ESC European Society of Cardiology

Europace (2020) **22**, 1635–1644
doi:10.1093/europace/euaa172

Machine learning does not improve upon traditional regression in predicting outcomes in atrial fibrillation: an analysis of the ORBIT-AF and GARFIELD-AF registries

Zak Loring [1,2]*, Suchit Mehrotra[3], Jonathan P. Piccini [1,2], John Camm[4], David Carlson[1], Gregg C. Fonarow [5], Keith A.A. Fox [6], Eric D. Peterson[1,2], Karen Pieper[1], and Ajay K. Kakkar[7,8]

[1]Duke Clinical Research Institute, Durham, NC, USA; [2]Division of Cardiology, Department of Medicine, Duke University Medical Center, 2301 Erwin Rd, DUMC 3845, Durham, NC 27710, USA; [3]Department of Statistics, North Carolina State University, Raleigh, NC, USA; [4]Cardiology Clinical Academic Group, St. George's University of London, London, UK; [5]Department of Medicine, UCLA Division of Cardiology, Los Angeles, CA, USA; [6]Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK; [7]Thrombosis Research Institute, London, UK; and [8]University College London, London, UK

Received 3 January 2020; editorial decision 23 May 2020; accepted after revision 26 May 2020; online publish-ahead-of-print 3 September 2020

Aims Prediction models for outcomes in atrial fibrillation (AF) are used to guide treatment. While regression models have been the analytic standard for prediction modelling, machine learning (ML) has been promoted as a potentially superior methodology. We compared the performance of ML and regression models in predicting outcomes in AF patients.

Methods and results The Outcomes Registry for Better Informed Treatment of Atrial Fibrillation (ORBIT-AF) and Global Anticoagulant Registry in the FIELD (GARFIELD-AF) are population-based registries that include 74 792 AF patients. Models were generated from potential predictors using stepwise logistic regression (STEP), random forests (RF), gradient boosting (GB), and two neural networks (NNs). Discriminatory power was highest for death [STEP area under the curve (AUC) = 0.80 in ORBIT-AF, 0.75 in GARFIELD-AF] and lowest for stroke in all models (STEP AUC = 0.67 in ORBIT-AF, 0.66 in GARFIELD-AF). The discriminatory power of the ML models was similar or lower than the STEP models for most outcomes. The GB model had a higher AUC than STEP for death in GARFIELD-AF (0.76 vs. 0.75), but only nominally, and both performed similarly in ORBIT-AF. The multilayer NN had the lowest discriminatory power for all outcomes. The calibration of the STEP modelswere more aligned with the observed events for all outcomes. In the cross-registry models, the discriminatory power of the ML models was similar or lower than the STEP for most cases.

Conclusion When developed from two large, community-based AF registries, ML techniques did not improve prediction modelling of death, major bleeding, or stroke.

Keywords Machine learning • Outcomes • Atrial fibrillation

Introduction

Stroke, bleeding, and death are important outcomes in patients with atrial fibrillation (AF) and treatment decisions are often dependent upon a given patient's risk for each ofthese outcomes.[1,2] While prediction models for these outcomes have improved over time, the discriminatory capacities of contemporary models are modest.[3,4] Despite their limitations, these risk models have become integral to patient care and stroke prevention therapy guidelines.[2,5]

* Corresponding author. Tel: +1 919 668 4649; fax: +1 919 681 9842. *E-mail address*: zak.loring@duke.edu

What's new?

- We compared the performance of machine learning (ML) and traditional regression models in predicting clinical outcomes using two large outpatient registries of more than 74 000 atrial fibrillation (AF) patients.
- The discrimination of the ML models was similar or worse than the stepwise regression models for nearly all outcomes.
- The stepwise regression models had better calibration than the ML models.
- In cross-registry validation, ML models performed as well or worse than stepwise regression in the majority of cases.
- When developed from two large, community-based AF registries, ML techniques did not improve prediction modelling of death, major bleeding, or stroke.

