ESC
European Society
of Cardiology

European Heart Journal - Quality of Care and Clinical Outcomes (2023) **9**, 16–23
https://doi.org/10.1093/ehjqcco/qcac013

ORIGINAL ARTICLE

Machine learning-based identification of risk-factor signatures for undiagnosed atrial fibrillation in primary prevention and post-stroke in clinical practice

Renate B. Schnabel[1,2,*], Henning Witt[3], Jochen Walker[4], Marion Ludwig[4], Bastian Geelhoed[1,2], Nils Kossack[5], Marie Schild[3], Robert Miller[3,6] and Paulus Kirchhof[1,2,7]

[1]Department of Cardiology, University Heart and Vascular Center Hamburg, Martinistraße 52, 20251 Hamburg, Germany; [2]DZHK (German Center for Cardiovascular Research), partner site Hamburg/Kiel/Luebeck, Germany; [3]Pfizer Pharma GmbH, Linkstraße 10, 10785 Berlin, Germany; [4]InGef - Institute for Applied Health Research Berlin GmbH, Spittelmarkt 12, 10117 Berlin, Germany; [5]WIG2 GmbH, Markt 8, 04109 Leipzig, Germany; [6]Faculty of Psychology, Technische Universität Dresden, Dresden, Germany; and [7]Institute of Cardiovascular Sciences, College of Medical and Dental Sciences, Medical School, University of Birmingham, Edgbaston, Birmingham UK

Received 18 November 2021; revised 24 January 2022; accepted 8 March 2022; online publish-ahead-of-print 18 April 2022

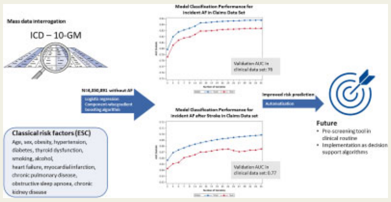
Aims Atrial fibrillation (AF) carries a substantial risk of ischemic stroke and other complications, and estimates suggest that over a third of cases remain undiagnosed. AF detection is particularly pressing in stroke survivors. To tailor AF screening efforts, we explored German health claims data for routinely available predictors of incident AF in primary care and post-stroke using machine learning methods.

Methods and results We combined AF predictors in patients over 45 years of age using claims data in the InGef database (n = 1 476 391) for (i) incident AF and (ii) AF post-stroke, using machine learning techniques. Between 2013–2016, new-onset AF was diagnosed in 98 958 patients (6.7%). Published risk factors for AF including male sex, hypertension, heart failure, valvular heart disease, and chronic kidney disease were confirmed. Component-wise gradient boosting identified additional predictors for AF from ICD-codes available in ambulatory care. The area under the curve (AUC) of the final, condensed model consisting of 13 predictors, was 0.829 (95% confidence interval (CI) 0.826–0.833) in the internal validation, and 0.755 (95% CI 0.603–0.890) in a prospective validation cohort (n = 661). The AUC for post-stroke AF was of 0.67 (95% CI 0.651–0.689) in the internal validation data set, and 0.766 (95% CI 0.731–0.800) in the prospective clinical cohort.

Conclusion ICD-coded clinical variables selected by machine learning can improve the identification of patients at risk of newly diagnosed AF. Using this readily available, automatically coded information can target AF screening efforts to identify high-risk populations in primary care and stroke survivors.

* Corresponding author. Tel: +49-1522-2816064, Fax: +49 (0)40 7410-55310, Email: r.schnabel@uke.de

Graphical Abstract Machine learning from health claims data robustly identifies patients at risk of AF in primary prevention and post stroke, which could be implemented in automated screening tools. The new algorithm is superior to classical risk factors available in the data set. ESC indicates European Society of Cardiology.


Mass data interrogation
ICD – 10-GM
N14,350,891 without AF
Logistic regression
Component-wise gradient boosting algorithm
Classical risk factors (ESC)
Age, sex, obesity, hypertension, diabetes, thyroid dysfunction, smoking, alcohol, heart failure, myocardial infarction, chronic pulmonary disease, obstructive sleep apnoea, chronic kidney disease
Model Classification Performance for Incident AF in Claims Data Set
Validation AUC in clinical data set: 78
Model Classification Performance for Incident AF after Stroke in Claims Data set
Validation AUC in clinical data set: 0.77
Number of Variables
Improved risk prediction
Automatization
Future
Pre-screening tool in clinical routine
Implementation as decision support algorithms

Keywords Atrial fibrillation • artificial intelligence • machine learning • risk prediction • stroke

Introduction

Projections in the European Union show that the number of adults >55 years of age with atrial fibrillation (AF) will double by 2060.[1,2] AF is associated with cardiovascular and cerebrovascular morbidities and mortality, including thromboembolic complications such as stroke, heart failure, cognitive decline and vascular dementia, and quantifiably impairments of quality of life.[3–5] AF onset is often asymptomatic, and 25–35% of the disease burden remains undiagnosed.[6,7] Many patients with AF are unaware of their disease, untreated, and at unnecessarily elevated risk of severe complications.[8,9] A stroke is often the first manifestation leading to diagnosis of AF, and at least one quarter of these strokes could have been prevented by earlier diagnosis of AF and initiation of anticoagulation.[10,11] An early AF diagnosis would furthermore enable earlier treatment of concomitant conditions, and consideration of rhythm control therapy with large potential benefits for public health.

Guidelines and experts recommend AF screening in primary prevention and post-stroke to enable timely diagnosis of AF.[12–14] To date, the 'gold-standard' for AF diagnosis is an electrocardiogram (ECG), however, mass screening with ECG is expensive and will require infrastructure to deal with incorrect diagnoses, particularly when the pre-test probability of having AF is low. Adapted and innovative approaches for cost-efficient screening programs are required.[12] To avoid the high costs of systematic screening and to reduce false-positive rates, the pre-selection of at-risk populations based on routine data is an attractive prospect. Traditional risk factor sets have been previously described,[12] but risk models that reliably predict AF in primary care and post-stroke remain sparse.[15] Supervised machine learning techniques enable automated selection of the most relevant risk factors from high-dimensional data while avoiding overfitting,[16] and thus enable constructing reliable risk factor-based screening algorithms.

