Atrial Fibrillation Burden Signature and Near-Term Prediction of Stroke: A Machine Learning Analysis

Lichy Han, B.S.[1], **Mariam Askari, B.S.**[2], **Russ B. Altman, M.D. Ph.D.**[3,4], **Susan K. Schmitt, Ph.D.**[2], **Jun Fan, M.S.**[2], **Jason P. Bentley, Ph.D.**[5], **Sanjiv M. Narayan, M.D. Ph.D.**[1,6], **Mintu P. Turakhia, M.D. M.A.S.**[2,6,7]

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA

[2]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA

[3]Department of Bioengineering, Stanford University

[4]Department of Genetics, Stanford University

[5]Quantitative Sciences Unit, Stanford University School of Medicine

[6]Department of Medicine, Stanford University School of Medicine

[7]Center for Digital Health, Stanford University School of Medicine, Stanford, CA

Abstract

Background—Atrial fibrillation (AF) increases the risk of stroke 5-fold and there is rising interest to determine if AF severity or burden can further risk stratify these patients, particularly for near-term events. Using continuous remote monitoring data from cardiac implantable electronic devices (CIED), we sought to evaluate if machine learned signatures of AF burden could provide prognostic information on near-term risk of stroke when compared to conventional risk scores.

Methods and Results—We retrospectively identified Veterans Health Administration (VA) serviced patients with CIED remote monitoring data and at least one day of device-registered AF. The first 30 days of remote monitoring in non-stroke controls were compared against the last 30 days of remote monitoring prior to stroke in cases. We trained three types of models on our data: 1) convolutional neural networks (CNN), 2) random forest (RF), and 3) L1 regularized logistic regression (LASSO). We calculated the CHA_2DS_2-VASc score for each patient and compared its performance against machine learned indices based on AF burden in separate test cohorts. Finally, we investigated the effect of combining our AF burden models with CHA_2DS_2-VASc. We identified 3,114 non-stroke controls and 71 stroke cases, with no significant differences in baseline characteristics. RF performed the best in the test dataset (AUC=0.662) and CNN in the validation dataset (AUC=0.702); whereas, CHA_2DS_2-VASc had an AUC of 0.5 or less in both datasets. Combining CHA_2DS_2-VASc with random forest and CNN yielded a validation AUC of 0.696 and test AUC of 0.634, yielding the highest average AUC on non-training data.

Address for Correspondence: Mintu Turakhia, M.D. M.A.S, VA Palo Alto Health Care System, 3801 Miranda Ave (111C), Palo Alto, CA 94304, Tel (650) 858-3932, mintu@stanford.edu.

Conclusions—This proof of concept study found that machine learning and ensemble methods that incorporate daily AF burden signature provided incremental prognostic value for risk stratification beyond CHA$_2$DS$_2$-VASc for near-term risk of stroke.

Atrial fibrillation (AF) affects three to five million people in the United States.[1] Stroke is the most devastating consequence of AF, accounting for 7% of all AF-associated deaths.[1] Among the 700,000 strokes per year in the U.S., 15% are associated with AF.[2] Oral anticoagulants, which can prevent stroke in AF but with a tradeoff of increased bleeding risk, are prescribed to patients identified by clinical risk scores based on age, comorbidities, and sex.[3,4] Unfortunately, these scores, derived from administrative claims or chart data, perform only modestly outside of their derivation cohorts, with C-statistics generally below 0.65.[5,6]

There is growing interest in assessing whether AF severity or burden of atrial fibrillation can further risk stratify patients. A *post-hoc* analysis of the ACTIVE-A and AVERROES trials found that permanent and persistent forms of AF conferred a greater risk of stroke compared to paroxysmal AF.[7] However, these patterns were identified based on clinical classification by the research coordinator or clinician, which has been shown to poorly reflect the true temporal pattern of AF based on continuous implantable AF monitoring.[7,8] Previously our group has found that transient AF increases the short-term risk of stroke in cardiac implantable electronic device (CIED) patients.[9] Although risk was increased, there was no empiric risk threshold observed based on any cut point of daily AF burden. Other CIED studies similarly shown no clear risk threshold.[10,11] One potential limitation of these analyses is that conventional biostatistical analysis of static baseline measures may not be well suited for time-varying AF burden.

Artificial intelligence (AI) approaches, including machine learning and neural networks, have been used to find patterns in complex datasets outside of health care[12] and in certain clinical applications, such as diagnostic imaging. Machine learning has been used to identify novel clinical phenotypes of AF based on clinical data.[13] However, machine learning has not been applied for clinical classification or risk stratification using continuous AF burden data. Patterns of paroxysmal AF may be complex and confer unique patterns of risk of hypercoagulability, thrombus formation, cardiac embolization, and other factors based on temporal pattern and duration, particularly when evaluating near-term risk of stroke.

Given previous findings of the transient increase in risk of stroke based on upward changes in AF in the 30 days preceding stroke, we sought to evaluate if machine learned signatures of AF burden in a similar time period could provide prognostic information on near-term risk of stroke when compared to conventional risk scores using continuous remote monitoring data from cardiac implantable electronic devices.

METHODS

We performed a retrospective cohort study consisting of 9,836 patients with cardiovascular implantable electronic devices (CIED) that registered in remotely monitoring managed by the Veterans Administration (VA) Health Care System between 2004 and 2009. The VA has a centralized remote monitoring program for over 30,000 devices, and 65% of devices

implanted in the VA are enrolled,[14] therefore encompassing the full denominator of remotely-monitored patients within the VA system.

Devices were limited to pacemakers and implantable cardioverter defibrillators with atrial leads from Medtronic Inc (Mounds View, MN), which is the most common device brand represented in the VA remote monitoring and also encodes AF burden as a day-level variable. The daily AF burden from these monitored devices was recorded through the VA centralized remote monitoring program, and the linkage of device data to medical records and electronic claims was performed as previously described.[9] Daily AF burden is stored as the cumulative amount of time during a full calendar day that the patient meets AF detection criteria and this value can combine multiple episodes of AF in a given 24-hour period.

