Feasibility and Evaluation of a Large-Scale External Validation Approach for Patient-Level Prediction in an International Data Network

Jenna Marie Reps (jreps@its.jnj.com)
Janssen Research and Development Titusville https://orcid.org/0000-0002-2970-0778
Ross Williams
Erasmus Universiteit Rotterdam
Seng Chan You
Ajou University School of Medicine and Graduate School of Medicine
Thomas Falconer
Columbia University Medical Center
Evan Minty
University of Calgary Cumming School of Medicine
Alison Callahan
Stanford University School of Medicine
Patrick B Ryan
Janssen Research and Development Titusville
Rae Woong Park
Ajou University School of Medicine and Graduate School of Medicine
Peter Rijnbeek
Erasmus MC

Research article

Keywords: patient-level prediction; prognostic model; external validation; transportability; collaborative network

Posted Date: July 20th, 2019

DOI: https://doi.org/10.21203/rs.2.11750/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on May 6th, 2020. See the published version at https://doi.org/10.1186/s12874-020-00991-3.

Abstract

Objective To demonstrate how the Observational Healthcare Data Science and Informatics (OHDSI) collaborative network and standardization can be utilized to externally validate patient-level prediction models at scale. Materials & Methods Five previously published prognostic models (ATRIA, CHADS2, CHADS2VASC, Q-Stroke and Framingham) that predict future risk of stroke in patients with atrial fibrillation were replicated using the OHDSI frameworks and a network study was run that enabled the five models to be externally validated across nine datasets spanning three countries and five independent sites. Results The five existing models were able to be integrated into the OHDSI framework for patient-level prediction and their performances in predicting stroke within 1 year of initial atrial fibrillation diagnosis for females were comparable with existing studies. The validation network study took 60 days once the models were replicated and an R package for the study was published to collaborators at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ExistingStrokeRiskExternalValidation. Discussion This study demonstrates the ability to scale up external validation of patient-level prediction models using a collaboration of researchers and data standardization that enable models to be readily shared across data sites. External validation is necessary to understand the transportability and reproducibility of prediction models, but without collaborative approaches it can take three or more years to be validated by one independent researcher. Conclusion In this paper we show it is possible to both scale-up and speed-up external validation by showing how validation can be done across multiple databases in less than 2 months.

Background

Observational healthcare data often contains longitudinal medical records for large heterogeneous populations. There has been increased interest in learning patient-level prediction models using these real world datasets with the aim of improving healthcare. These patient-level prediction models can be used to identify high-risk subgroups that could benefit from interventions. It is important to ensure a model has good performance before it is used clinically.

Models are often internally validated using the development dataset by withholding a subset of that data from the model training stage so that it can be used for evaluating the model performance. The majority of patient-level prediction models will report internal validation. External validation is accomplished by evaluating the model on a new dataset (that is different from the development dataset). Few published patient-level prediction models are externally validated and research has shown that it often takes three or more years for external validation to occur once a model is published [1].

External validation of a patient-level prediction model can provide useful insights into the model. When a model is validated on a population that has similar characteristics to the development data population the generalizability of the model is investigated (i.e., how well the model performs when making predictions on similar patients). When a model is validated on a population that has different characteristics to the development data population the transportability of the model is investigated (i.e.,

how well the model performs on different patients). Many observational datasets are not representative of the whole population, so the transportability performance of the model discovered during external validation is important to know when identifying who the model can be broadly applied to. For example, some clinical guidelines recommend treatment stratification for patients based on applying a simple risk score model that was developed on a small population and the transportability on the general population may never have been studied.

External validation is a slow process due to the difficulty finding suitable data to replicate a prediction model on and difficulty replicating a prediction model (e.g., writing code to correctly extract the same model covariates from the new data). Often published papers lack the information required to replicate the model or subjectivity (e.g., in defining medical conditions or variables) can be an issue causing models to be replicated incorrectly. This prevents efficient and large-scale external validation which likely slows down clinical uptake of published patient-level prediction models or results in the models being applied clinically to patient populations where the model transportability is unknown.