Machine learning (ML) has emerged as a powerful technique for analysing complex analytic problems. Machine learning algorithms use non-linear, highly interactive combinations of predictors to uncover novel patterns that may improve predictive performance.[6] However, ML algorithms are by their very nature more complex and less easily understood by clinicians. Despite the rapid expansion of ML techniques being applied to different types of data, to date, there have been few head-to-head comparisons of ML vs. traditional multivariable modelling. A study of hospitalized patients from five hospitals found that an ML model (random forest) for clinical deterioration performed better than logistic regression models (using either linear predictor terms or restricted cubic splines) or the commonly used Modified Early Warning Score.[7] Large outcomes studies have also shown improved prediction of cardiovascular events with ML models compared to established risk scores.[8,9] Using ML for prediction of heart failure readmissions has shown mixed results with some studies showing higher discriminatory power with ML models compared to logistic regression,[10] and others showing largely similar performance among ML and traditional regression.[11] The predictive accuracy of ML developed prediction models have not yet been directly compared to traditional regression modelling of stroke, bleeding, or death.

We used data from two very large community-based AF registries to examine whether ML was superior to traditional regression modelling for AF outcomes. The Outcomes Registry for Better Informed Treatment of Atrial Fibrillation (ORBIT-AF)[12,13] and the Global Anticoagulant Registry in the FIELD (GARFIELD-AF)[14] registries capture patient demographics, comorbidities, treatments, and outcomes. We compared the performance of ML algorithms to traditional multivariable regression techniques to determine which method provided better predictive performance in these large, structured data registries.

Methods

Study population

We analysed patients included in the ORBIT-AF, ORBIT-AF II, and GARFIELD-AF registries, the details of which have been previously published.[12,13,15] In brief, ORBIT-AF and ORBIT-II AF enrolled AF patients followed in outpatient practices and followed prospectively every 6 months for a minimum of 2 years. ORBIT-AF included 10 137 patients enrolled from 176 US practices between 29 June 2010 and 9 August 2011; ORBIT-AF II included 13 394 patients (unique from the ORBIT-AF cohort) enrolled from 244 US practices from February 2013 through 12 July 2016. Patients with complete baseline data and at least one follow-up encounter were included in the present analysis. GARFIELD-AF is a prospective, multicentre, international registry of patients with newly diagnosed AF and at least one additional risk factor for stroke. A total of 52 032 prospectively enrolled patients with follow-up provided from 35 countries were enrolled between March 2010 and July 2015. These patients were followed for a minimum of 2 years with data collection every 4 months for the first 2 years. The study protocol was reviewed and approved by the Duke University Medical Center Institutional Review Board (IRB) and the IRB at each enrolling centre and this study complies with the Declaration of Helsinki. The data underlying this article were provided by Ortho-McNeil Janssen Scientific Affairs, LLC (ORBIT-AF) and the Thrombosis Research Institute (GARFIELD-AF) by permission. Data will be shared on request to the corresponding author with permission of the respective parties.

Predictors and outcomes

Baseline variables as reported on the registries' case report forms were used as potential predictors. The final list of variables considered for all models were: age, sex, race, body mass index, diabetes mellitus (DM), hyperlipidaemia, hypertension, history of bleeding (gastrointestinal bleeding only for ORBIT-AF), chronic obstructive pulmonary disease (COPD), cancer, liver disease, peripheral vascular disease, coronary artery disease (CAD), significant valvular heart disease, heart failure (HF), cognitive impairment/dementia, anaemia, smoking status, drug abuse, alcohol abuse, frailty, type of AF (new onset, paroxysmal, persistent, permanent), heart rate, systolic and diastolic blood pressure, haemoglobin, and estimated glomerular filtration rate. In GARFIELD-AF, region (grouped into Europe, Latin America, and Asia, with Australia, Egypt and South Africa grouped together as 'Rest of the world') was also considered. In total, 30 variables were considered in ORBIT-AF and 32 in GARFIELD-AF.