The device uses the number and timing of atrial events between ventricular events to determine evidence for AF episodes, where, for 32 ventricular intervals, at least 2 atrial sensed events must occur per ventricular interval. Additionally, the median of the 12 most recent sensed atrial intervals must be shorter than that of the programmed AF detection interval (nominally 350 ms). Previous studies have evaluated the AF detection algorithm extensively and shows the device to quantify AF burden with 99% accuracy.[15–17]

We used linked administrative claims data and electronic health records of multiple VA centralized data sets. These data include the VA National Patient Care Database with demographic, outpatient, inpatient, and long-term care administrative data, the VA Decision Support System national pharmacy extract with patient level data on inpatient and outpatient medication administration and costs, the VA Fee Basis Inpatient and Outpatient data sets and the VA Vital Status File with validated combined demographic data from the VA, Social Security Administration, and Medicare.

We excluded patients with oral anticoagulation at baseline and those with an ICD-9 code indicating a transient ischemic attack or stroke due to occlusion of the carotid artery or its branches (Figure 1). Subjects with at least one day of device-registered AF in these time intervals were included. From these data, 30-day intervals were created for each subject (Figure 2). We selected the first 30 days of remote monitoring in non-stroke controls, to be compared against the last 30 days of remote monitoring prior to stroke in cases. These 30 days of AF burden served as features for our machine learning models. Two cases in the database were not eligible for inclusion due to not having 30 consecutive days of remote monitoring preceding the index stroke. This study was approved by the Stanford University School of Medicine Institutional Review Board and local VA Research & Development Committee. The data, analytic methods, and study materials will not be made available to other researchers for purposes of reproducing the results.

Supervised Machine Learning using Daily AF Burden

All machine learning analyses were performed using R 3.4.3 (GNU public license) and Python 2.7.9 (public license, Corporation for National Research Initiatives, Reston, VA). Cases and controls were each randomly divided into three groups: 50% for training, 20% for validation and tuning, and 30% for testing (Figure 2). The case and control subsets were

then combined to form the training, validation, and test datasets, thus maintaining the outcome rate of each dataset.

We trained three types of models on our data: 1) convolutional neural networks (CNN), 2) random forest (RF), and 3) L1 regularized logistic regression (LASSO) (Figure 2). We implemented these classifiers using keras with tensorflow backend, randomForest[18], and glmnet[19] packages, respectively. Although recurrent neural networks are also frequently used for sequential data, we chose convolutional neural networks because they are adept at extracting patterns across neighboring features, enabling us to detect patterns from AF incidence distributions over several days. Detection of such features enables us to gain more insight into the feature engineering process of the neural network and allows us to better interpret the risk predictions in the context of certain AF burden patterns. Random forest is an ensemble method for partitioning data using many decision trees, which can be highly successful in situations where decision boundaries are irregular.[20] Logistic regression is a linear classifier frequently used for classification in medical literature,[21] and the L1 penalty was imposed to control for the correlated nature of time series features and to prevent overfitting.[22]

We used the validation dataset to tune model hyperparameters (Figure 2). For CNN, we assessed the following hyperparameters: number of CNN layers (1, 2, 3), learning rate (1e-6, 1e-5, 1e-4), dropout rate (0.7, 0.9), number of filters (8, 16), kernel size (3, 5), batch size during training (128, 256), and class weights (1:1, 1:30). 1:30 was chosen to approximate the outcome rate of 0.037 of cases to controls in the training dataset. We additionally tested the same class weights in the random forest and LASSO models and used 10-fold cross validation on the training dataset to determine the optimal value for the penalty parameter. The final architecture of the CNN was a 3 layer model with a learning rate of 1e-6, dropout rate of 0.9, kernel size of 5, batch of 256, and class weights of 1:30. All models were evaluated for their ability to discriminate cases from controls using the area under the curve (AUC) of the receiver operating characteristic (ROC) (Figure 2). For calculating the sensitivity and specificity, we chose the optimal cutoff point on the ROC curve from the training data and applied it to the validation and test data. We then compared ROC curves using the bootstrap method. We did not apply cross-validation to the outer model due to the limited number of positive cases.

To assess AF burden feature importance, we extracted the mean decrease in the Gini coefficient from the random forest classifier, which represents the feature importance score for each day of the AF interval. For CNN, we performed gradient ascent to search for the AF pattern that maximized each filter's output. We then normalized the output to visualize each filter. By doing so, we extracted AF patterns discovered by our CNN, which may yield insight into types of AF patterns that may confer greater stroke risk.

Logistic Regression with CHA$_2$DS$_2$-VASc

We calculated the CHA$_2$DS$_2$-VASc score for each patient to compare the performance of this clinical risk score with our machine learning approaches based on AF burden. We used logistic regression to predict stroke from the CHA$_2$DS$_2$-VASc score as a continuous variable and compared its performance by the AUC.

Ensemble Models Combining CHA$_2$DS$_2$-VASc with AF Burden

We further investigated the effect of combining our AF burden models with CHA$_2$DS$_2$-VASc. Typically, in ensemble methods, each individual model receives a vote toward the final classification of each patient. We incorporated this by averaging the probabilities of all classifiers constructed, so that each classifier contributed equally to the prediction.

RESULTS

We identified 3,114 non-stroke controls and 71 cases from our cohort of VA patients with remote monitoring data. Baseline characteristics of AF patients with and without stroke are presented in Table 1. Notably, there was no significant difference in CHA$_2$DS$_2$-VASc score between AF patients with stroke and controls.

