the feature importance evaluation supports this contention. To our knowledge, this phenomenon has not been described in EHR-analysis approaches, many of which apply some form of dimensionality reduction[61], although further exploration is likely necessary.

Second, we found that for a rare condition like 6-month incident AF (1.2% of the total population), oversampling to rebalance the data was superior to using the imbalanced dataset and undersampling, regardless of the classification algorithm applied. In fact, the majority of the ML approaches tested failed to provide any added improvement (F1 0.0 and AUC 0.5) without sample rebalancing. This finding is likely reflective of the increase in power obtained with oversampling, although further work is needed to understand why this particular approach was superior.

Finally, we found that random forest classification provided the optimal classification algorithm, with better classification than a deep neural network approach when the input was raw features. This finding demonstrates that although deep learning approaches may be superior for classification of structured datasets, such as in image[62] or voice recognition[63], they are not always optimal over other 'standard' ML algorithms, and highlights the importance of examining all approaches for each classification problem, rather than assuming a giving approach is optimal.

Strengths

In addition to the insights above, there are several additional strengths noteworthy in this investigation. First, all models were created using a harmonization scheme (OMOP common data model) that could allow for direct application and validation to data mapped from a separate EHR. Such harmonization allows for the opportunity to explore transfer-learning[64] approaches, which could provide additional insight into similar and divergent AF risk factors across populations. Second, we conducted a systematic approach to identify the best dimensionality reduction, resampling, and classification algorithm for this outcome. Further work in other outcomes is needed to determine if the combination we identified for predicting 6-month incident AF is also optimal for prevalent or longer-term AF prediction, as well as for outcomes that are more or less common than AF. Finally, we examined a dataset of over 2 million subjects, which provided more than enough sample size from our single institution to conduct cross-validation and out-of-sample validation. This power from use of big data is possible by the unique circumstances of our relationship with Google Cloud Platform, although many other EHRs are moving to the cloud, providing further opportunities for development and testing.

Limitations

There were several limitations in this study, many of which are the subject of future, more targeted investigations. For one, our study included a very simple method for the temporal relationships between features in our dataset, which did not account for time-varying effects or censoring. An AF event that occurs the day after an encounter is modeled the same as one occurring the day before a subsequent encounter, and a diagnosis or medication that was given

one month before the AF diagnosis was weighted the same as one given 4 years prior. While we suggest that the approach we employed for this investigation is reasonable based on the typical 6-month follow-up schedule for patients seen in cardiology clinic, we realize that additional information about temporal risk will be needed for more accurate prediction approaches. More sophisticated methods, such as recurrent neural networks[61, 65] or parametric survival functions[66] could provide more accurate prediction in future investigations. A second weakness is that we excluded some additional data elements, such as lab values and diagnostic test results, which may have had prognostic value for predicting 6-month AF[67, 68]. Some of these values have been difficult to harmonize across datasets via OMOP-CDM, and others suffer from high variability in inter-institutional measurement, such as echo measures of diastolic function[57, 69]. Nonetheless, there are many additional biomarkers[67, 68] likely to have a more 'biological' relationship to risk of AF than a diagnostic code, and future applications that include this information would be expected to provide both predictive and inferential knowledge about risk of AF. Finally, although the systematic, harmonized approach we employed in this study holds potential for cross-institutional validation, much work is needed in terms of data sharing before actual testing can be performed. Our group and others are working in this direction, and the hope is that sometime in the not-to-distant future, all EHRs will incorporate a 'standard' risk prediction model for AF and many other conditions.

In conclusion, we studied the development of an ML model to predict 6-month risk of AF using harmonized EHR data and found that the combination of raw feature inputs, SMOTE oversampling, and random forest classification provided superior prediction than other models, including one with known clinical risk factors. Further work is needed to explore the technical and clinical applications of this approach model to improving outcomes.

Acknowledgements

This work was funded by grants from the National Institute of Health/NHLBI (MAR: 5K23 HL127296).

References

1. Majeed A, Moser K and Carroll K. Trends in the prevalence and management of atrial fibrillation in general practice in England and Wales, 1994-1998: analysis of data from the general practice research database. *Heart*. 2001;86:284-8.
2. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C, Isasi CR, Jimenez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Mackey RH, Matsushita K, Mozaffarian D, Mussolino ME, Nasir K, Neumar RW, Palaniappan L, Pandey DK, Thiagarajan RR, Reeves MJ, Ritchey M, Rodriguez CJ, Roth GA, Rosamond WD, Sasson C, Towfighi A, Tsao CW, Turner MB, Virani SS, Voeks JH, Willey JZ, Wilkins JT, Wu JH,

Alger HM, Wong SS and Muntner P. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation*. 2017;135:e146-e603.

3. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB and Levy D. Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation*. 1998;98:946-52.

4. Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ and Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. *JAMA*. 1994;271:840-4.

5. The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation. The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators. *N Engl J Med*. 1990;323:1505-11.

6. Stroke Prevention in Atrial Fibrillation Study. Final results. *Circulation*. 1991;84:527-39.

7. Warfarin versus aspirin for prevention of thromboembolism in atrial fibrillation: Stroke Prevention in Atrial Fibrillation II Study. *Lancet*. 1994;343:687-91.

8. Petersen P, Boysen G, Godtfredsen J, Andersen ED and Andersen B. Placebo-controlled, randomised trial of warfarin and aspirin for prevention of thromboembolic complications in chronic atrial fibrillation. The Copenhagen AFASAK study. *Lancet*. 1989;1:175-9.

9. Ezekowitz MD, Bridgers SL, James KE, Carliner NH, Colling CL, Gornick CC, Krause-Steinrauf H, Kurtzke JF, Nazarian SM, Radford MJ and et al. Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. Veterans Affairs Stroke Prevention in Nonrheumatic Atrial Fibrillation Investigators. *N Engl J Med*. 1992;327:1406-12.

10. Connolly SJ, Laupacis A, Gent M, Roberts RS, Cairns JA and Joyner C. Canadian Atrial Fibrillation Anticoagulation (CAFA) Study. *J Am Coll Cardiol*. 1991;18:349-55.

11. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. *Arch Intern Med*. 1994;154:1449-57.

12. Hart RG, Pearce LA and Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann Intern Med*. 2007;146:857-67.

13. van Walraven C, Hart RG, Singer DE, Laupacis A, Connolly S, Petersen P, Koudstaal PJ, Chang Y and Hellemons B. Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. *JAMA*. 2002;288:2441-8.

14. Cooper NJ, Sutton AJ, Lu G and Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Arch Intern Med*. 2006;166:1269-75.

15. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD and Wallentin L. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2009;361:1139-51.

16. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, Becker RC, Nessel CC, Paolini JF, Berkowitz SD, Fox KA and Califf RM. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med*. 2011;365:883-91.
17. Connolly SJ, Eikelboom J, Joyner C, Diener HC, Hart R, Golitsyn S, Flaker G, Avezum A, Hohnloser SH, Diaz R, Talajic M, Zhu J, Pais P, Budaj A, Parkhomenko A, Jansky P, Commerford P, Tan RS, Sim KH, Lewis BS, Van Mieghem W, Lip GY, Kim JH, Lanas-Zanetti F, Gonzalez-Hermosillo A, Dans AL, Munawar M, O'Donnell M, Lawrence J, Lewis G, Afzal R and Yusuf S. Apixaban in patients with atrial fibrillation. *N Engl J Med*. 2011;364:806-17.
18. Jaakkola J, Mustonen P, Kiviniemi T, Hartikainen JE, Palomäki A, Hartikainen P, Nuotio I, Ylitalo A and Airaksinen KE. Stroke as the First Manifestation of Atrial Fibrillation. *PLoS One*. 2016;11:e0168010.
19. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, Doubeni CA, Epling JW, Jr., Kemper AR, Kubik M, Landefeld CS, Mangione CM, Silverstein M, Simon MA, Tseng CW and Wong JB. Screening for Atrial Fibrillation With Electrocardiography: US Preventive Services Task Force Recommendation Statement. *Jama*. 2018;320:478-484.
20. Jonas DE, Kahwati LC, Yun JDY, Middleton JC, Coker-Schwimmer M and Asher GN. Screening for Atrial Fibrillation With Electrocardiography: Evidence Report and Systematic Review for the US Preventive Services Task Force. *Jama*. 2018;320:485-498.
21. Freedman B, Camm J, Calkins H, Healey JS, Rosenqvist M, Wang J, Albert CM, Anderson CS, Antoniou S, Benjamin EJ, Boriani G, Brachmann J, Brandes A, Chao TF, Conen D, Engdahl J, Fauchier L, Fitzmaurice DA, Friberg L, Gersh BJ, Gladstone DJ, Glotzer TV, Gwynne K, Hankey GJ, Harbison J, Hillis GS, Hills MT, Kamel H, Kirchhof P, Kowey PR, Krieger D, Lee VWY, Levin LA, Lip GYH, Lobban T, Lowres N, Mairesse GH, Martinez C, Neubeck L, Orchard J, Piccini JP, Poppe K, Potpara TS, Puererfellner H, Rienstra M, Sandhu RK, Schnabel RB, Siu CW, Steinhubl S, Svendsen JH, Svennberg E, Themistoclakis S, Tieleman RG, Turakhia MP, Tveit A, Uittenbogaart SB, Van Gelder IC, Verma A, Wachter R and Yan BP. Screening for Atrial Fibrillation: A Report of the AF-SCREEN International Collaboration. *Circulation*. 2017;135:1851-1867.
22. Kocher R, Emanuel EJ and DeParle NA. The Affordable Care Act and the future of clinical medicine: the opportunities and challenges. *Ann Intern Med*. 2010;153:536-9.
23. Vis C, Mol M, Kleiboer A, Buhrmann L, Finch T, Smit J and Riper H. Improving Implementation of eMental Health for Mood Disorders in Routine Practice: Systematic Review of Barriers and Facilitating Factors. *JMIR mental health*. 2018;5:e20.
24. Dexheimer JW, Talbot TR, Sanders DL, Rosenbloom ST and Aronsky D. Prompting clinicians about preventive care measures: a systematic review of randomized controlled trials. *Journal of the American Medical Informatics Association : JAMIA*. 2008;15:311-20.
25. Sadoughi F and Erfannia L. Health Information System in a Cloud Computing Context. *Studies in health technology and informatics*. 2017;236:290-297.