The aim of our study was to identify a set of routinely available AF and stroke-related AF risk predictors, integrate these using predictive models and German claims data from over four million patients, and then show their discriminatory ability and generalizability in clinical cohorts.

Methods

Study design

We defined the following objectives: identify risk factors for (i) incident AF (first objective), and (ii) undiagnosed AF in patients with ischemic stroke (second objective). These risk factors were used to generate predictive screening models for early AF detection and undiagnosed AF in stroke patients. Then, we (iii) externally validated the performance of the developed models using data from an accurately phenotyped clinical cohort.

Training cohort

A retrospective cohort study using insurance claims data from the InGef research database was used to identify AF predictors. This database contains anonymized longitudinal data from approximately 60 German statutory health insurance providers (SHIs). A data set with 4 350 891 patients was extracted as a representative sample (regarding age and sex) of the German population between January 1st, 2010 and December 31st, 2016. Data from patients, healthcare providers, and the corresponding SHI are anonymized. The InGef database includes information on in- and out-patient treatment, prescribed and dispensed medications, sick leave and benefits, prescribed and dispensed medical devices and therapies, as well as demographic information such as age, sex, and location.[17]

The observation period of 01.01.2013 to 31.12.2015, was used to identify the individual index date (the first date of documented AF or stroke). Patients without AF or stroke were given a random pseudo index date during the observation period.[18]

To exclude patients with AF or stroke prior to the index date, a baseline period was defined. Due to the way the health claims data were recorded, 1095 days pre-index date was defined for inpatient diagnoses (documented by date), and 12 quarters prior to the quarter containing the index date was defined for outpatient diagnoses (documented by quarter). We split our data set into two parts: data from 01.01.2013–01.12.2015 were used to derive the models, while data from 01.01.2016–31.12.2016 were used as an internal validation set to assess the models' predictive accuracy.

We included patients 45 years of age or older at index date, since AF is an age-related disease. Patients were assigned to one of four study groups, based on the first and second objectives. Each patient was included only once and grouped according to first index event. Inclusion and exclusion criteria are provided in Table S1. AF diagnosis was identified by the ICD-10-GM codes I48.0, I48.1, I48.2, or I48.9 (as valid in 2013). To identify stroke diagnosis, we used the ICD-10-GM code I63*.

Validation cohort

Finally, we applied the screening models in a prospective cohort of consecutive ambulatory patients with cardiovascular risk factors older than 18 years. Patients' demographics, health status, cardiovascular risk factors, medical history, and environmental factors were collected during enrolment using a standardized interview, medical records, and a 12-lead ECG. For external validation, we used the discharge ICD-10-GM codes of $n = 661$ patients. We also enrolled post-stroke patients ($n = 162$) for validation of AF prediction after stroke, undergoing the same data collection and receiving a 7-day rhythm monitoring (Novacor, R.Test Evolution 4). The study was approved by the local ethics committee and complied with the Declaration of Helsinki. Written informed consent was collected from all study participants.

Incident AF during one-year follow-up was diagnosed, if at least two physicians trained in ECG reading confirmed the rhythm abnormality after enrolment using information from the clinic visit, outside physician, hospital records, ambulatory ECG reports, or the 7-day rhythm monitoring in post-stroke patients. Prevalent AF was diagnosed if the participant reported having AF, was treated as an in- or out-patient for AF, or medical records contained an AF diagnosis.

Statistical methods for identifying and validating risk factors

Risk factors for incident and previously undetected AF were identified based on patients' demographics, verified ambulatory diagnoses, primary or secondary hospital discharge diagnosis, and drugs dispensed during the baseline period. Two different sets of risk factors were used. The first set of risk factors contained known risk factors as published in the ESC guidelines, classical risk factors model (Table S4).[12] Their association with incident and unrecognized AF was assessed by fitting a logistic regression model. Then, novel risk factors were determined using three-digit ICD-10-GM codes, and four- or five-digit ATC (anatomical, therapeutic, chemical classification) codes, healthcare claims data-based risk model. The predictive value of these additional risk factors for both outcomes was optimized by using a component-wise gradient boosting algorithm. As a base learner, we used a linear model with a step size of 0.05. The number of iterations was tuned using bootstrap technique, with a minimal Akaike Information Criterion (AIC) as a stop criterion.[19] A certain number of features (i.e. ICD-codes or ATC-codes) had to be present in both groups for them to be considered for boosting. In addition, significance and medical plausibility were considered. Accordingly, the final selection of risk factors was based on both the background medical knowledge on the relationship of the variable with outcome and their statistical associations.

In order to avoid overfitting and to ensure variable selection, the algorithm was programmed to stop before convergence ("early stopping").

We assessed model performance by the total area under the receiver operating characteristic (ROC) curve (AUC). A binary classification analysis was used to compare observed and expected prediction of both respondents (AF diagnosis) and non-respondents, in the sample. This process provided sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). The optimum cut-off point was chosen for best trade-off between sensitivity and specificity by Youden's Index.[20]

Two sensitivity analyses were performed: the first included patients at least 65 years of age at index date, and the second excluded patients with an oral anticoagulant (OAC) prescription in the baseline period. Both analyses were based on a model that used all available information in the InGef research database to define risk factors.