Outcomes of interest included stroke, major bleeding, and death within 1 year of enrolment. Stroke for ORBIT-AF is defined as a new, sudden, focal neurologic deficit persisting for greater than 24 h and is not due to a readily identifiable, non-vascular cause (e.g. seizure). For GARFIELD-AF, the endpoint is the combined endpoint of primary ischaemic stroke or secondary haemorrhagic ischaemic stroke or systemic embolism. Major bleeding was defined based on the International Society of Thrombosis and Haemostasis.[16] The bleeding event included primary intracerebral Haemorrhage in GARFIELD-AF.

Prediction model

All models considered were fit using either the R (R foundation, Vienna, Austria) or Python (www.python.org) programming languages. The data were split 80:20 into training and tests sets, and area under the curves were calculated for the prediction of the outcomes of stroke, major bleeding, and death. For the ML methods, we further split the training dataset to estimate optimal tuning parameters via cross-validation.

The stepwise multivariable logistic regression model ('stepwise model') used a logit link and was estimated using the *step* function in R to perform stepwise elimination. The logistic regression models were fit to the occurrence of each outcome over available follow-up. Missingness was handled with single imputation.[17] The predictive capacity of the regression model was estimated via the mean value and 95% confidence interval for the C-statistic over 75 cross-validation iterations. The ML methods were fit using the scikit-learn Python library.[18] The ML models

tested in this study included random forests (RF), gradient boosting (GB), and two neural net (NN) structures. The RFs were fit using 500 estimators and a minimum of five samples per leaf. For GB, classification trees were used with a maximum depth of 3 as the weak base learner, a learning rate of 0.1. A 15-fold Monte Carlo cross-validation was used to find the optimal number of estimators (with a maximum set at 100) and 25% of the training data was subsampled to fit each weak learner. Each NN used early stopping, RELU activation, the ADAM optimizer, and a maximum of 200 iterations to fit. Neural net (1) used three layers with 5, 4, and 3 neurons, respectively, while NN (2) used one layer with only seven neurons. For a detailed mathematical description of each method, we refer the reader to the references.[19,20] After estimating the optimal tuning parameters on the training data, the model was fit on the whole training data set, an out-of-sample C-statistic was calculated on the test set. The predictive capacity of the models was estimated in the same way as the regression models, via the mean value and 95% confidence interval for the C-statistic over 75 cross-validation iterations. Additionally, model performance for both the regression and ML models were evaluated with calibration plots comparing expected and actual event rates for outcomes for one of the train/test splits. In order for the ML models to maintain stability, event rates were artificially increased via resampling with replacement. thus, the ML calibration curves may be distorted due to the models over-estimating event risk.

To assess the external validity of the models, a cross-registry analysis was performed using only variables that were common to both the ORBIT-AF and GARFIELD-AF registries. Using this more limited set of variables, stepwise multivariable logistic regression, RF and GB models were generated in one population then tested in the other population (i.e. models developed in the ORBIT-AF registry were tested in the GARFIELD-AF registry and vice versa). C-statistics for the ML methods were compared to the stepwise model using the DeLong test.[21] The added value of the ML techniques compared to the stepwise model was assessed using the net reclassification index (NRI).[22] The NRI measures the number of additional proportion of events that are correctly identified using one model compared to another (event NRI) as well as the number of non-events correctly identified (non-event NRI). The overall NRI is the sum of these two values.

Results

Study population

Baseline characteristics for the patients included in this study from the ORBIT-AF and GARFIELD-AF registries are shown in *Table 1*. In the ORBIT-AF I and II registry patients, the median age was 73 [interquartile range (IQR) 65–80], were 58% male and were predominantly white ($N = 19\,903$, 87%). The most common comorbidities included hypertension ($N = 18\,474$, 81%), CAD ($N = 6990$, 30%), and DM ($N = 6294$, 27%) and 1478 (6%) patients were active smokers at baseline. The majority of patients had paroxysmal or new-onset AF ($N = 16\,172$, 69%) with 6588 (28%) patients having persistent or permanent AF. After a median Follow-up of 540 days (IQR 360–783), there were 1871 deaths (5.6 per 100 patient-years), 1323 major bleeding events (3.9 per 100 patient-years), and 178 strokes (0.5 per 100 patient-years).