A collaborative approach to external model validation has been proposed to enable extensive evaluation of prediction models [2]. The Observational Healthcare Data Science and Informatics (OHDSI) network is a community of researchers that are working towards the common goal of improving the analysis of observational data. The OHDSI community have developed standardizations that enable efficient collaboration across research sites. The main standardization is the common data structure and vocabulary used by all collaborators known as the Observational Medical Outcomes Partnership (OMOP) common data model. The OMOP common data model ensures all researchers have their data in the same structure so analysis codes such as SQL can be shared across sites. This has enabled the development of analysis packages in R for causal inference and patient-level prediction that can be used by any researcher with data in the OMOP common data model. The OHDSI collaborative network, common data model and patient-level prediction package now present the opportunity to scale up external validation.

The aim of this study is to demonstrate that the OHDSI tools and OMOP common data model can improve the ability to perform external validation. Instead of taking years to externally validate a model, it may now be possible to apply a prediction models to a large number of datasets in a short period of time. We picked the problem of predicting stroke in atrial fibrillation patients as there are multiple existing models that are used clinically, namely ATRIA (no prior stroke model) [3], Framingham (no prior stroke model) [4] , $CHADS_2$ [5], $CHADS_2$-VASc [6] and Q-Stroke (female model) [7]. We show these models can be replicated using the OHDSI standardization and then externally validated across numerous data sites at scale.

Methods

Existing stoke prediction models

We chose the problem of predicting stroke in patients with atrial fibrillation as it has been well studied and is one of the only prediction problems to have been extensively validated. Therefore, we have ample benchmarks to compare to this study's validations. The existing models we replicated were ATRIA, $CHADS_2$, CHA_2DS_2-VASc, Framingham and Q-Stroke.

The Anticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) [3] model was developed on a cohort of 7,284 patients who were 18+ and had an atrial fibrillation outpatient diagnosis during 1997 or 1998 and internally validated on 3,643 patient hold out set obtaining an AUROC of 0.72. In the same paper, the authors also externally validated the model on a cohort of 33,247 patients aged 21+ with inpatient or outpatient atrial fib or flutter during 2006-2009, obtaining an AUROC of 0.7. The Congestive heart failure, Hypertension, Age > 75, Diabetes, prior Stroke/transient ischemic attack ($CHADS_2$) score [5] was developed by combining two other stroke prediction models (using the variables from these models and assigning points) and was validated on 1,733 patients aged 65 to 95 years who had nonrheumatic AF. The $CHADS_2$ score obtained an AUROC of 0.81 on this population. The CHA_2DS_2-VASc score [6] is another score based model that was developed using knowledge of risk factors. The model was validated on a cohort of 1,577 patients who were 18+ and had AF during 2003 to 2004 from 35 countries. The model obtained an AUROC of 0.61 for this patient population. The Framingham score [4] model was based on a Cox model developed using data from 705 patients aged 55 to 94 with initial atrial fibrillation. The internal validation, using a bootstrap approach, showed an AUROC of 0.66. The Q-Stroke [7] model was developed using primary care data from the UK consisting of 3, 549, 478 patients aged 25-84 with no prior stroke or anticoagulation use (except aspirin) and internally validated on 1, 897, 168 similar patients. When applying the model to predict the 10-year risk of stroke in female patients with AF at baseline, the AUROC was 0.65.

The existing models include a small number of variables, Table 1 summarizes the variables included in each model. Some of the variables are unlikely to be available in claims data and these are highlighted in red. A large number of Q-Stroke variables are not commonly found in claims data (or are UK specific), so this model is difficult to replicate in external non-UK databases. Although the internal validation for some of the models was as high as 0.8, previous external validation of the models tends to achieve an AUROC between 0.6 and 0.7.

The complete definitions for each variable (sets of SNOMED CT or RXNorm codes) are provided in Appendix A.