26. Jensen RE, Rothrock NE, DeWitt EM, Spiegel B, Tucker CA, Crane HM, Forrest CB, Patrick DL, Fredericksen R, Shulman LM, Cella D and Crane PK. The role of technical advances in the adoption and integration of patient-reported outcomes in clinical care. *Medical care*. 2015;53:153-9.
27. Casey JA, Schwartz BS, Stewart WF and Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual review of public health*. 2016;37:61-81.
28. Huang Y, Lee J, Wang S, Sun J, Liu H and Jiang X. Privacy-Preserving Predictive Modeling: Harmonization of Contextual Embeddings From Different Sources. *JMIR medical informatics*. 2018;6:e33.
29. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E and Solbrig HR. A Consensus-Based Approach for Harmonizing the OHDSI Common Data Model with HL7 FHIR. *Studies in health technology and informatics*. 2017;245:887-891.
30. Jiang G, Evans J, Oniki TA, Coyle JF, Bain L, Huff SM, Kush RD and Chute CG. Harmonization of detailed clinical models with clinical study data standards. *Methods of information in medicine*. 2015;54:65-74.
31. Ocana K and de Oliveira D. Parallel computing in genomic research: advances and applications. *Advances and applications in bioinformatics and chemistry : AABC*. 2015;8:23-35.
32. Korb O, Finn PW and Jones G. The cloud and other new computational methods to improve molecular modelling. *Expert opinion on drug discovery*. 2014;9:1121-31.
33. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature*. 2015;521:436-44.
34. Stallkamp J, Schlipsing M, Salmen J and Igel C. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition. *Neural networks : the official journal of the International Neural Network Society*. 2012;32:323-32.
35. Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, den Heeten A and Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*. 2017;35:303-312.
36. Minarro-Gimenez JA, Marin-Alonso O and Samwald M. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*. 2014;205:584-8.
37. Zhu Q, Li X, Conesa A and Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*. 2018;34:1547-1554.
38. Moravcik M, Schmid M, Burch N, Lisy V, Morrill D, Bard N, Davis T, Waugh K, Johanson M and Bowling M. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*. 2017;356:508-513.
39. Schaeffer J, Burch N, Bjornsson Y, Kishimoto A, Muller M, Lake R, Lu P and Sutphen S. Checkers is solved. *Science*. 2007;317:1518-22.
40. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T and Hassabis D. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484-9.

41. Ramirez J, Monasterio V, Minchole A, Llamedo M, Lenis G, Cygankiewicz I, Bayes de Luna A, Malik M, Martinez JP, Laguna P and Pueyo E. Automatic SVM classification of sudden cardiac death and pump failure death from autonomic and repolarization ECG markers. *J Electrocardiol*. 2015;48:551-7.
42. Beaulieu-Jones BK and Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of biomedical informatics*. 2016;64:168-178.
43. Luo Y, Li Z, Guo H, Cao H, Song C, Guo X and Zhang Y. Predicting congenital heart defects: A comparison of three data mining methods. *PLoS One*. 2017;12:e0177811.
44. Amini P, Ahmadinia H, Poorolajal J and Moqaddasi Amiri M. Evaluating the High Risk Groups for Suicide: A Comparison of Logistic Regression, Support Vector Machine, Decision Tree and Artificial Neural Network. *Iranian journal of public health*. 2016;45:1179-1187.
45. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW and Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical care medicine*. 2016;44:368-74.
46. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E and Dudley JT. Artificial Intelligence in Cardiology. *J Am Coll Cardiol*. 2018;71:2668-2679.
47. Makadia R and Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *EGEMS (Washington, DC)*. 2014;2:1110.
48. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313:504-7.
49. Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357.
50. Chai KE, Anthony S, Coiera E and Magrabi F. Using statistical text classification to identify health information technology incidents. *Journal of the American Medical Informatics Association : JAMIA*. 2013;20:980-5.
51. J D and M. G. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 2006:233–40.
52. Hripcsak G and Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA*. 2005;12:296-8.
53. Agresti A and Coull BA. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*. 1998;52:119-126.
54. Lakhani P and Langlotz CP. Automated detection of radiology reports that document non-routine communication of critical or significant results. *Journal of digital imaging*. 2010;23:647-57.
55. Nattel S, Shiroshita-Takeshita A, Brundel BJ and Rivard L. Mechanisms of atrial fibrillation: lessons from animal models. *Prog Cardiovasc Dis*. 2005;48:9-28.