In the GARFIELD-AF registry, the median age was 71 (63–78), with 56% of patients being male and were predominantly white ($N = 24\,603$, 63%). Frequency of comorbid DM, hypertension and history of stroke as well as tobacco use was similar and the majority of patients had either new onset or paroxysmal AF ($N = 27\,937$, 72%). Follow-up in this cohort was truncated at 1 year over which time there were 1567 deaths (3.0 per 100 patient-years), 349 major bleeding events (0.7 per 100 patient-years), and 473 strokes (0.9 per 100 patient-years).

Model discrimination

Over the 75 iterations of the stepwise regression models, variables were included in models with varying frequency. The variables that were most frequently included in the models for each of the three outcomes for each cohort are shown in *Figure* 1A *and B*. Differences in the registry elements resulted in inclusion of different parameters for each of the registries. For example, in the prediction models for death, 100% of the iterations of stepwise regression for both registries included age, sex, current smoking status, CAD, congestive heart failure, diabetes, peripheral vascular disease, dementia/cognitive impairment, heart rate, blood pressure, and renal function. However, the models generated in the ORBIT-AF population also included former smoking status, cancer, valvular heart disease, haemoglobin, COPD, and frailty which were not available in the GARFIELD-AF registry. The GARFIELD-AF models similarly included history of acute coronary syndrome and medications that were not available in the ORBIT-AF registry. Of note, some variables that were present in both registries were included in one registries model but not the other. AF type was in 100% of the ORBIT-AF registry models for death, and none of the GARIFELD AF models. Cirrhosis, gastrointestinal bleeding, history of stroke, and race were included in 100% of the GARFIELD-AF models but none of the ORBIT-AF models.

The C-statistics for all the models are listed in *Table 2* and depicted in the Figure 2. C-statistics were highest for death and lowest for stroke in all models. Compared with stepwise regression, all tested ML models except for the GB model demonstrated lower C-statistics for major bleeding and all except the single-layer NN in the GARFIELD-AF population underperformed for stroke prediction. For death, the random forest models had similar discrimination as the stepwise models. The GB model was similar to the stepwise model in the ORBIT-AF population and provided slightly better discrimination than the stepwise model in the GARFIELD-AF population. The multilayer NN had the worst discrimination for all outcomes.

Model calibration

The calibration plots for the stepwise regression model as well as the RF and GB ML models are presented in *Figure 3*. The calibration of the stepwise model was best for all endpoints. The RF and GB models were best calibrated for the outcome of death but underestimated event rates for all outcomes.

Cross-registry analysis

The performance of each modelling technique on the subset of variables common to both ORBIT-AF and GARFIELD-AF are presented in *Table 3*. Due to the poor performance of the NNs, only the stepwise regression, random forest and GB models were evaluated in this common data model. Similar to the models developed from the more complete variable list, C-statistics were highest for death. The GB model trained in ORBIT-AF and tested in GARFIELD-AF had statistically significantly better discrimination than the stepwise model for the outcomes of death and major bleeding ($P < 0.001$ for both), though the magnitude of improvement was small. When the GB