Baseline characteristics were analyzed using demographic (age, sex) and clinical (comorbidities, hospitalizations, medication) characteristics of the patients in all treatment groups. Variables were derived from the InGef research database for the respective baseline period. Table S7 provides an overview of all variables assessed for the description of the study population, and for the inclusion in the multivariable models. The data set contained information on age, sex, and ICD-10-GM codes in ambulatory care, since more data is available in the ambulatory setting and reflects primary care. For assessment of the goodness of fit of our models, we calculated the Hosmer–Lemeshow test.[21]

We further assessed the AUC for two frequently used risk prediction scores in AF, the $CHADS_2$ (one point each for heart failure, hypertension, age ≥75 years, diabetes mellitus, 2 points for prior stroke/transient ischemic attack)[22] and HATCH (hypertension [1 point], age ≥75 years [1 point], transient ischemic attack or stroke [2 points], chronic obstructive pulmonary disease [1 point], and heart failure [2points]) scores.[23]

A *P*-value of $p < 0.01$ was considered significant. Data analysis was carried out by InGef; data management and statistical analyses were performed using SAS 9.3 (SAS Institute Inc.) and R 3.4.1.

Results

Participants

A total of 1 476 391 patients met the inclusion criteria for objective one (Table S2), and 98 958 incident AF patients were identified in the InGef database from 2013–2016. After applying the inclusion criteria, 88 111 AF patients were suitable for analysis.

In the 2010–2016 observation period, 29 155 patients had an incident ischemic stroke without AF, and 19 019 fulfilled the inclusion criteria and remained for analysis (Table S3). In the same period, 4 653 patients were diagnosed with AF post-stroke.

Baseline Characteristics

Baseline demographic and medical characteristics of the study groups obtained from the InGef research database and for the validation cohort are shown in *Table 1*.

Incident AF patients were older (74.7 vs. 61.1 years) and were more likely men. Comorbidities such as diabetes mellitus, hypertension, valvular heart disease, heart failure, or chronic kidney disease, occurred more frequently in the AF-group than in the non-AF group. The distribution of the modifiable risk factors of alcohol consumption or tobacco use was similar between groups.

Patients who had a stroke followed by a newly diagnosed AF were on average 6 years older than patients without a subsequent AF diagnosis (77.3 vs. 71.2 years). They were more likely women, and more often had cardiovascular comorbidities (except for myocardial infarction).

Known ESC risk predictors

All classical risk predictors were statistically significantly associated with incident AF, with age as the strongest predictor. Men were more likely to develop AF, as well as patients with cardiovascular

Table 1 Baseline characteristics of the study population (for model derivation)

	Incident AF		Post-stroke		External validation	
Characteristics	**No AF**	**AF**	**No AF**	**AF**	**No stroke**	**Post stroke**
	***n* = 1 115 485**	***n* = 66 697**	***n* = 14 001**	***n* = 3419**	***n* = 661**	**N = 162**
Demographics						
Age (mean±SD)	61.1±11.5	74.7±11.5	71.2±11.9	77.3±10.0	65±10	63±11
Age <65 (%)	64.6	18.0	30.2	11.7	44.6	53.3
Age 65-74 (%)	19.8	25.7	25.4	21.8	36.6	31.1
Age >74 (%)	15.6	56.4	44.5	66.5	18.8	15.6
Women (%)	51.9	46.0	43.7	52.1	34.6	36.5
Underlying diseases						
Hypertension (%)	49.3	81.8	73.9	83.1	66.6	60.5
Myocardial infarction (%)	1.2	3.8	2.7	2.7	13.9	12.6
Heart failure (%)	4.7	21.3	12.5	20.7	32.7	3.0
Valvular heart disease	5.8	18.6	10.5	15.7	10.7	11.4
Thyroid dysfunction (%)	24.6	26.7	23.0	23.9	23.1	18.6
Obesity (%)	16.0	23.5	18.2	18.5	25.6	22.2
Diabetes mellitus (%)	15.9	34.7	33.9	35.9	13.2	13.2
Chronic obstructive pulmonary disease (%)	8.1	16.7	13.1	14.2	12.0	7.8
Sleep apnoea (%)	3.8	6.0	4.6	4.1	11.2	5.4
Chronic kidney disease (%)	3.3	12.1	9.4	10.6	11.5	3.0
Tobacco use (%)	8.2	7.2	10.7	6.5	12.4	33.5
Alcohol consumption (%)	2.1	2.8	4.1	2.8	5.6	10.8
Stroke (%)	0.5	4.0	3.8	3.8	0.0	100
Rheumatoid arthritis (%)	2.9	4.6	3.9	4.1	0.8	0.0
Atherosclerosis (%)	6.0	16.1	14.4	15.7	3.3	68.9
Other cardiac arrhythmias (%)	7.1	23.2	9.7	17.3	8.2	2.4
Hospitalization within 2 quarters prior to diagnosis (%)	11.7	25.9	22.6	19.8	n.a.	n.a.

The mean age±standard deviation and the percentage for categorical variables are provided.
AF stands for atrial fibrillation, n.a. indicates not available.

conditions. Patients with obesity, diabetes, COPD, kidney disease, sleep apnoea, tobacco use, or alcohol consumption developed AF more often (Table S5). The AUC of the model was 0.804 (95% CI 0.802–0.806), with the optimal cut-off value according to Youden Index being 0.0644, leading to a sensitivity of 77% and a specificity of 70%. A cut-off value of 0.0197 gave a sensitivity of 95% and a specificity of 35%.

In the post-stroke model, age was the strongest predictor of post-stroke AF. Men were less likely to be diagnosed with AF post-stroke. The presence of cardiovascular conditions, except for myocardial infarction, significantly increased the risk of incident stroke followed by AF. Myocardial infarction, diabetes mellitus, chronic kidney disease, and tobacco use were negatively associated with AF diagnosed after incident stroke (Table S5). The combined ESC risk factors achieved an AUC of 0.658 (95% CI 0.647–0.669), with an optimal cut-off value of 0.173, a sensitivity of 77%, and a specificity of 46%. A cut-off value of 0.0905 resulted in a sensitivity of 95% and a specificity of 18%.