56. Nattel S and Harada M. Atrial remodeling and atrial fibrillation: recent advances and translational perspectives. *J Am Coll Cardiol*. 2014;63:2335-45.
57. Rosenberg MA and Manning WJ. Diastolic dysfunction and risk of atrial fibrillation: a mechanistic appraisal. *Circulation*. 2012;126:2353-62.
58. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, Sinner MF, Sotoodehnia N, Fontes JD, Janssens AC, Kronmal RA, Magnani JW, Witteman JC, Chamberlain AM, Lubitz SA, Schnabel RB, Agarwal SK, McManus DD, Ellinor PT, Larson MG, Burke GL, Launer LJ, Hofman A, Levy D, Gottdiener JS, Kaab S, Couper D, Harris TB, Soliman EZ, Stricker BH, Gudnason V, Heckbert SR and Benjamin EJ. 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:e000102.
59. Roselli C, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM, Almgren P, Alonso A, Anderson CD, Aragam KG, Arking DE, Barnard J, Bartz TM, Benjamin EJ, Bihlmeyer NA, Bis JC, Bloom HL, Boerwinkle E, Bottinger EB, Brody JA, Calkins H, Campbell A, Cappola TP, Carlquist J, Chasman DI, Chen LY, Chen YI, Choi EK, Choi SH, Christophersen IE, Chung MK, Cole JW, Conen D, Cook J, Crijns HJ, Cutler MJ, Damrauer SM, Daniels BR, Darbar D, Delgado G, Denny JC, Dichgans M, Dorr M, Dudink EA, Dudley SC, Esa N, Esko T, Eskola M, Fatkin D, Felix SB, Ford I, Franco OH, Geelhoed B, Grewal RP, Gudnason V, Guo X, Gupta N, Gustafsson S, Gutmann R, Hamsten A, Harris TB, Hayward C, Heckbert SR, Hernesniemi J, Hocking LJ, Hofman A, Horimoto A, Huang J, Huang PL, Huffman J, Ingelsson E, Ipek EG, Ito K, Jimenez-Conde J, Johnson R, Jukema JW, Kaab S, Kahonen M, Kamatani Y, Kane JP, Kastrati A, Kathiresan S, Katschnig-Winter P, Kavousi M, Kessler T, Kietselaer BL, Kirchhof P, Kleber ME, Knight S, Krieger JE, Kubo M, Launer LJ, Laurikka J, Lehtimaki T, Leineweber K, Lemaitre RN, Li M, Lim HE, Lin HJ, Lin H, Lind L, Lindgren CM, Lokki ML, London B, Loos RJF, Low SK, Lu Y, Lyytikainen LP, Macfarlane PW, Magnusson PK, Mahajan A, Malik R, Mansur AJ, Marcus GM, Margolin L, Margulies KB, Marz W, McManus DD, Melander O, Mohanty S, Montgomery JA, Morley MP, Morris AP, Muller-Nurasyid M, Natale A, Nazarian S, Neumann B, Newton-Cheh C, Niemeijer MN, Nikus K, Nilsson P, Noordam R, Oellers H, Olesen MS, Orho-Melander M, Padmanabhan S, Pak HN, Pare G, Pedersen NL, Pera J, Pereira A, Porteous D, Psaty BM, Pulit SL, Pullinger CR, Rader DJ, Refsgaard L, Ribases M, Ridker PM, Rienstra M, Risch L, Roden DM, Rosand J, Rosenberg MA, Rost N, Rotter JI, Saba S, Sandhu RK, Schnabel RB, Schramm K, Schunkert H, Schurman C, Scott SA, Seppala I, Shaffer C, Shah S, Shalaby AA, Shim J, Shoemaker MB, Siland JE, Sinisalo J, Sinner MF, Slowik A, Smith AV, Smith BH, Smith JG, Smith JD, Smith NL, Soliman EZ, Sotoodehnia N, Stricker BH, Sun A, Sun H, Svendsen JH, Tanaka T, Tanriverdi K, Taylor KD, Teder-Laving M, Teumer A, Theriault S, Trompet S, Tucker NR, Tveit A, Uitterlinden AG, Van Der Harst P, Van Gelder IC, Van Wagoner DR, Verweij N, Vlachopoulou E, Volker U, Wang B, Weeke PE, Weijs B, Weiss R, Weiss S, Wells QS, Wiggins KL, Wong JA, Woo D, Worrall BB, Yang PS, Yao J, Yoneda ZT, Zeller T, Zeng L, Lubitz SA, Lunetta KL and Ellinor PT. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat Genet*. 2018.
60. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, Palsson A,