Table 1 Baseline characteristics

	ORBIT-AF	GARFIELD-AF
N	22 760	52 032
Age (years)	73 (65–80)	71 (63–78)
Male	13 208 (58%)	29 042 (56%)
Race		
White	19 903 (87%)	32 005 (63%)
Black	1127 (5%)	243 (0.5%)
Hispanic	1073 (5%)	3392 (7%)
Other	657 (3%)	15 108 (30%)
Diabetes mellitus	6294 (27%)	11 546 (22%)
Hypertension	18 474 (81%)	39 610 (76%)
History of bleeding[a]	1463 (6%)	1416 (3%)
Cirrhosis	–	294 (<1%)
COPD	3065 (13%)	–
Cancer	4857 (21%)	–
Liver disease	470 (2%)	–
PVD	2358 (10%)	2859 (6%)
History of stroke	2864 (13%)	3878 (8%)
History of systemic embolism	–	335 (<1%)
CAD	6990 (31%)	11 253 (22%)
History of stent placement	–	3-550 (7%)
History of CABG	2600 (11%)	1625 (3%)
Cognitive impairment/dementia	489 (2%)	764 (2%)
Obstructive sleep apnoea	4045 (18%)	805 (2%)
Current smoker	1478 (6%)	4201 (11%)
Drug abuse	270 (1%)	–
Alcohol abuse[b]	861 (4%)	1026 (2%)
AF type		
New onset	6708 (29%)	23 331 (45%)
Paroxysmal	9464 (42%)	14 307 (28%)
Persistent	3255 (14%)	7758 (15%)
Permanent	3333 (15%)	6630 (13%)
Heart rate	72 (63–81)	84 (70–105)
Systolic BP	126 (116–138)	130 (120–145)
Diastolic BP	74 (67–80)	80 (70–88)
Haemoglobin	13.6 (12.3–14.7)	–
BMI	30 (26–24)	27 (24-31)
Estimated GFR	70.55 (56.15–86.29)	–
Chronic kidney disease[c]	6609 (31%)	5355 (10%)
VKA	8749 (38%)	20 183 (39%)
DOAC	10 214 (45%)	14 123 (28%)
Antiplatelet (with or without OAC)	9006 (40%)	18 104 (35%)

Values reported as median (interquartile range) or *N* (%) as appropriate.
AF, atrial fibrillation; BMI, body mass index; BP, blood pressure; CABG, coronary artery bypass graft; CAD, coronary artery disease; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; DOAC, direct-acting oral anticoagulant; GFR, glomerular filtration rate; OAC, oral anticoagulant; PVD, peripheral vascular disease; VKA, vitamin K antagonist.
[a]History of gastrointestinal bleeding only for ORBIT I and II.
[b]Alcohol abuse is defined as heavy alcohol use in GARFIELD-AF.
[c]Moderate to severe CKD only for GARFIELD-AF, any CKD for ORBIT-AF.

model was trained in GARFIELD-AF and tested in ORBIT-AF, it had similar discrimination for death and worse discrimination for major bleeding ($P < 0.0001$) or stroke ($P = 0.02$). Calibration curves for the cross-registry models are presented in the Supplementary material online, *Figures*. All the models showed the best calibration for death. In the ML models trained in ORBIT-AF and tested in GARFIELD-AF (Supplementary material online, *Figure S1*), the risk of stroke was underestimated and risk of major bleeding overestimated. The

Figure 1 Stepwise model parameter frequency. Frequency with which each parameter was included in the 75 iterations of the stepwise regression model for the ORBIT-AF cohort (*A*) and GARFIELD-AF cohort (*B*). ACS, acute coronary syndrome; AF, atrial fibrillation; BMI, body mass index; BP, blood pressure; CAD, coronary artery disease; CHF, chronic heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; GFR, glomerular filtration rate; GIB, gastrointestinal bleeding; NOAC, non-vitamin K antagonist oral anticoagulants; PVD, peripheral vascular disease; TIA, transient ischaemic attack; VTE, venous thromboembolism.

Table 2 C-statistics for ML models in ORBIT-AF and GARFIELD-AF populations

Outcome	Stepwise	Random forest	Gradient boosting	Multi-layer neural network	Single-layer neural network
ORBIT-AF population					
Death	0.801 (0.798–0.804)	0.797 (0.794–0.800)	0.802 (0.799–0.805)	**0.651** (0.643–0.659)	**0.788** (0.779–0.797)
Major bleeding	0.711 (0.707–0.715)	**0.699** (0.695–0.703)	**0.702** (0.698–0.706)	**0.584** (0.579–0.589)	**0.692** (0.682–0.702)
Stroke	0.671 (0.660–0.682)	**0.618** (0.608–0628)	**0.639** (0.629–0.649)	**0.563** (0.551–0.575)	**0.630** (0.613–0.647)
GARFIELD-AF population					
Death	0.752 (0.749–0.755)	0.752 (0.749–0.755)	**0.762** (0.759–0.765)	**0.731** (0.728–0.734)	0.758 (0.755–0.781)
Major bleeding	0.647 (0.641–0.653)	**0.630** (0.624–0.636)	0.643 (0.637–0.649)	**0.631** (0.624–0.638)	**0.632** (0.625–0.639)
Stroke	0.660 (0.656–0.664)	**0.638** (0.633–0643)	**0.649** (0.645–0.653)	**0.522** (0.516–0.528)	0.653 (0.648–0.658)