Machine Learning Using AF Burden

Applying our AF burden based models to our data generated c-statistics between 0.6 and 0.7 in derivation and validation cohorts (Figure 3, Table 2), compared to CHA$_2$DS$_2$-VASc which yielded an AUC of 0.5 or less in derivation and validation cohorts (Figure 3, Table 2). Compared to CHA$_2$DS$_2$-VASc, CNN provided superior discrimination in the validation cohort, with an AUC of 0.702 (p = 0.0003). Random forest analysis outperformed CNN in the test cohort, but with a high degree of overfitting to the training data which was not observed using the CNN. After utilizing a calibration plot (20 bins) on the test dataset to further evaluate performance, the ensemble method, while underconfident in that maximum probability is 40–50%, exhibited an increasing observed event percentage with each bin with the best spread. (Supplemental Figure). Notably, there is heavy class imbalance, with most events being predicted from controls and an overall small N. Other methods, including the CHA$_2$DS$_2$-VASc score, tended to either have no relationship or were clustered at low or high predicted probabilities.

Models Combining CHA$_2$DS$_2$-VASc and AF Burden

Combining CHA$_2$DS$_2$-VASc with random forest and CNN yielded a validation AUC of 0.696 and test AUC of 0.634, yielding the highest average AUC on non-training data. Compared to single classifiers, the ensemble classifier maintained good performance on the validation dataset and improved the AUC achieved in the test cohort from 0.524 with CHA$_2$DS$_2$-VASc alone to 0.634. Exploring the raw probability values in the test set revealed that though stroke cases tended to be associated with a higher probability score, the classifier tended to be underconfident, likely stemming from class imbalance.

Both sensitivity and specificity improved in the ensemble method over using only CHA$_2$DS$_2$-VASc. The single method classifiers tended to have cutoff points that balanced sensitivity and specificity, while using CHA$_2$DS$_2$-VASc alone or in the ensemble method let to prioritizing specificity.

Assessing Important AF Burden Features

The random forest feature importance scores (Figure 4) revealed that the amount of AF on day 30 was the most important feature, followed by day 29. The other 28 days had lower

importance scores, although with a cyclic pattern (for instance, higher importance scores on days 1, 8, and 17).

These cyclic patterns arose again when we evaluated the filters of our CNN model. Our final CNN had 3 layers with 16 filters each, and filters for the first two layers are shown in Figure 5. The first layer shows filters that appear sensitive to different patterns at the beginning and end of 30-day intervals (Figure 5A). Filters in the second layer appear mostly to evaluate periodicity in AF burden (Figure 5B). For example, filter 6 in the second layer appears to be looking for 3-day long cycles, with increasing importance attributed to the later days. Filter 12, on the other hand, focuses on rapid alterations in AF burden which were uniform throughout 30 day monitoring windows. The filters of the third layer (not shown) were similar to those in the second layer. Thus, CNN filters learned different types of AF burden patterns, and appeared to recognize AF burden in the first and last few days, and on alternating cycles.

DISCUSSION

In this work we present an approach that applies machine learning to granular data on daily AF burden in a large well-characterized cohort of AF patients with and without stroke, to build classification models for short-term stroke prediction. We found that a variety of machine learning techniques that evaluated the AF burden signature were superior for discrimination of stroke compared to the CHA_2DS_2-VASc score. In particular, ensemble models such as random forest attained the best performance in non-training data, outperforming CHA_2DS_2-VASc by up to 0.14 in the AUC which achieved statistical significance for selected machine learning methods. Additional ensemble models that included CHA_2DS_2-VASc also maintained good performance on non-training data and may be useful extensions to CHA_2DS_2-VASc alone. These findings suggest that the AF burden signature, which indicates temporal patterns in AF onset and offset over 30 days which extend prior 'averages' of AF burden, could provide incremental prognostic value for risk stratification, particularly for near-term risk of stroke.

Comparison to Conventional Risk Stratification

Conventional risk stratification for stroke in AF relies mostly on clinical variables derived largely from administrative data and have not incorporated measures of AF burden or duration. A recent registry attempted to identify stroke risk based on AF burden, but relied on multivariable regression and AF prespecified burden thresholds to bin and aggregate patients.[10] A limitation of the approach is that cutpoints may not represent physiological risk thresholds. Additionally, they do not account for changes in burden over time but rather a static measure on the day of baseline ascertainment. In our machine learning analysis, we explored both linear and nonlinear combinations of variables to find the optimal boundary between stroke cases and non-stroke controls in high dimensions rather than prespecifying a series of arbitrary AF burden thresholds, thereby allowing us to discover unforeseen combinations of variables which improved model discrimination. We focused on near-term disk of stroke, predicated on our prior work that indicated that near-term risk of stroke was not static and was temporally and transiently influenced by AF burden.[9] Despite the

improved classification seen in this study, there may be challenges in longer-term prediction, as prior device detected AF studies have shown a large amount of temporal discordance between AF and the timing of ischemic stroke.[23] Nonetheless, with machine learning approaches, we can apply our models in real-time to determine stroke risk continuously for the subsequent days to weeks to months, rather than previous approaches which evaluate stroke risk in the following years simultaneously or indefinitely. Such an approach could potentially allow for real-time monitoring and active changes in anti-coagulation therapy based on a patient's current risk profile. A rhythm-guided approach in patients with insertable cardiac monitors and paroxysmal atrial fibrillation was previously piloted and demonstrated a 94% reduction in anticoagulation exposure but was not powered for stroke events.[24,25] However, a previous trial of rhythm-guided vs sustained anticoagulation using warfarin in patients with CIEDs did meet its primary endpoint[23], but the trial relied on arbitrary cutpoints of AF burden. If confirmed in larger cohorts, incorporation of AF signatures rather than burden alone may help to improve short-term risk stratification to dynamically guide therapy.