Prediction task

Within a target population of female patients with newly diagnosed atrial fibrillation and no prior stroke predict who will develop a stroke 1 to 365 days after initial diagnosis of atrial fibrillation.

Target populations definitions

The existing models were applied to two target populations. Both target populations consisted of female patients newly diagnosed with atrial fibrillation and no prior stroke or anticoagulant use but target population 1 was patients ages 65 or older and target population 2 was all ages.

Target population 1: The target populations was defined as females aged 65-95 with either:

- 2 atrial fibrillation records
- 1 atrial fibrillation in an inpatient setting
- 1 atrial fibrillation with an ECG within 30 days prior

and at least 730 days prior database observation and no prior stroke and no prior anticoagulant.

Target population 2: The target populations was defined as females with either:

- 2 atrial fibrillation records
- 1 atrial fibrillation in an inpatient setting
- 1 atrial fibrillation with an ECG within 30 days prior

and at least 730 days prior database observation and no prior stroke and no prior anticoagulant.

Outcome definition

The stroke outcome was defined as:

- An ischemic or hemorrhagic stroke recorded with an inpatient or ER visit

The code sets used to define atrial fibrillation, ECG and ischemic or hemorrhagic stroke are presented in Appendix B. The full analysis code (data creation and model evaluation) is available at: https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ExistingStrokeRiskExternalValidation

Time-at-risk

We predicted stroke occurring 1 day until 365 days after the initial atrial fibrillation start date.

Sensitivity analysis

As using anticoagulants can impact a patient's risk of stroke, in addition to the two target populations we did a sensitivity analysis where we removed people who had an anticoagulant prescription record that may have intervened in the stroke development. In the additional sensitivity target populations we censored patients at the point an anticoagulant was recorded, so any patient with an anticoagulant during the time-at-risk period was effectively removed from the target population unless they had a stroke prior to the anticoagulant,

Statistical analysis

The prediction model performances were evaluated using the area under the receiver operating characteristic (AUROC) curve. Discriminative measure and confidence intervals were also calculated when the data are small.

Datasets

The datasets used to evaluate the models are:

IBM MarketScan® Commercial Database (CCAE) is a United States employer-sponsored insurance health plans claims database. The database contains claims (e.g. inpatient, outpatient, and outpatient pharmacy) from private healthcare coverage to employees, their spouses, and dependents, so patients are aged 65 or younger.

IBM MarketScan® Medicare Supplemental Database (MDCR) represents health services of retirees in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. The patients are aged 65 or older.

IBM MarketScan® Multi-State Medicaid Database (MDCD) contains adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity.

Optum© De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum Claims) is an adjudicated administrative health claims database for members with private health insurance. The population is primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+ years old) however ages are capped at 90 years.

Optum© de-identified Electronic Health Record Dataset (Optum EHR) is a US electron health record containing clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP).

Stanford Translational Research Integrated Database Environment (STRIDE) is a clinical data warehouse that supports clinical and translational research at Stanford University. This resource includes the EHR data of approximately 2 million adult and pediatric patients cared for at either the Stanford Hospital or the Lucile Packard Children's hospital. This study was completed on an OMP-CDM adherent instance of STRIDE.

Columbia University Medical Center's (CUMC) OMOP CDM data come from New York Presbyterian hospital's clinical data warehouse. The database comprises EHR data on approximately 5 million patients and includes information such as diagnoses, procedures, lab measurements and prescriptions.

Ajou Univeresity School Of Medicine (AUSOM) is a database in the format of OMOP-CDM version 5 for entire EHR data from 1994 to 2018 of Korean tertiary hospital, Ajou university hospital. It contains medical record of about 2.9 million patients.

The Integrated Primary Care Information (IPCI) is an electronic health care databases containing patients of Dutch general practitioners.

Each site had institutional review board approval for the analysis approval for the analysis, or used deidentified data and thus the analysis was determined not to be human subjects research and informed consent was not deemed necessary at any site.