C-statistics for each model type and outcomes of death, major bleeding and stroke for the ORBIT-AF (upper panel) and GARFIELD-AF (lower panel) populations. Values in bold indicate C-statistics which are statistically significantly different from the stepwise model with red values less than the stepwise model and green values greater than the stepwise model.

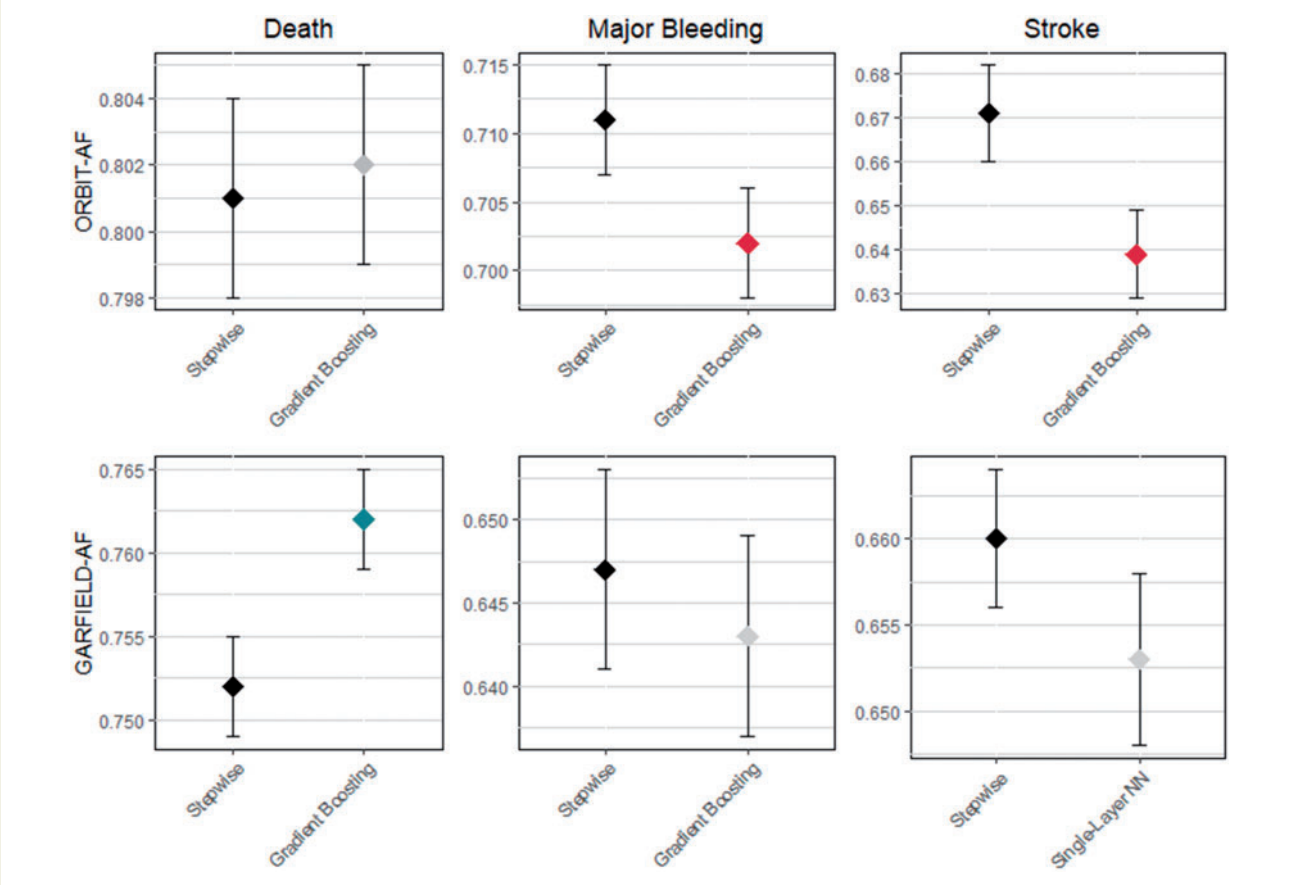

Figure 2 C-statistics and 95% confidence intervals for prediction models developed using stepwise regression and the best performing machine learning model for major clinical outcomes in two large atrial fibrillation registries.

opposite trend was observed in the models trained in GARFIELD-AF and tested in ORBIT-AF (Supplementary *Figure S2*).

We assessed the NRI when using RF and GB models compared to the stepwise model (*Table 4*). In the models trained in ORBIT-AF and tested in GARFIELD-AF, both the RF and GB models correctly identified fewer events and non-events for death. For major bleeding and stroke, the RF model correctly identified more events but misclassified more non-events resulting in an overall NRI only slightly above zero for both (0.038 for major bleeding, 0.024 for stroke). The GB model correctly identified 34.0% more bleeding events but misclassified 32.6% more non-events resulting in an overall NRI slightly greater than zero (0.014). The GB model identified fewer events and