Limitations

While this is a proof-of-concept study, there remain significant limitations. First, this is a population of patients with implantable pacemakers and defibrillators, who are at higher risk of development and progression of AF, of stroke, and also of competing causes for death. Findings may therefore not generalize to patients without implantable devices. Second, we did not account for time varying exposures such as medication (anticoagulation, antiarrhythmic drugs) or rhythm-restoring procedures such as ablation or cardioversion. Third, this is a predominately male veteran cohort and data may not be generalizable to women, which notably could also affect discrimination for scores using $CHA_2DS_2\text{-}VASc$, which includes sex in the score calculation. Fourth, although the overall cohort is large, the number of stroke cases is small, limited by data availability and stroke incidence. Fifth, we used a selected number of patients who were not treated with oral anticoagulation despite the presence of atrial fibrillation and during a period over which professional society guidelines would have recommended treatment based on $CHA_2DS_2\text{-}VASc$ in most case and control patients in our study. This may represent the Achilles' Heel of trying to develop new stroke risk stratification tools in AF, regardless of whether clinical, device, or other data are used. Finally, $CHA_2DS_2\text{-}VASc$ was not designed or calibrated for short-term prediction, although we felt that it was the most appropriate reference risk score against which to test. Validation of prediction methods in other cohorts would help to confirm calibration and generalizability, although the present cohort remains the largest cohort of remote monitoring linked to comprehensive electronic health record and claims data.

CONCLUSIONS

In summary, this proof of concept study found that machine learning and ensemble methods that incorporate daily AF burden signatures from time-varying patterns of AF incidence in 30-day windows has the potential to improve or refine short-term stroke prediction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, de Ferranti SD, Ferguson JF, Fornage M, Gillespie C, Isasi CR, Jiménez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Lutsey PL, Mackey JS, Matchar DB, Matsushita K, Mussolino ME, Nasir K, O'Flaherty M, Palaniappan LP, Pandey A, Pandey DK, Reeves MJ, Ritchey MD, Rodriguez CJ, Roth GA, Rosamond WD, Sampson UKA, Satou GM, Shah SH, Spartano NL, Tirschwell DL, Tsao CW, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS, Muntner P, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. Circulation. 2018;137:e67–e492. [PubMed: 29386200]

2. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation: a major contributor to stroke in the elderly. The Framingham Study. Arch Intern Med. 1987;147:1561–1564. [PubMed: 3632164]

3. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The Euro Heart Survey on atrial fibrillation. Chest. 2010;137:263–272. [PubMed: 19762550]

4. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of Clinical Classification Schemes Results From the National Registry of Atrial Fibrillation. Jama. 2001;285:2864–2870. [PubMed: 11401607]

5. Fang MC, Go AS, Chang Y, Borowsky L, Pomernacki NK, Singer DE. Comparison of risk stratification schemes to predict thromboembolism in people with nonvalvular atrial fibrillation. Journal of the American College of Cardiology. 2008;51:810–815. [PubMed: 18294564]

6. Piccini JP, Stevens SR, Chang Y, Singer DE, Lokhnygina Y, Go AS, Patel MR, Mahaffey KW, Halperin JL, Breithardt G, Hankey GJ, Hacke W, Becker RC, Nessel CC, Fox KAA, Califf RM. Renal dysfunction as a predictor of stroke and systemic embolism in patients with nonvalvular atrial fibrillation: Validation of the R2CHADS2 index in the ROCKET AF. Circulation. 2013;127:224–232. [PubMed: 23212720]

7. Vanassche T, Lauw MN, Eikelboom JW, Healey JS, Hart RG, Alings M, Avezum A, Díaz R, Hohnloser SH, Lewis BS, Shestakovska O, Wang J, Connolly SJ. Risk of ischaemic stroke according to pattern of atrial fibrillation: Analysis of 6563 aspirin-treated patients in active-a and averroes. European Heart Journal. 2015;36:281–287. [PubMed: 25187524]

8. Charitos EI, Pürerfellner H, Glotzer TV, Ziegler PD. Clinical classifications of atrial fibrillation poorly reflect its temporal persistence: insights from 1,195 patients continuously monitored with implantable devices. J Am Coll Cardiol. 2014;63:2840–2848. [PubMed: 24814497]

9. Turakhia MP, Ziegler PD, Schmitt SK, Chang Y, Fan J, Than CT, Keung EK, Singer DE. Atrial Fibrillation Burden and Short-Term Risk of Stroke: Case-Crossover Analysis of Continuously Recorded Heart Rhythm from Cardiac Electronic Implanted Devices. Circulation: Arrhythmia and Electrophysiology. 2015;8:1040–1047. [PubMed: 26175528]

10. Boriani G, Glotzer TV, Santini M, West TM, De Melis M, Sepsi M, Gasparini M, Lewalter T, Camm JA, Singer DE. Device-detected atrial fibrillation and risk for stroke: an analysis of >10 000 patients from the SOS AF project (Stroke preventiOn Strategies based on Atrial Fibrillation information from implanted devices). European Heart Journal. 2014;35:508–516. [PubMed: 24334432]

11. Swiryn S, Orlov MV, Benditt DG, DiMarco JP, Lloyd-Jones DM, Karst E, Qu F, Slawsky MT, Turkel M, Waldo AL, RATE Registry Investigators. Clinical Implications of Brief Device-Detected Atrial Tachyarrhythmias in a Cardiac Rhythm Management Device Population: Results from the Registry of Atrial Tachycardia and Atrial Fibrillation Episodes. Circulation. 2016;134:1130–1140. [PubMed: 27754946]

12. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature. 2015;518:529–533. [PubMed: 25719670]

13. Inohara T, Shrader P, Pieper K, Blanco RG, Thomas L, Singer DE, Freeman JV, Allen LA, Fonarow GC, Gersh B, Ezekowitz MD, Kowey PR, Reiffel JA, Naccarelli GV, Chan PS, Steinberg BA, Peterson ED, Piccini JP. Association of of Atrial Fibrillation Clinical Phenotypes With Treatment Patterns and Outcomes: A Multicenter Registry Study. JAMA Cardiol. 2018;3:54–63. [PubMed: 29128866]