Sensitivity analyses

In the first sensitivity analysis of patients aged 65 years or older, age showed a lower magnitude of association, but remained the strongest predictor. The AUC reached 0.713. The second sensitivity analyses excluded patients with an oral anticoagulant (OAC) prescription in the baseline period; the absence of an OAC prescription did not markedly changed the association between the pre-defined risk factors and AF occurrence.

Predictors for incident AF

The boosting procedure led to an initial model of 43 predictors, among which the classic ESC risk factors showed the strongest associations with incident AF. Novel predictors consisted of cardiovascular conditions that strongly increased risk of incident AF. Patients with hemiplegia or paroxysmal tachycardia were 3.04 (95% confidence interval (CI) 2.86–3.23) and 2.20 (95% CI 2.11–2.30) times more likely to be diagnosed with AF, respectively. Diseases of the digestive system were negatively associated with AF (e.g. having functional dyspepsia lowered the risk of AF by about 10%). The final reduced model consisted of 13 variables that enhanced predictive accuracy (*Table 2*). The classification performance of the final reduced model is shown in parallel for the derivation and internal validation data set (*Figure 1*). The ROC-curve for the novel risk factor model provided an AUC of 0.829 (95% confidence interval (CI) 0.826–0.833), using the validation data set is shown in Figure S1. Both, the

$CHADS_2$ and HATCH scores, achieved an AUC of 0.779. According to the Youden Index, the optimal cut-off value was 0.0526, leading to a sensitivity of 80% and a specificity of 72%. A cut-off value of 0.0155 resulted in a sensitivity of 95% and a specificity of 42%. The PPV and NPV were 0.13 and 0.98, respectively. The Brier Score was 0.045 in the validation data. The Hosmer–Lemeshow test *P*-value in the validation data set was <0.0001. The healthcare claims data-based risk model achieved an AUC of 0.755 (95% CI 0.603–0.890) in the prospective, external validation cohort. The model based on classical risk factors according to ESC achieved an AUC of 0.734 (95% CI 0.697–0.770) ($n = 661$). Compared to prior scores.

Predictors for post-stroke AF

Using gradient boosting, the initial model for AF after stroke included 36 variables. In addition to classical cardiovascular risk factors, the presence of paroxysmal tachycardia, pulmonary heart diseases, or other cardiac arrhythmias significantly increased the risk of post-stroke AF. Pre-existing diseases of the nervous system (such as disorders of the trigeminal nerve or spinal cord), and presence of nutritional and metabolic diseases were inversely associated with AF after stroke. The final reduced model was based on 13 variables (*Table 3*). The AUC achieved a value of 0.670 (95% CI 0.651–0.689) in the internal validation data set (Figure S2), 0.766 (95% CI 0.731–0.80) in the prospective clinical cohort. According to the Youden Index, the optimal cut-off value was 0.1685, leading to a sensitivity of 76% and a specificity of 50%. The PPV and NPV were 0.25 and 0.91, respectively. The Brier Score was 0.138 in the validation data. The Hosmer–Lemeshow test *P*-value in the post stroke validation data set was 0.929. The new predictive model achieved a sensitivity of 95% and a specificity of 18%, with a cut-off value of 0.0727.

Discussion

For the prediction of incident AF in primary care and post-ischemic stroke, machine learning techniques confirmed known risk factors for AF and identified novel risk predictors, such as right heart disease, other cerebrovascular diseases, and breathing abnormalities. Other comorbidities were associated with an overall reduced likelihood of newly diagnosed post-stroke AF, such as intestinal diverticular disease, and orthopedic problems of the knee or dorsalgia. An AF risk model based on these factors available in large administrative databases predicted incident AF well and could complement information derived from classical cardiovascular risk factors. Using known and novel predictors, our data further sets the stage for predicting post-stroke AF, and possibly refined screening (Graphical Abstract).

Risk factors were identified in two different settings, each representing patients at sufficient risk of undetected AF that merit screening. Screening in older patients, and after the occurrence of stroke, has been recommended.[12,13] The current standard of AF detection using ECG is not feasible for screening on a large scale, and selection of patients for extended rhythm monitoring has remained inconclusive. Traditional sets of risk factors and medical scoring schemes such as the CHARGE-AF score are limited in their predictive performance and often require electrocardiographic parameters or laboratory measurements.[24] Refinement and simplification is required for optimal and efficient patient management,[25] and can be achieved by automated processing of readily available routine data.

Our results in a claims data set of over one million patients indicate that the most important published risk factors for incident AF are reflected by the corresponding ICD-10-GM codes available in ambulatory care. The associations are consistent with epidemiolog-