Blondal T, Sulem P, Backman VM, Hardarson GA, Palsdottir E, Helgason A, Sigurjonsdottir R, Sverrisson JT, Kostulas K, Ng MC, Baum L, So WY, Wong KS, Chan JC, Furie KL, Greenberg SM, Sale M, Kelly P, MacRae CA, Smith EE, Rosand J, Hillert J, Ma RC, Ellinor PT, Thorgeirsson G, Gulcher JR, Kong A, Thorsteinsdottir U and Stefansson K. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*. 2007;448:353-7.

61. Pham T, Tran T, Phung D and Venkatesh S. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*. 2017;69:218-229.

62. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB and Kim N. Deep Learning in Medical Imaging: General Overview. *Korean journal of radiology*. 2017;18:570-584.

63. Agarwalla S and Sarma KK. Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech. *Neural networks : the official journal of the International Neural Network Society*. 2016;78:97-111.

64. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, Liebovitz D, Sun J, Denny J and Malin B. Building bridges across electronic health record systems through inferred phenotypic topics. *Journal of biomedical informatics*. 2015;55:82-93.

65. Yang H, Pan Z and Tao Q. Robust and Adaptive Online Time Series Prediction with Long Short-Term Memory. *Computational intelligence and neuroscience*. 2017;2017:9478952.

66. Zhang Z. Parametric regression model for survival data: Weibull regression model as an example. *Annals of translational medicine*. 2016;4:484.

67. Rosenberg MA, Maziarz M, Tan AY, Glazer NL, Zieman SJ, Kizer JR, Ix JH, Djousse L, Siscovick DS, Heckbert SR and Mukamal KJ. Circulating fibrosis biomarkers and risk of atrial fibrillation: The Cardiovascular Health Study (CHS). *Am Heart J*. 2014;167:723-8 e2.

68. Patton KK, Ellinor PT, Heckbert SR, Christenson RH, DeFilippi C, Gottdiener JS and Kronmal RA. N-terminal pro-B-type natriuretic peptide is a major predictor of the development of atrial fibrillation: the Cardiovascular Health Study. *Circulation*. 2009;120:1768-74.

69. Rosenberg MA, Gottdiener JS, Heckbert SR and Mukamal KJ. Echocardiographic diastolic parameters and risk of atrial fibrillation: the Cardiovascular Health Study. *Eur Heart J*. 2012;33:904-12.

	No AF	6-month Incident AF
Number (%)	2.3M (98.85%)	26K (1.15%)
Age (Mean ± SD)	42.48 ± 22.28	71.14 ± 16.28
Female sex (%)	54.79%	45.2%
#1 code	320128 Essential hypertension 15.6% SNOMED:59621000	320128 Essential hypertension 44.66% SNOMED:59621000
#2 code	254761 Cough 10.03% SNOMED:49727002	432867 Hyperlipidemia 32.42% SNOMED:55822004
#3 code	200219 Abdominal pain 9.89% SNOMED:21522001	77670 Chest pain 17.08% SNOMED:29857009
#4 code	432867 Hyperlipidemia 9.23% SNOMED:55822004	318800 Gastroesophageal reflux disease 14.94% SNOMED:235595009
#5 code	257011 Acute upper respiratory infection 8.87% SNOMED:54398005	312437 Dyspnea 14.91% SNOMED:267036007
#6 code	77670 Chest pain 8.86% SNOMED:29857009	201826 Type 2 diabetes mellitus 14.65% SNOMED:44054006
#7 code	25297 Acute pharyngitis 8.15% SNOMED:363746003	42872402 Coronary arteriosclerosis in native artery 13.38% SNOMED:1641000119107
#8 code	378253 Headache 7.67% SNOMED:25064002	40481919 Coronary atherosclerosis 13.35% SNOMED:443502000
#9 code	442077 Anxiety disorder 7.66% SNOMED:197480006	439926 Malaise and fatigue 12.80% SNOMED:271795006
#10 code	436096 Chronic pain 7.39% SNOMED:82423001	254761 Cough 12.38% SNOMED:49727002

Table 1. Description of UCHealth cohort by AF diagnosis. Provided are the mean age and gender by 6 month AF diagnosis, as well as the top 10 diagnosis codes, according to whether AF was diagnosed within a 6-month period.