14. Sung RK, Massie BM, Varosy PD, Moore H, Rumsfeld J, Lee BK, Keung E. Long-term electrical survival analysis of Riata and Riata ST silicone leads: National Veterans Affairs experience. Heart Rhythm. 2012;9:1954–1961. [PubMed: 22871583]

15. Purerfellner H, Gillis AM, Holbrook R, Hettrick DA. Accuracy of atrial tachyarrhythmia detection in implantable devices with arrhythmia therapies. Pacing Clin Electrophysiol. 2004;27:983–992. [PubMed: 15271020]

16. Passman RS, Weinberg KM, Freher M, Denes P, Schaechter A, Goldberger JJ, Kadish AH. Accuracy of mode switch algorithms for detection of atrial tachyarrhythmias. J Cardiovasc Electrophysiol. 2004;15:773–777. [PubMed: 15250860]

17. Swerdlow CD, Schsls W, Dijkman B, Jung W, Sheth NV, Olson WH, Gunderson BD. Detection of atrial fibrillation and flutter by a dual-chamber implantable cardioverter-defibrillator. For the Worldwide Jewel AF Investigators. Circulation. 2000;101:878–885. [PubMed: 10694527]

18. Liaw A, Weiner M. Classification and Regression by randomForest. R News. 2002;2:18–22.

19. Helleputte T, Gramme P. LiblineaR: linear predictive models based on the LIBLINEAR C/C++ library. R package version. 2015;1.94–2.

20. Breiman L. Random Forests. Machine Learning (). 2001;45:5–32.

21. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. Journal of clinical epidemiology. 2001;54:979–985. [PubMed: 11576808]

22. Lee S-I, Lee H, Abbeel P, Ng AY. Efficient L1 Regularized Logistic Regression. AAAI. 2006;401–408.

23. Brambatti M, Connolly SJ, Gold MR, Morillo CA, Capucci A, Muto C, Lau CP, Van Gelder IC, Hohnloser SH, Carlson M, Fain E, Nakamya J, Mairesse GH, Halytska M, Deng WQ, Israel CW, Healey JS, ASSERT Investigators. Temporal relationship between subclinical atrial fibrillation and embolic events. Circulation. 2014;129:2094–2099. [PubMed: 24633881]

24. Passman R, Leong-Sit P, Andrei A-C, Huskin A, Tomson TT, Bernstein R, Ellis E, Waks JW, Zimetbaum P. Targeted Anticoagulation for Atrial Fibrillation Guided by Continuous Rhythm Assessment With an Insertable Cardiac Monitor: The Rhythm Evaluation for Anticoagulation With Continuous Monitoring (REACT.COM) Pilot Study. J Cardiovasc Electrophysiol. 2016;27:264–270. [PubMed: 26511221]

25. . Martin DT, Bersohn MM, Waldo AL, Wathen MS, Choucair WK, Lip GYH, Ip J, Holcomb R, Akar JG, Halperin JL, IMPACT Investigators. Randomized trial of atrial arrhythmia monitoring to guide anticoagulation in patients with implanted defibrillator and cardiac resynchronization devices. Eur Heart J. 2015;36:1660–1668. [PubMed: 25908774]

CLINICAL PERSPECTIVE

What is Known?

- Recent clinical trials found that permanent and persistent forms of atrial fibrillation (AF) lead to an increased risk in stroke when compared to paroxysmal AF.

- Transient AF may increase the short-term risk of stroke in patients with cardiac implantable electronic devices (CIED).

What the Study Adds:

- This proof-of-concept illustrates the feasibility of combining standard risk scores that currently guide disease management with remote monitoring AF burden data.