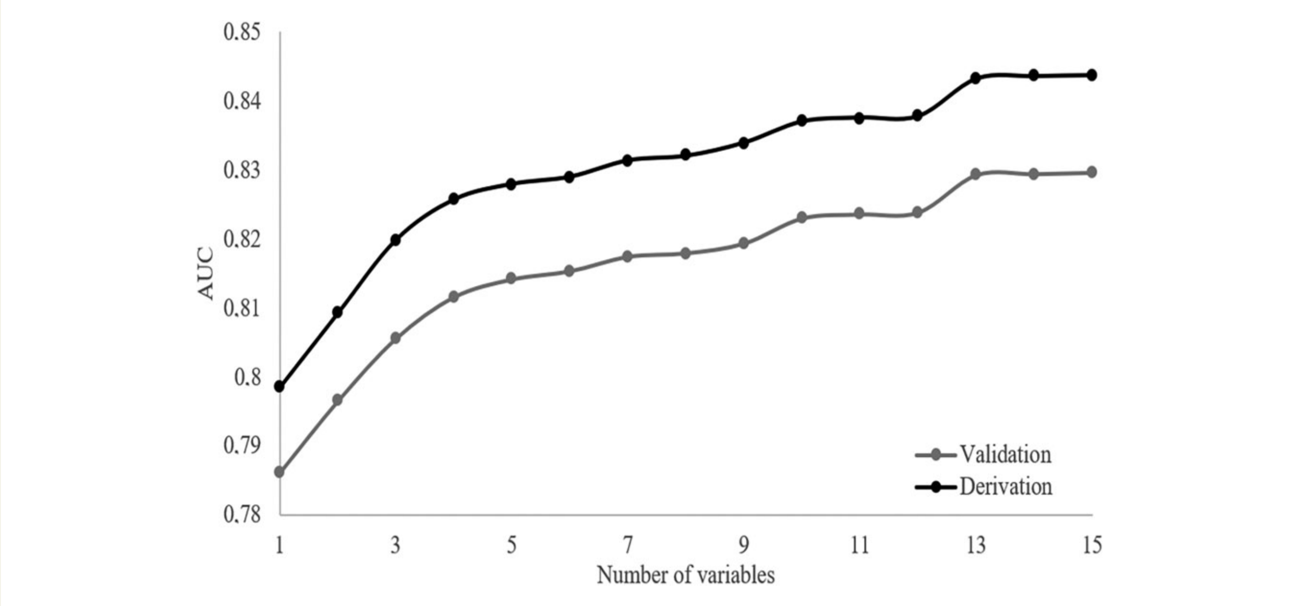


Figure 1 Classification performance (area under the receiver operating characteristic curve; AUC) of a model consisting of risk factors for incident AF according to complexity (number of variables). The order of the included variables is age, male sex, hypertension, treated, heart failure, valvular heart disease, chronic kidney disease, stroke not specified as hemorrhage or infarction, hemiplegia, other pulmonary heart diseases, paroxysmal tachycardia, other cardiac arrhythmias, ulcer of lower limb, not elsewhere classified, personal history of medical treatment.

Table 2 Covariable-adjusted odds ratios for incident atrial fibrillation in the final model derived from boosting technique

Predictor	OR (95% CI)	*P*-value
Age 45–49	0.08 (0.08–0.09)	<0.0001
Age 50–54	0.13 (0.12–0.13)	<0.0001
Age 55–59	0.19 (0.18–0.20)	<0.0001
Age 60–64	0.30 (0.29–0.31)	<0.0001
Age 65–69	0.46 (0.45–0.48)	<0.0001
Age 70–74	0.71 (0.69–0.73)	<0.0001
Age 80–84	1.45 (1.41–1.49)	<0.0001
Age 85–89	1.91 (1.85–1.97)	<0.0001
Age 90+	2.24 (2.15–2.34)	<0.0001
Men	1.52 (1.49–1.54)	<0.0001
Hypertension, treated	1.76 (1.72–1.79)	<0.0001
Heart failure, treated	1.54 (1.50–1.58)	<0.0001
Valvular heart disease	1.42 (1.39–1.46)	<0.0001
Chronic kidney disease	1.21 (1.18–1.24)	<0.0001
Stroke, not specified as hemorrhage or infarction	2.43 (2.29–2.57)	<0.0001
Hemiplegia	3.04 (2.86–3.23)	<0.0001
Other pulmonary heart diseases	1.60 (1.51–1.69)	<0.0001
Paroxysmal tachycardia	2.20 (2.11–2.30)	<0.0001
Other cardiac arrhythmias	2.11 (2.07–2.16)	<0.0001
Ulcer of lower limb, not elsewhere classified	1.65 (1.56–1.73)	<0.0001
Personal history of medical treatment	1.62 (1.58–1.65)	<0.0001

Odds ratios are provided with 95% confidence intervals and *P*-values associated with incident AF derived from boosting technique.

Table 3 Covariable-adjusted odds ratios for post stroke AF in the final model derived from boosting technique

Predictor	OR (95% CI)	*P*-value
Age 45–49	0.19 (0.13–0.28)	<0.0001
Age 50–54	0.25 (0.19–0.33)	<0.0001
Age 55–59	0.23 (0.18–0.30)	<0.0001
Age 60–64	0.43 (0.36–0.51)	<0.0001
Age 65–69	0.52 (0.44–0.62)	<0.0001
Age 70–74	0.82 (0.72–0.94)	0.0043
Age 80–84	1.20 (1.06–1.36)	0.0044
Age 85–89	1.34 (1.17–1.54)	<0.0001
Age 90+	1.49 (1.26–1.77)	<0.0001
Men	0.85 (0.78–0.92)	<0.0001
Hypertension, treated	1.32 (1.19–1.47)	<0.0001
Heart failure, treated	1.25 (1.11–1.40)	0.0002
Chronic kidney disease	0.79 (0.70–0.89)	<0.0001
Disorders of lipoprotein metabolism and other lipidaemias	0.84 (0.78–0.92)	<0.0001
Pulmonary heart diseases	1.83 (1.41–2.38)	<0.0001
Cardiac arrhythmias	1.68 (1.51–1.88)	<0.0001
Other cerebrovascular diseases	0.74 (0.65–0.84)	<0.0001
Diverticular disease of intestine	0.76 (0.67–0.86)	<0.0001
Internal derangement of knee	0.65 (0.53–0.80)	<0.0001
Dorsalgia	0.81 (0.75–0.88)	<0.0001
Breathing abnormalities	0.77 (0.68–0.89)	0.0002

Odds ratios are provided with 95% confidence intervals.

ical data and previous risk prediction models for AF, and validate the concept that comorbidities determined from a claims database may provide reasonably accurate risk prediction.[20,26] Our results suggest that age and male sex are the most important risk factors for AF. As shown in previous publications, the presence of cardiovascular conditions (e.g. hypertension, heart failure), was positively associated with incident AF.[27–29] By using machine learning techniques, we identified additional predictors for incident AF, such as pulmonary heart disease and ulcers of the lower limb, which reflect high comorbidity levels in older patients with AF. Prior stroke and hemiplegia also strongly contributed to the prediction of incident AF. Although the additional predictive value provided by each of these newly identified variables was small, the overall AUC in the claims data validation set improved the classification performance substantially, with an AUC of 0.829. The AUC was comparable to prior published scores such as the CHADS$_2$ and HATCH scores.[22,23] Modelling the risk of incident AF, we found that variables routinely collected in primary care are sufficient to reliably predict onset of AF.