Table 2. Comparison of Stacked Autoencoders

Architecture of Stacked Autoencoders	Validation Error
26000-13000-6000-2000-500-2000-6000-13000-26000	0.035
26000-13000-6000-2000-500-2000-6000-13000-26000, dropout	0.049
26000-10000-500-10000-26000,dropout	0.038
26000-10000-2000-500-2000-10000-26000, dropout, sigmoid activation	0.007
26000-500-26000	0.067
26000-10000-500-10000-26000, dropout, sigmoid activation	0.045
26000-10000-2000-100-2000-10000-26000, dropout, sigmoid activation	0.019

Table 3. Comparison of Resampling Strategies. Sampling comparison from Random Forest model; performed using SAE encoded as specified.

		F1 Score	AUC	Training time
Oversampling				
	Random (SAE encoded)	0.033	0.56	65.4 minutes
	SMOTE (SAE encoded)	0.033	0.57	66.7 minutes
Undersampling				
	Random (SAE encoded)	0.036	0.61	4.6 minutes
	Cluster centroid (SAE encoded)	0.033	0.56	65.4 minutes
None		0.00	0.5	10.2 minutes

Table 4. Comparison of machine learning approaches. Using SMOTE resampling technique and all features. F1 and AUC calculated from model applied to held-out testing set (20%); training time is for training of training set (80%)

	F1 Score	AUC	Training time
Naïve Bayes	0.034	0.57	0.1 minutes
Logistic regression with L2 regularization	0.039	0.61	2.1 minutes
RF	0.116	0.65	1489.1 minutes
Shallow NN	0.00	0.5	2.9 minutes
Deep NN	0.050	0.53	252.8 minutes
GBM	0.041	0.64	109.8 minutes

Figure 1. Study sampling design. *Note that for comparison of stacked autoencoders (unsupervised learning), the training set was split into an additional validation set (10% of training set). See *Methods* for details.

Figure 2. A. First two principal components of codes (color labeled by case and control). Red = control, green = case. **B. Variance explained by PC.**

Figure 3. A. Precision-recall curve for optimal model. B. ROC curve for optimal model.