- Machine learning techniques that evaluate AF burden signature are superior for discrimination of stroke when compared to the CHA_2DS_2-VASc score. Ensemble models, such as random forest, incorporating CHA_2DS_2-VASc performed well and can be extensions to using CHA_2DS_2-VASc alone.
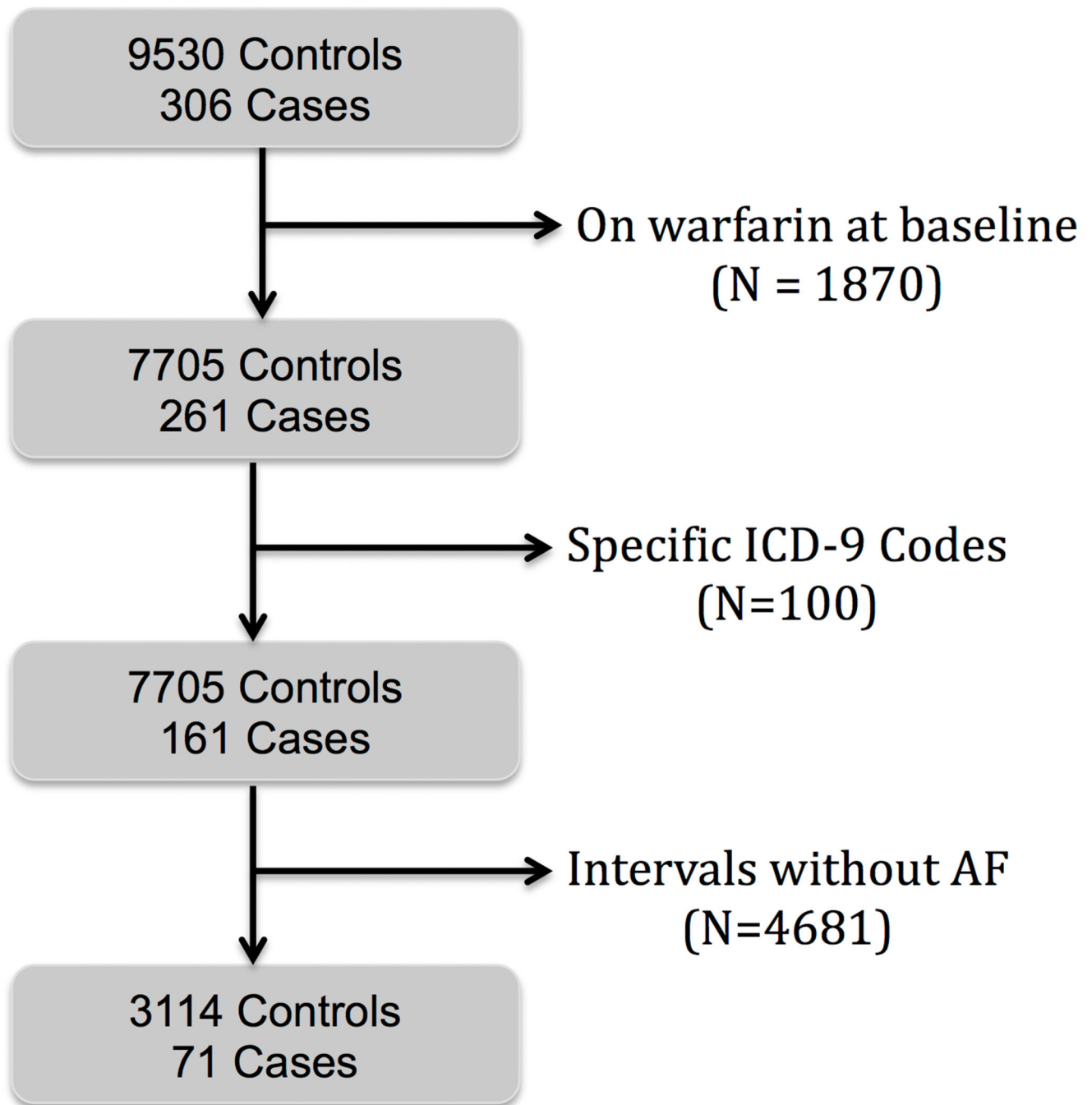
Figure 1.
Overview of subject selection.
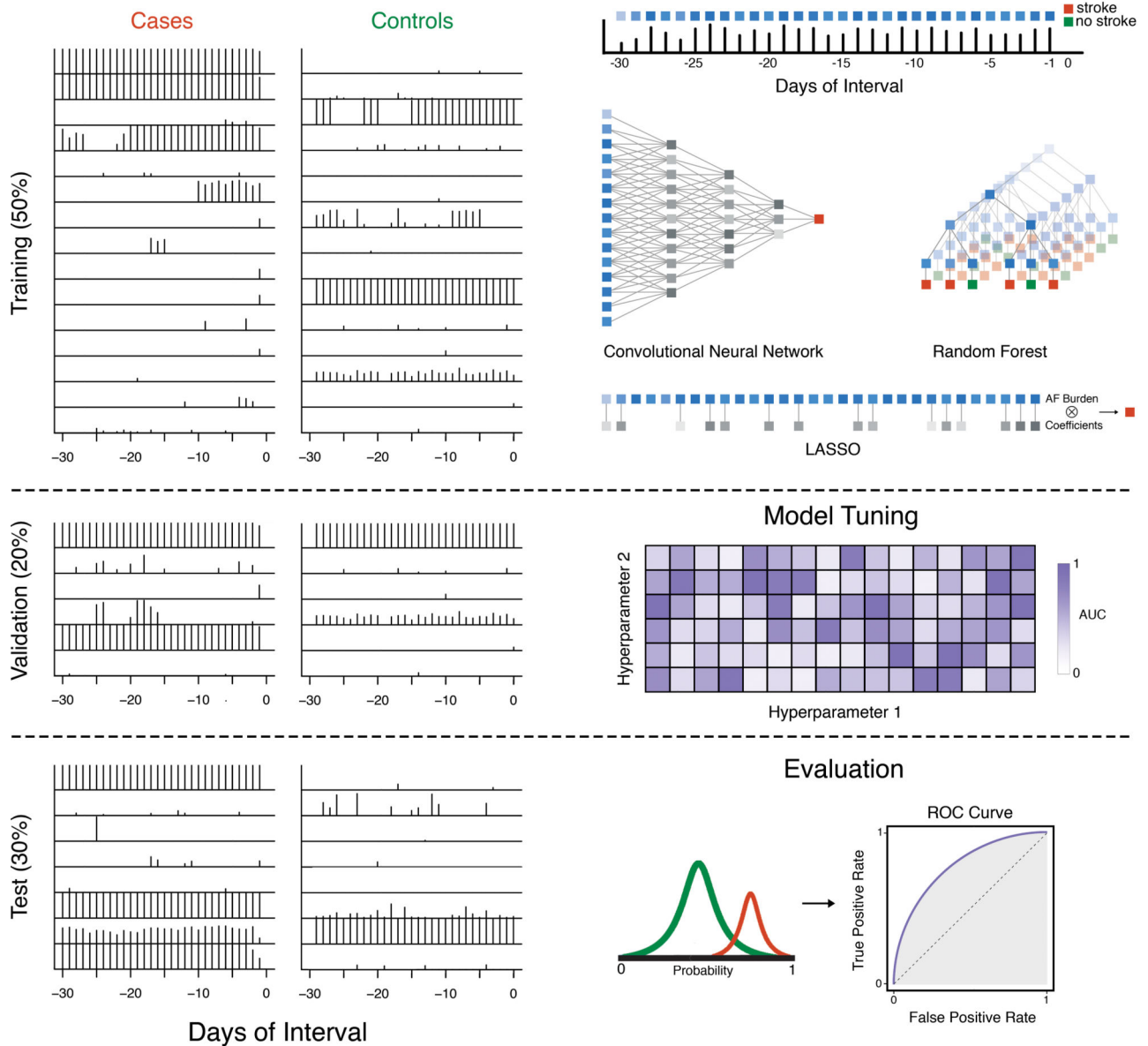
Figure 2. Workflow.
The entire dataset is first split into training (50%), validation (20%) and test (30%) datasets. Then, each day of AF burden in the 30-day interval is considered a feature, which is used as inputs to three machine learning models (CNN, random forest, LASSO) on the training data. The validation dataset is used to tune these models. Trained models are subsequently evaluated on the test dataset.
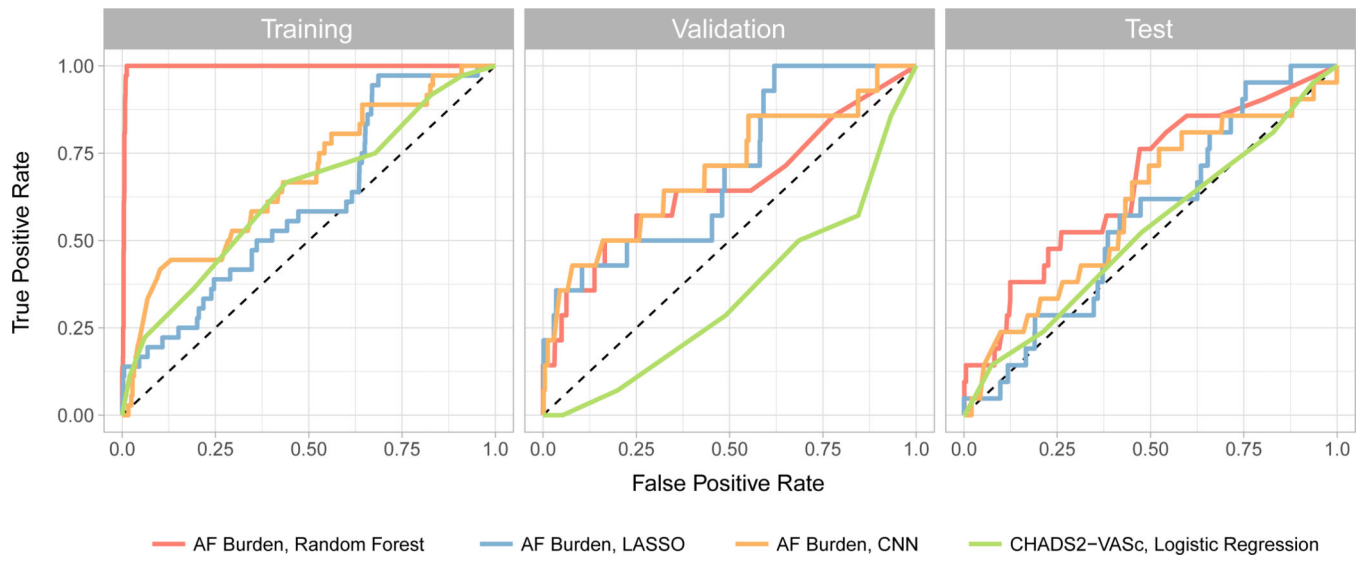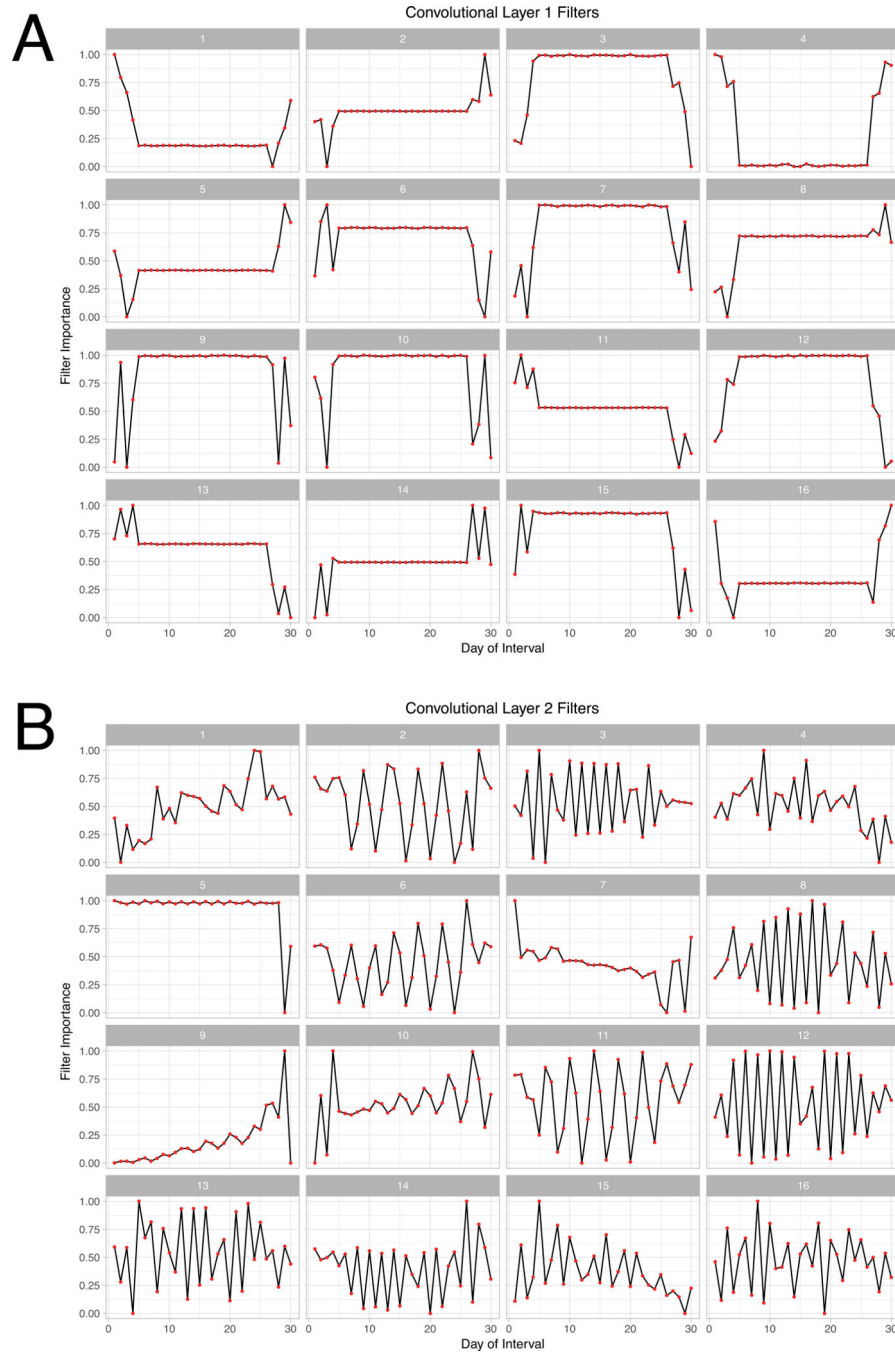
Figure 3.
ROC curves for AF burden classifiers (random forest, LASSO, CNN) and CHA_2DS_2-VASc with logistic regression in the training, validation, and test datasets.