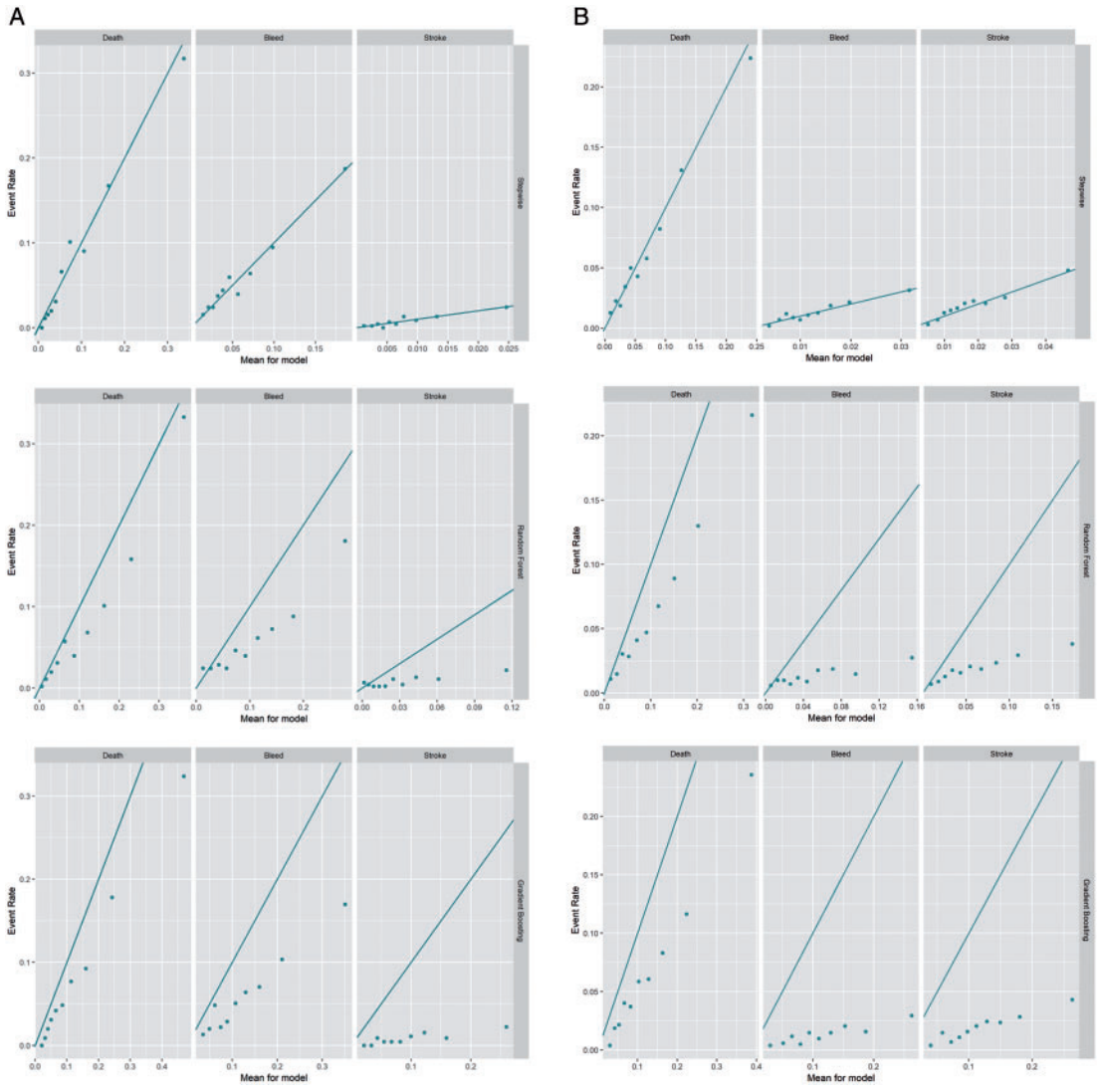

Figure 3 Calibration plots. Plots comparing predicted event rates (*x*-axis) and observed event rates (*y*-axis) for death (left), bleeding (middle), and stroke (right). Stepwise (top), random forest (middle), and gradient boosting (bottom) model results presented for both the ORBIT-AF cohort (*A*) and GARFIELD-AF cohort (*B*).

non-events for stroke. In the models trained in GARFIELD-and tested in ORBIT-AF, both the RF and GB models correctly identified more death events (3.4% and 16.2%, respectively), but incorrectly identified non-events more frequently resulting in negative overall NRIs. The RF model better identified 1.2% more bleeding events and 2.7% more non-bleeding events (overall NRI 0.039). The GB model for bleeding and both the GB and RF models for stroke did a poorer job in identifying both events and non-events.

Discussion

In this study of two, large contemporary AF registries which included more than 74 000 patients from more than 1000 practices across the world, stepwise regression models performed as well or better than ML for prediction of stroke, major bleeding, or death. All the models studied performed best at predicting death and worst for stroke. Of the ML modelling methods studied, the multilayer neural net had the lowest performance for all