Results And Discussion

IPCI did not contain inpatient stroke records, so the models were unable to be evaluated on this dataset. The incidence of stroke in each of the remaining dataset target populations is presented in Table 2. The incidence of stroke during the 1 year following atrial fibrillation diagnosis in the various target populations ranged from approximately 1% in CCAE, STRIDE, AUSOM and Optum EHR to 5% in MDCD and CUMC. Excluding patients with an anticoagulant after atrial fibrillation who did not have a prior stroke increased the incidence rate for all databases except CCAE and STRIDE. This suggests many people under 65 who have a stroke within a year of initial atrial fibrillation diagnosis had a prior

anticoagulant. This may be a consequence of different treatment of patients with atrial fibrillation who are under 65 compared to being 65 and older. Atrial fibrillation patients who are given an anticoagulant when they are younger than 65 may have other risk factors prompting the use of an anticoagulant.

The results of the discriminative ability of the five existing models across all seven datasets that had inpatient stroke recorded are presented in Table 3. As the AUSOM and STRIDE datasets had outcome counts less than 100, we report the performance in Table 3 but do not include it in the aggregate summaries due to uncertainty in the estimates as a result of small sample sizes.

Across the datasets with sufficient outcome counts, ATRIA obtained a mean AUROC of 0.61 (range 0.57-0.64) on the female patients aged 65 or older and a mean AUROC of 0.63 (range 0.58-0.65) on the female patients of all ages. $CHADS_2$ obtained a mean AUROC of 0.58 (range 0.54-0.60) on the female patients aged 65 or older and a mean AUROC of 0.61 (range 0.56-0.63) on the female patients of all ages. CHA_2DS_2VASc obtained a mean AUROC of 0.60 (range 0.55-0.62) on the female patients aged 65 or older and a mean AUROC of 0.63 (range 0.58-0.65) on the female patients of all ages. Framingham obtained a mean AUROC of 0.57 (range 0.55-0.59) on the female patients aged 65 or older and a mean AUROC of 0.60 (range 0.56-0.62) on the female patients of all ages. Q-Stroke obtained a mean AUROC of 0.55 (range 0.53-0.56) on the female patients aged 65 or older and a mean AUROC of 0.57 (range 0.54-0.61) on the female patients of all ages.

The sensitivity analysis shows the AUROC performance of models when removing patients with an anticoagulant and no stroke or an anticoagulant prior to stroke is comparable or better. This makes sense, for example consider the hypothetical situation where a clinical risk model correctly assigns a high risk to a patient who will have a stroke, but this high risk leads to a clinician giving the patient anticoagulants before the stroke that prevent the stroke occurring. In this situation the model's performance will be negatively impacted because of the intervention as the model was correct to assign a high risk but was wrong due to the intervention preventing the stroke. This raises the issue of evaluating models that are already being used clinically or prediction problems where existing guidelines are used to identify patients who should being given preventative medicine. A fair evaluation is simple when there is no clinical intervention, but complex when preventative medicine exists for the outcome.

The validation performance of the models replicated using the OHDSI patient-level prediction framework and validated across the OHDSI network are comparable with other published results. The Q-Stroke model performed the worst out of all the existing models, but this is likely due to many variables of that model being specific to the UK or not well recorded in claims data. This may indicate that Q-Stroke is not transportable to the US population. In addition, the performances of the models were worse when applied to older females as age is a key predictor in many of the models. In future work it would be interesting to investigate applying machine learning methods to learn more advanced models for predicting stroke in older patients with atrial fibrillation and no prior stroke.

The external validation was performed over 60 days by five different research sites. This greatly scales up the current process for external validation that takes more than three years on average for one other researcher to implement the model. The large-scale external validation was only possible because i) the OMOP common data model and OHDSI standardizations enable sharing of analysis code and ii) collaboration that is possible due to the OHDSI network. We recommend researchers who develop prediction models gain insight into their model's transportability by utilizing the OHDSI network's external validation ability. All that is required is to replicate their models using the OHDSI Patient-level prediction framework, which would also enable other researchers to readily implement the model.