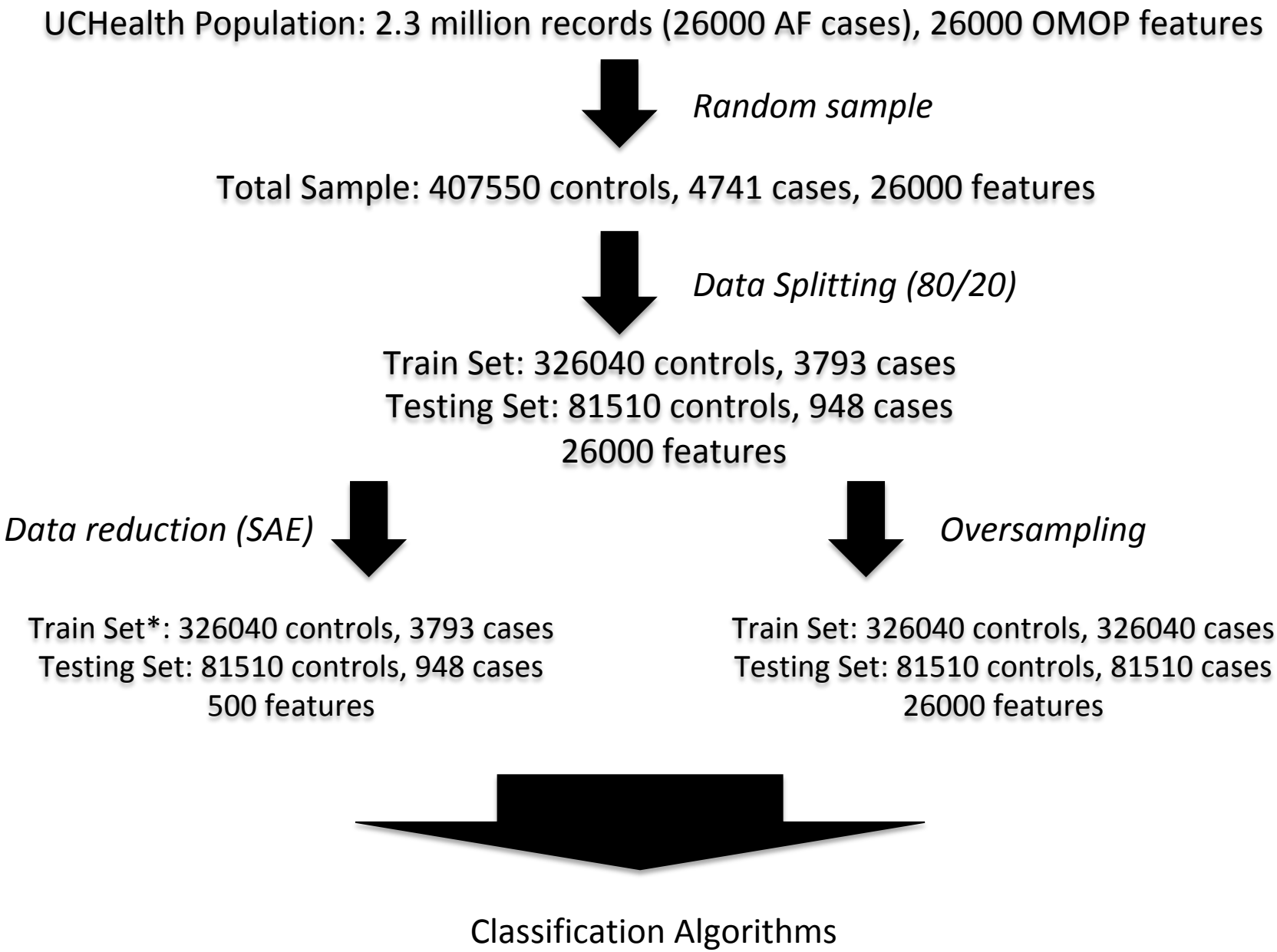
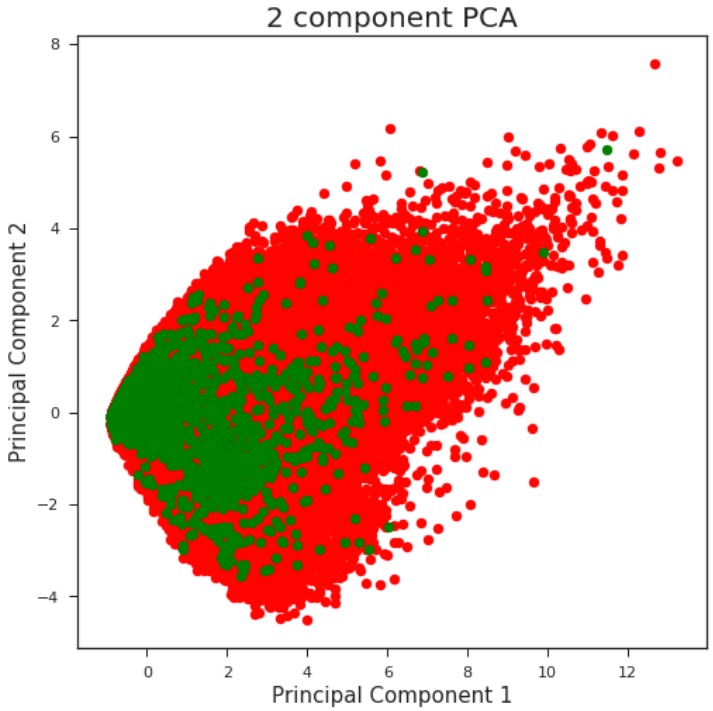

UCHealth Population: 2.3 million records (26000 AF cases), 26000 OMOP features
Random sample
Total Sample: 407550 controls, 4741 cases, 26000 features
Data Splitting (80/20)
Train Set: 326040 controls, 3793 cases
Testing Set: 81510 controls, 948 cases
26000 features
Data reduction (SAE)
Oversampling
Train Set*: 326040 controls, 3793 cases
Testing Set: 81510 controls, 948 cases
500 features
Train Set: 326040 controls, 326040 cases
Testing Set: 81510 controls, 81510 cases
26000 features
Classification Algorithms