Feature Importance: Day of Interval
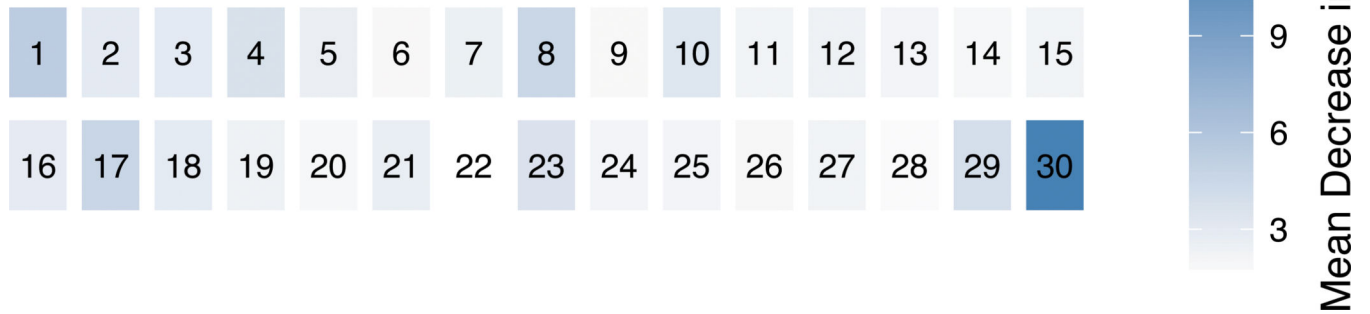
Figure 4.
Heatmap of the mean decrease in the Gini coefficient from the random forest classifier for each day of AF in the 30-day interval.