The main limitation of this study was the correct replication of existing models. The reason external validation rarely occurs is that many published models lack certain details such as how to define variables, as code lists are often not published. As a best practice patient-level prediction models should provide full definitions for all variables in the model and provide the model. We used the model's variable definitions when published, but when not available we used our own code sets to define the variables. Another limitation in this study is the limited target populations investigated. We chose females aged 65 or older with no prior stroke as that was the intersection of criteria used when developing the five existing stroke models but we also wanted to see the impact of restricting to older patients (as many models use age as a variable), so we included a second target population of all females with no prior stroke. In future work it would be interesting to investigate the performances of the models across many different target populations.

Conclusions

In this paper we demonstrated the ability to scale-up external validation by using a collaborative network where researchers share a common data structure. The existing prediction models were validated on 9 databases across 5 sites within two months. We recommend that researchers utilize the OHDSI network to externally validate their models at scale to gain insight into the generalizability and transportability of their models.

In addition, the results show that the existing stroke in atrial fibrillation models do not perform well at predicting stroke in the target population of older females in datasets we investigated. This prompts further research into whether a better model can be developed.

References

1. Siontis, GC., Tzoulaki, I., Castaldi, PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*.2015;*68*(1):25-34.
2. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605
3. Singer, DE., Chang, Y., Borowsky, LH., et al. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *J Am Heart*

*Assoc.*2013;*2*(3):e000250

4. Wang, TJ., Massaro, JM., Levy, D., et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. *JAMA*. 2003;*290*(8):1049-1056.
5. Gage, BF., Waterman, AD., Shannon, W., et al. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. 2001;*285*(22):2864-2870.
6. Lip, GY., Nieuwlaat, R., Pisters, R., et al. 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*(2):263-272.
7. Hippisley-Cox, J., Coupland, C. and Brindle, P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ*. 2013;*346*:f2573.
8. van den Ham, HA., Klungel, OH., Singer, DE., et al. Comparative performance of ATRIA, CHADS2, and CHA2DS2-VASc risk scores predicting stroke in patients with atrial fibrillation: results from a national primary care database. *J Am Coll Cardiol*. 2015;*66*(17):1851-1859.
9. Aspberg, S., Chang, Y., Atterman, A., et al. Comparison of the ATRIA, CHADS2, and CHA2DS2-VASc stroke risk scores in predicting ischaemic stroke in a large Swedish cohort of patients with atrial fibrillation. *Eur Heart J.*2016;*37*(42):3203-3210.
10. Chao, TF., Liu, CJ., Tuan, TC., et al. Comparisons of CHADS2 and CHA2DS2-VASc scores for stroke risk stratification in atrial fibrillation: which scoring system should be used for Asians?. *Heart Rhythm*. 2016;*13*(1):46-53.
11. Parmar, P., Krishnamurthi, R., Ikram, MA., et al. The Stroke Riskometer™ app: validation of a data collection tool and stroke risk predictor. *Int J Stroke*. 2015;*10*(2):231-244.
12. Van Staa, TP., Setakis, E., Di Tanna, GL, et al. A comparison of risk stratification schemes for stroke in 79 884 atrial fibrillation patients in general practice. *J Thromb Haemost*. 2011;*9*(1):39-48.

References

FUNDING STATEMENT

No sources of funding were used to assist in the preparation of this study.

COMPETING INTERESTS STATEMENT

Jenna Reps is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. Patrick Ryan is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. Peter Rijnbeek works for a research group who received unconditional research grants from

Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, Servier. None of these grants result in a conflict of interest to the content of this paper.

CONTRIBUTORSHIP STATEMENT

JMR lead and all others contributed to the conception and design of the work, the analysis and the interpretation of data for the work. All authors contributed in drafting, revising and approving the final version.