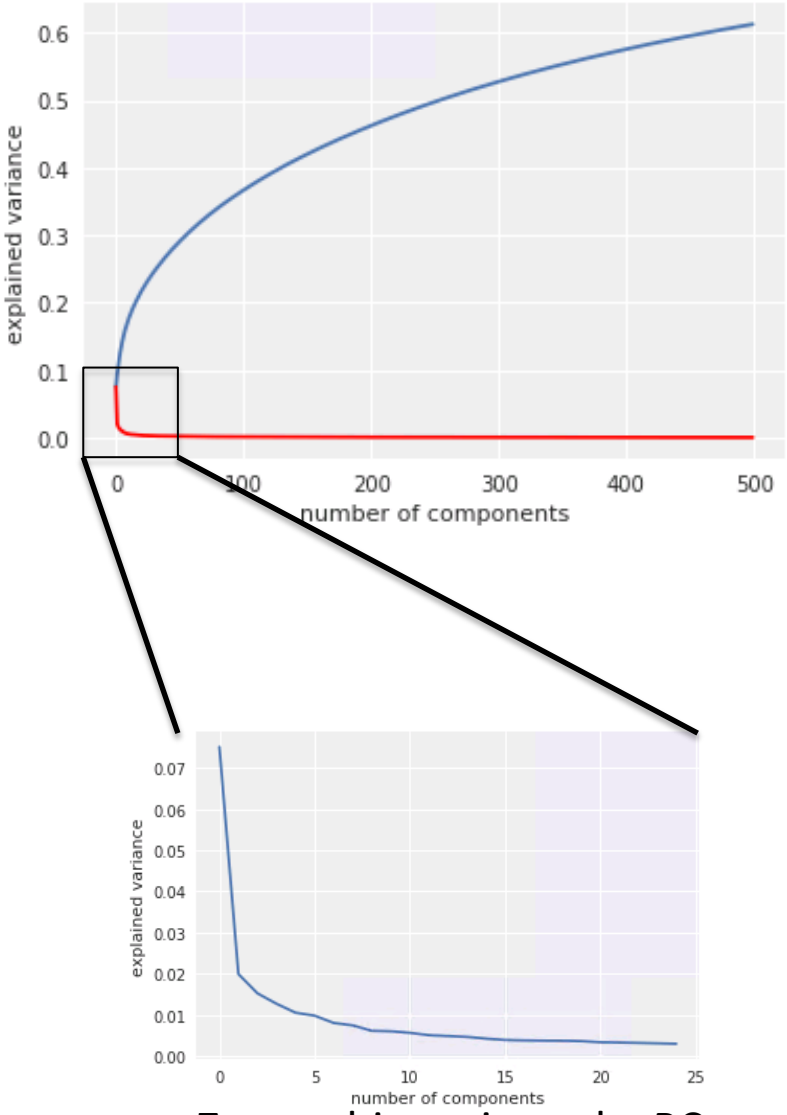
Figure 1.

Figure 2.

A. PCA plot

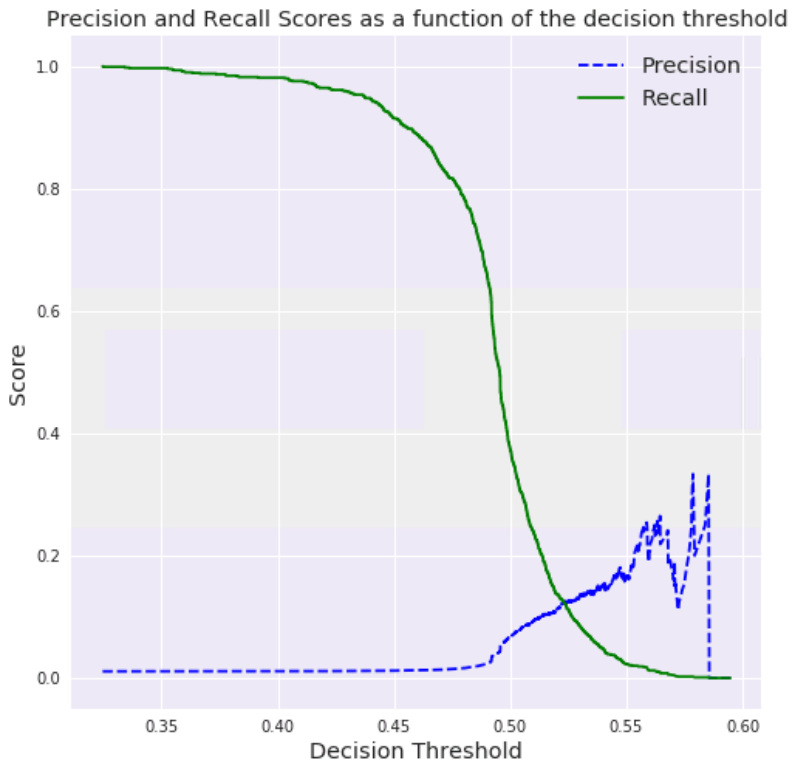


B. Variance explained by PC



Zoomed-in variance by PC

Figure 3.


Precision and Recall Scores as a function of the decision threshold
Precision
Recall
Score
Decision Threshold
1.0
0.8
0.6
0.4
0.2
0.0
0.35
0.40
0.45
0.50
0.55
0.60


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
AUC = 0.65
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0