Figure 5. Filters from layer 1 and 2 of our convolutional neural network.
Filters for the first two layers of the CNN classifier. Each filter represents the AF pattern that maximizes the output of that filter. Filters from layer 1 are focused on changes at the beginning and end of the AF interval, and filters from layer 2 appear to look for periodic patterns.

Table 1.

Baseline characteristics of patients with and without stroke

	Stroke Cases	Non-Stroke Controls	
Demographics	**(N=71)**	**(N=3114)**	**p-value**
Age, mean ± SD	69.0 ± 11.0	67.6 ± 10.3	0.32
Male, n(%)	70 (98.6)	3068 (98.5)	1.00
Prior MI, n(%)	6 (8)	418 (13)	0.29
Hypertension, n(%)	54 (76)	2196 (71)	0.36
Heart Failure, n(%)	45 (63)	2154 (69)	0.30
Diabetes Mellitus Type 2, n(%)	33 (46)	1252 (40)	0.33
Peripheral Artery Disease, n(%)	12 (17)	293 (9)	0.04
Coronary Artery Disease, n(%)	51 (72)	2156 (69)	0.70
Charlson Comorbidity Index, mean ± SD	2.7 ± 2.0	2.7 ± 1.8	0.98
Selim Comorbidity Score, mean± SD	4.9 ± 3.0	5.2 ± 2.9	0.51
CHA_2DS_2-VASc Score, mean ± SD	3.5 ± 1.9	3.2 ± 1.6	0.24
CHA_2DS_2-VASc Score by Group			
0, n(%)	4 (5.6)	240 (7.7)	
1, n(%)	9 (12.7)	285 (9.2)	
2, n(%)	9 (12.7)	444 (14.3)	
3, n(%)	10 (14.1)	710 (22.8)	
4, n(%)	20 (28.1)	815 (26.2)	
5, n(%)	8 (11.3)	424 (13.6)	
6, n(%)	6 (8.5)	132 (4.2)	
7, n(%)	5 (7.0)	63 (2.0)	
8, n(%)	0 (0)	1 (0.03)	
9, n(%)	0 (0)	0 (0)	

Table 2.

Classification AUCs in the training, validation, and test datasets. P-values shown are in comparison to CHA_2DS_2-VASc.

	CHA_2DS_2-VASc	Random Forest	CNN	LASSO	Ensemble
Training AUC	0.634	0.996	0.678	0.606	0.996
Validation AUC	0.346	0.653	0.702	0.701	0.696
P-Value		0.007	0.0003	0.0004	0.001
Sensitivity	0.071	0.571	0.571	0.500	0.143
Specificity	0.799	0.684	0.717	0.666	0.984
Test AUC	0.524	0.662	0.600	0.564	0.634
P-Value		0.120	0.640	0.650	0.250
Sensitivity	0.238	0.523	0.429	0.381	0.143
Specificity	0.788	0.630	0.667	0.635	0.984