Tables

Variable	ATRIA	Framingham	$CHADS_2$	CHA_2DS_2VASc	Q-Stroke
Age 85+	x				
Age 75-84	x				
Age 65-74	x			x	
Age 60-62		x			
Age 63-66		x			
Age 67-71		x			
Age 72-74		x			
Age 75-77		x			
Age 78-81		x			
Age 82-85		x			
Age 86-90		x			
Age 91-93		x			
Age >93		x			
Age 75+			x	x	
Female	x	x		x	
Diabetes	x	x	x	x	x
CHF	x		x		x
Prior Stroke or TIA		x	x	x	
Hypertension	x		x	x	x
Systolic blood pressure		x			x
Total cholesterol:HDL cholesterol ratio					x
Townsend deprivation score					x
Proteinuria	x				
eGFR<45 or ESRD	x				
Vascular disease				x	
CHF or LV disease				x	
Smoking status					x

Ethnicity					x
CHD					x
Family history of congestive heart failure					x
Atrial fibrillation					x
Rheumatoid arthritis					x
Chronic renal disease					x
Valvular heart disease					x
Internal AUROC					
	0.72	0.66	0.82	0.61	0.65
External AUROCs					
UK EMR 2015 [8]	0.7 (0.69-0.71)	-	0.68 (0.67-0.69)	0.68 (0.67-0.69)	-
Swedish EMR 2016 [9]	0.71 (0.70-0.71)	-	0.69 (0.69-0.70)	0.69 (0.69-0.70)	-
Taiwan 2016 [10]	-	-	0.66	0.70	-
New Zealand, Russia and the Netherlands 2014 [11]	-	0.70 (0.68-0.73)	-	-	0.71 (0.69-0.73)
UK EMR 2010 [12]	-	0.65 (0.63-0.68)	0.66 (0.64-0.68)	0.67 (0.65-0.69)	-

Table 1 Existing models for predicting stroke risk

	Incidence % (Target population size)							
Target Population	CCAE	MDCD	MDCR	Optum claims	Optum EHR	CUMC	AUSOM	STRIDE
T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants	-	4.95 (25,880)	4.40 (89,156)	4.07 (110,905)	1.30 (149,906)	5.46 (3,664)	2.61 (268)	1.37 (3,366)
T2: Females with atrial fibrillation no prior stroke or anticoagulants	1.33 (61,224)	4.61 (33,262)	-	3.49 (139,376)	1.13 (189,815)	4.71 (4,969)	1.76 (455)	1.28 (4,456)
Sensitivity T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	-	5.04 (23,586)	5.26 (56,511)	4.48 (78,353)	1.44 (99,212)	5.99 (2,872)	4.17 (144)	1.29 (2,094)
Sensitivity T2: Females with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	1.28 (46,054)	4.69 (29,546)	-	3.73 (100,757)	1.22 (128,409)	5.04 (3,910)	2.73 (256)	1.22 (2,786)

Table 2 Target population size in each dataset and the incidence of stroke within 1 year of initial atrial fibrillation diagnosis