Almost 20% of patients with stroke had a subsequent diagnosis of AF. These data are in line with smaller, earlier studies summarized in meta-analyses, that indicated AF incidence post-stroke was 11–24%.[11] These numbers highlight the relevance of AF detected in routine care in an unselected stroke cohort. An AF diagnosis often affects treatment course, in particular the initiation of OAC.[30,31] More intensive screening for AF post-stroke might have increased the number of patients diagnosed with AF to up to 24%.[11] To date, a systematic screening for AF, for example with a 24h Holter ECG, is not routine in stroke patients.[32] Our study underlines the importance of post-stroke AF and the potential relevance of screening, as selection indicators of patients for more intense post-stroke AF screening are largely unknown.[12]

We were able to predict post-stroke AF incidence with fair accuracy: besides advanced age and classic cardiovascular risk factors, prevalent cardiovascular diseases such as heart failure, structural heart defects and cardiac arrhythmias were strong predictors of post-stroke AF. Our findings extend the current knowledge of the association of heart disease with newly detected AF post-stroke.[33] In patients with pre-existing disorders of the nervous system other than stroke, including trigeminal nerve or spinal cord diseases, post-stroke AF incidence was less likely. In accordance with our study, an increased risk of post-stroke AF in women compared to men has been observed in a Swedish nationwide registry and has been explained by the higher average age in women with stroke.[34] The unexpected inverse contribution of chronic kidney disease and dyslipidemia compared to primary prevention may indicate that patients with these diagnoses receive more clinical attention and AF may have been diagnosed earlier, i.e. prior to stroke.

The combination of these predictors demonstrated fair sensitivity in identifying patients who developed post-stroke AF and may help flag patients for more intensive AF screening, making targeted screening more efficient and practicable.

Both final models, in primary prevention and post-stroke, were successfully replicated in contemporary clinical cohorts in the ambulatory setting and suggest the external validity of our findings.

Our study demonstrates that machine learning techniques can accurately identify patients at increased risk of AF based on readily available routine data. In the future, adaptive algorithms could be implemented as primary care decision support tools to promote opportunistic AF screening, or as electronic decision support for effective post-stroke AF screening.[35]

Limitations

Our findings must be interpreted in light of some limitations. Although the InGef research database covers more than four million insured members of SHIs across Germany, representativeness cannot be guaranteed. However, the very good reproducibility of the classical risk factors in the claims data set and the validation in independently phenotyped clinical cohorts indicates that the models' performances are robust and are likely generalizable. Since a further limitation is that the validation sample is a clinical cohort of cardiovascular patients, good discriminatory ability in this independent cohort shows the strength of the initial model. AF and stroke are correlated with geographical, ethnic, and socioeconomic factors not sufficiently reflected in the database.[36] Some risk indicators that have consistently been related to AF, such as natriuretic peptides,[37] glomerular filtration rate,[38] and electrocardiographic alterations are not available in the claims database. On the other hand, our models are based on relatively uniform information available in practice, which can automatically be extracted from the health records with little additional cost. Thus, they may serve as a benchmark for implementing targeted AF screening of increasing clinical significance due to the substantial increases in AF prevalence.

Known clinical conditions and cardiovascular risk factors can be reproduced in claims data and jointly predict of incident AF, also in the post-stroke setting. The prediction performance can be improved by adding novel clinical variables, identified by machine learning. The clinical validation of the described novel set of AF risk predictors indicates that incorporating easily available and broad information on underlying comorbidities strongly enhances the prediction of AF-onset. Our risk-factor model relies on readily available data, and according decision support systems could be implemented as a pre-screening tool in primary care and post-stroke clinical routine.

Supplementary material

Supplementary material is available at *European Heart Journal—Quality of Care and Clinical Outcomes* online.

Acknowledgements

We thank all study participants of the clinical cohorts for their willingness to provide data for this research project and we are indebted to all coworkers for their enthusiastic commitment. Daniel Engler, MSc provided editorial support for the submission.

Conflicts of interest: This research was funded by Pfizer Pharma GmbH. The authors had complete autonomy for the process of establishing the protocol, carrying out the analyses and interpreting the results, including full publishing rights to the results, without limitation. RBS discloses consulting and lecture fees from BMS/Pfizer (unrelated to the present study). RBS and PK are members of Kompetenznetzwerk Vorhofflimmern e.V., which is a paid consultant to Pfizer in connection with the development of the manuscript. RBS and PK did not receive any financial compensation for the data interpretation and development of the manuscript from Kompetenznetzwerk Vorhofflimmern e.V. PK receives research support for basic, translational, and clinical research projects from European Union, British Heart Foundation, Leducq Foundation, Medical Research Council (UK), and German Centre for Cardiovascular Research, from several drug and device companies active in atrial fibrillation, and has received honoraria from several such companies in the past, but not in the last three years. PK is listed as inventor on two patents held by University of Birmingham (Atrial Fibrillation Therapy WO 2015140571, Markers for Atrial Fibrillation WO 2016012783). HW, MS, and RM are employees of Pfizer. ML and JW are employees of InGef - Institute for Applied Health Research Berlin GmbH, which executes data analysis and manuscript preparation for WIG2 GmbH. NK is an employee of WIG2 GmbH, which is a paid consultant to Pfizer.