Target Population	Model	CCAE	MDCD	MDCR	Optum claims	Optum EHR	CUMC	AUSOM	STRIDE
T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants	ATRIA	-	0.57 (0.55-0.58)	0.63 (0.62-0.64)	0.61	0.62	0.64 (0.60-0.68)	0.60 (0.33-0.87)	0.49 (0.40-0.58)
	$CHADS_2$	-	0.54 (0.53-0.56)	0.60 (0.59-0.61)	0.59	0.60	0.60 (0.57-0.64)	0.51 (0.27-0.75)	0.48 (0.39-0.57)
	CHA_2DS_2VASc	-	0.55 (0.53-0.57)	0.60 (0.59-0.61)	0.59	0.62	0.62 (0.58-0.66)	0.53 (0.32-0.74)	0.52 (0.42-0.62)
	Framingham	-	0.55 (0.53-0.56)	0.59 (0.58-0.60)	0.56	0.58	0.57 (0.53-0.61)	0.46 (0.24-0.69)	0.57 (0.49-0.65)
	Q-Stroke	-	0.53 (0.52-0.55)	0.56 (0.55-0.57)	0.55	0.56	0.56 (0.51-0.60)	0.56 (0.29-0.84)	-
T2: Females with atrial fibrillation no prior stroke or anticoagulants	ATRIA	0.62 (0.60-0.64)	0.58 (0.56-0.59)	-	0.65	0.65	0.65 (0.62-0.69)	0.73 (0.58-0.89)	0.52 (0.44-0.60)
	$CHADS_2$	0.61 (0.59-0.62)	0.56 (0.55-0.57)	-	0.62	0.63	0.62 (0.59-0.66)	0.63 (0.43-0.83)	0.50 (0.42-0.57)
	CHA_2DS_2VASc	0.63 (0.61-0.65)	0.58 (0.56-0.59)	-	0.64	0.65	0.64 (0.61-0.68)	0.73 (0.60-0.85)	0.55 (0.47-0.62)
	Framingham	0.61	0.56	-	0.61	0.62	0.61	0.61	0.57

		(0.59-0.63)	(0.55-0.58)				(0.57-0.64)	(0.47-0.76)	(0.50-0.64)
	Q-Stroke	0.61 (0.59-0.63)	0.54 (0.53-0.56)	-	0.57	0.58	0.57 (0.53-0.61)	0.63 (0.39-0.88)	-
Sensitivity T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	ATRIA	-	0.56 (0.55-0.58)	0.63 (0.62-0.64)	0.61 (0.61-0.62)	0.63 (0.61-0.64)	0.66 (0.62-0.70)	0.69 (0.43-0.95)	0.55 (0.47-0.62)
	$CHADS_2$	-	0.54 (0.53-0.56)	0.61 (0.60-0.62)	0.59 (0.58-0.60)	0.61 (0.59-0.62)	0.62 (0.58-0.66)	0.61 (0.36-0.85)	0.51 (0.38-0.63)
	CHA_2DS_2VASc	-	0.55 (0.54-0.57)	0.61 (0.60-0.62)	0.59 (0.58-0.60)	0.63 (0.61-0.64)	0.64 (0.60-0.68)	0.64 (0.45-0.83)	0.55 (0.42-0.67)
	Framingham	-	0.54 (0.53-0.56)	0.59 (0.58-0.60)	0.57 (0.56-0.57)	0.59 (0.58-0.61)	0.59 (0.55-0.63)	0.53 (0.30-0.77)	0.51 (0.40-0.61)
	Q-Stroke	-	0.53 (0.52-0.55)	0.57 (0.55-0.58)	0.55 (0.54-0.56)	0.57 (0.55-0.58)	0.56 (0.52-0.61)	0.61 (0.30-0.92)	-
Sensitivity T2: Females with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	ATRIA	0.63 (0.61-0.66)	0.58 (0.56-0.59)	-	0.67	0.67	0.68 (0.64-0.71)	0.79 (0.63-0.94)	0.53 (0.43-0.63)
	$CHADS_2$	0.62 (0.60-0.65)	0.56 (0.55-0.58)	-	0.64	0.65	0.65 (0.61-0.68)	0.72 (0.53-0.91)	0.51 (0.41-0.62)
	CHA_2DS_2VASc	0.65	0.58	-	0.65	0.67	0.67	0.81	0.55

		(0.62 -0.67)	(0.56 -0.59)				(0.63-0.70)	(0.71-0.90)	(0.44-0.65)
	Framingham	0.63 (0.61 -0.65)	0.56 (0.55 -0.58)	-	0.63	0.64	0.63 (0.59-0.67)	0.67 (0.52-0.82)	0.52 (0.43-0.61)
	Q-Stroke	0.62 (0.60 -0.64)	0.55 (0.53 -0.56)	-	0.58	0.6	0.58 (0.54-0.62)	0.68 (0.42-0.94)	-

Table 3 Discrimination performance of the existing models across the datasets

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.docx