Funding

Pfizer Pharma GmbH. RBS has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 648131); the German Center for Cardiovascular Research (DZHK) (81Z1710103); the European Union's Horizon 2020 research and innovation program (grant agreement no. 847770, AFFECT-EU); and ERACoSysMed3 (031L0239).

Declaration of Helsinki

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional ethics committee (PV5705) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data availability statement

The data used in this study cannot be made available in the manuscript, the supplemental files, or in a public repository due to German data protection laws (Bundesdatenschutzgesetz). To facilitate the replication of results, anonymized data used for this study are stored on a secure drive at the Institute for Applied Health Research Berlin (InGef). Access to the raw data used in this study can only be provided to external parties under the conditions a cooperation contract and can be accessed upon request, after written approval (info@ingef.de), if required.

References

1. Krijthe BP, Kunst A, Benjamin EJ, Lip GY, Franco OH, Hofman A, Witteman JC, Stricker BH, Heeringa J. Projections on the number of individuals with atrial

fibrillation in the European Union, from 2000 to 2060. *Eur Heart J* 2013;**34**: 2746–2751.

2. Schnabel RB, Yin X, Gona P, Larson MG, Beiser AS, McManus DD, Newton-Cheh C, Lubitz SA, Magnani JW, Ellinor PT, Seshadri S, Wolf PA, Vasan RS, Benjamin EJ, Levy D. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. *Lancet* 2015;**386**:154–162.
3. Ball J, Carrington MJ, McMurray JJ, Stewart S. Atrial fibrillation: profile and burden of an evolving epidemic in the 21st century. *Int J Cardiol* 2013;**167**:1807–1824.
4. Ball J, Carrington MJ, Stewart S. Mild cognitive impairment in high-risk patients with chronic atrial fibrillation: a forgotten component of clinical management? *Heart* 2013;**99**:542–547.
5. Dietzel J, Haeusler KG, Endres M. Does atrial fibrillation cause cognitive decline and dementia? *Europace* 2018;**20**:408–419.
6. Svennberg E, Engdahl J, Al-Khalili F, Friberg L, Frykman V, Rosenqvist M. Mass Screening for Untreated Atrial Fibrillation: The STROKESTOP Study. *Circulation* 2015;**131**:2176–2184.
7. Rho RW, Page RL. Asymptomatic atrial fibrillation. *Prog Cardiovasc Dis* 2005;**48**: 79–87.
8. Lowres N, Neubeck L, Redfern J, Freedman SB. Screening to identify unknown atrial fibrillation. A systematic review. *Thromb Haemostasis* 2013;**110**:213–222.
9. Samol A, Masin M, Gellner R, Otte B, Pavenstädt HJ, Ringelstein EB, Reinecke H, Waltenberger J, Kirchhof P. Prevalence of unknown atrial fibrillation in patients with risk factors. *Europace* 2013;**15**:657–662.
10. Lin HJ, Wolf PA, Benjamin EJ, Belanger AJ, D'Agostino RB. Newly diagnosed atrial fibrillation and acute stroke. The Framingham Study. *Stroke* 1995;**26**:1527–1530.
11. Sposato LA, Cipriano LE, Saposnik G, Ruíz Vargas E, Riccio PM, Hachinski V. Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis. *Lancet Neurol* 2015;**14**:377–387.
12. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener HC, Heidbuchel H, Hendriks J, Hindricks G, Manolis AS, Oldgren J, Popescu BA, Schotten U, Van Putte B, Vardas P. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2016;**37**:2893–2962.
13. 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 L, 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, Yan BP. Screening for Atrial Fibrillation: a Report of the AF-SCREEN International Collaboration. *Circulation* 2017;**135**:1851–1867.
14. Schnabel RB, Haeusler KG, Healey JS, Freedman B, Boriani G, Brachmann J, Brandes A, Bustamante A, Casadei B, Crijns H, Doehner W, Engström G, Fauchier L, Friberg L, Gladstone DJ, Glotzer TV, Goto S, Hankey GJ, Harbison JA, Hobbs FDR, Johnson LSB, Kamel H, Kirchhof P, Korompoki E, Krieger DW, Lip GYH, Løchen ML, Mairesse GH, Montaner J, Neubeck L, Ntaios G, Piccini JP, Potpara TS, Quinn TJ, Reiffel JA, Ribeiro ALP, Rienstra M, Rosenqvist M, Themistoclakis S, Sinner MF, Svendsen JH, Van Gelder IC, Wachter R, Wijeratne T, Yan B. Searching for Atrial Fibrillation Poststroke: a White Paper of the AF-SCREEN International Collaboration. *Circulation* 2019;**140**:1834–1850.
15. Linker DT, Murphy TB, Mokdad AH. Selective screening for atrial fibrillation using multivariable risk models. *Heart* 2018;**104**:1492–1499.
16. James GM, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. NY: Springer-Verlag New York; 2013.
17. Andersohn F, Walker J. Characteristics and external validity of the German Health Risk Institute (HRI) Database. *Pharmacoepidemiol Drug Saf* 2016;**25**:106–109.
18. Jacob J, Schmedt N, Hickstein L, Galetzka W, Walker J, Enders D. Comparison of approaches to select a propensity score matched control group in the absence of an obvious start of follow up for this group: an example study on the economic impact of the DMP Bronchial Asthma. *Gesundheitswesen* 2020;**82**:S151–s157.
19. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974;**19**:716–723.
20. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;**3**:32–35.
21. Hosmer DW, Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*: John Wiley & Sons; 2013.
22. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA* 2001;**285**:2864–2870.
23. de Vos CB, Pisters R, Nieuwlaat R, Prins MH, Tieleman RG, Coelen RJ, van den Heijkant AC, Allessie MA, Crijns HJ. Progression from paroxysmal to persistent atrial fibrillation clinical correlates and prognosis. *J Am Coll Cardiol* 2010;**55**: 725–731.
24. 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, Kääb S, Couper D, Harris TB, Soliman EZ, Stricker BH, Gudnason V, Heckbert SR, Benjamin EJ. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc* 2013;**2**:e000102.
25. Fabritz L, Crijns H, Guasch E, Goette A, Haeusler K, Kotecha D, Lewalter T, Meyer C, Potpara T, Rienstra M, Schnabel R, WIllems S, Breithardt G, Camm A, Chan A, Chua W, de Melis M, Dimopoulou C, Dobrev D, Easter C, Eckardt L, Haase D, Hatem S, Healey J, Heijman J, Hohnloser S, Huebner T, Ilyas B, Isaacs A, Kutschka I, Leclerq C, Lip G, Marinelli E, Merino J, Mont L, Nabauer M, Oldgren J, Purerfellner H, Ravens U, Savelieva I, Sinner M, Sitch A, Smolnik R, Steffel J, Stein K, Stoll M, Svennberg E, Thomas D, van Gelder I, Vardar B, Wakili R, Wieloch M, Zeemering S, Ziegler P, Heidbuchel H, Hindricks G, Schotten U, Kirchhof P. Dynamic risk assessment to improve quality of care in patients with atrial fibrillation: The 7th AFNET/EHRA Consensus Conference. *Europace*, in press 2020.
26. Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, Ambrose M, Alonso A. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *Am J Cardiol* 2011;**107**:85–91.
27. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., Newton-Cheh C, Yamamoto JF, Magnani JW, Tadros TM, Kannel WB, Wang TJ, Ellinor PT, Wolf PA, Vasan RS, Benjamin EJ. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet (London, England)* 2009;**373**:739–745.
28. Chambless LE, Heiss G, Shahar E, Earp MJ, Toole J. Prediction of ischemic stroke risk in the Atherosclerosis Risk in Communities Study. *Am J Epidemiol* 2004;**160**: 259–269.
29. Furberg CD, Psaty BM, Manolio TA, Gardin JM, Smith VE, Rautaharju PM. Prevalence of atrial fibrillation in elderly subjects (the Cardiovascular Health Study). *Am J Cardiol* 1994;**74**:236–241.
30. Edwards JD, Kapral MK, Fang J, Saposnik G, Gladstone DJ. Underutilization of Ambulatory ECG monitoring after stroke and Transient Ischemic Attack: missed opportunities for Atrial Fibrillation detection. *Stroke* 2016;**47**: 1982–1989.
31. Sanna T, Diener HC, Passman RS, Di Lazzaro V, Bernstein RA, Morillo CA, Rymer MM, Thijs V, Rogers T, Beckers F, Lindborg K, Brachmann J. Cryptogenic stroke and underlying atrial fibrillation. *N Engl J Med* 2014;**370**:2478–2486.
32. Ntaios G, Papavasileiou V, Milionis H, Makaritsis K, Vemmou A, Koroboki E, Manios E, Spengos K, Michel P, Vemmos K. Embolic strokes of undetermined source in the athens stroke registry. *Stroke* 2015;**46**:2087–2093.
33. Rizos T, Horstmann S, Dittgen F, Täger T, Jenetzky E, Heuschmann P, Veltkamp R. Pre-existing heart disease underlies newly diagnosed Atrial Fibrillation after acute ischemic stroke. *Stroke* 2016;**47**:336–341.
34. Friberg L, Rosenqvist M, Lindgren A, Terént A, Norrving B, Asplund KJS. High prevalence of atrial fibrillation among patients with ischemic stroke. *Stroke* 2014;**45**:2599–2605.
35. Orchard J, Neubeck L, Freedman B, Li J, Webster R, Zwar N, Gallagher R, Ferguson C, Lowres N. eHealth tools to provide structured assistance for Atrial Fibrillation screening, management, and guideline-recommended therapy in metropolitan general practice: The AF - SMART Study. *J Am Heart Assoc* 2019;**8**:e010959.
36. Misialek JR, Rose KM, Everson-Rose SA, Soliman EZ, Clark CJ, Lopez FL, Alonso A. Socioeconomic status and the incidence of atrial fibrillation in whites and blacks: the Atherosclerosis Risk in Communities (ARIC) study. *J Am Heart Assoc* 2014;**3**:e001159.
37. Sinner MF, Stepas KA, Moser CB, Krijthe BP, Aspelund T, Sotoodehnia N, Fontes JD, Janssens AC, Kronmal RA, Magnani JW, Witteman JC, Chamberlain AM, Lubitz SA, Schnabel RB, Vasan RS, Wang TJ, Agarwal SK, McManus DD, Franco OH, Yin X, Larson MG, Burke GL, Launer LJ, Hofman A, Levy D, Gottdiener JS, Kääb S, Couper D, Harris TB, Astor BC, Ballantyne CM, Hoogeveen RC, Arai AE, Soliman EZ, Ellinor PT, Stricker BH, Gudnason V, Heckbert SR, Pencina MJ, Benjamin EJ, Alonso A. B-type natriuretic peptide and C-reactive protein in the prediction of atrial fibrillation risk: the CHARGE-AF Consortium of community-based cohort studies. *Europace* 2014;**16**:1426–1433.
38. Bansal N, Zelnick LR, Alonso A, Benjamin EJ, de Boer IH, Deo R, Katz R, Kestenbaum B, Mathew J, Robinson-Cohen C, Sarnak MJ, Shlipak MG, Sotoodehnia N, Young B, Heckbert SR. eGFR and Albuminuria in relation to risk of incident Atrial Fibrillation: a meta-analysis of the Jackson Heart Study, the Multi-Ethnic Study of Atherosclerosis, and the Cardiovascular Health Study. *Clin J Am Soc Nephrol* 2017;**12**:1386–1398.