endpoints; whereas, GB performed the best for all endpoints. The stepwise regression model was better calibrated than the ML models. When evaluated across registries, the stepwise model demonstrated similar or better discrimination for most endpoints. The ML methods more frequently misclassified events and non-events when compared with the stepwise models. These results suggest that when analysing two real-world registries with structured data, stepwise regression models perform at least as well if not better than ML models for predicting outcomes.

Table 3 C-statistics for cross-registry models

Outcome	Model	Trained in ORBIT-AF		Trained in GARFIELD-AF	
		Tested in GARFIELD-AF		Tested in ORBIT-AF	
		C-Statistic	*P*-value	C-Statistic	*P*-value
Death	Stepwise	0.737		0.779	
	Random forest	0.731	0.03	0.773	0.07
	Gradient boosting	**0.744**	**0.0008**	0.781	0.40
Major bleeding	Stepwise	0.576		0.656	
	Random forest	0.578	0.85	**0.618**	**<0.0001**
	Gradient boosting	**0.597**	**0.0009**	**0.638**	**<0.0001**
Stroke	Stepwise	0.650		0.713	
	Random forest	**0.602**	**<0.0001**	**0.644**	**<0.0001**
	Gradient boosting	**0.622**	**<0.0001**	**0.696**	**0.02**

C-statistics for models generated with data elements common to both the ORBIT-AF and GARFIELD-AF registries, trained in one then tested in the other. Values in bold indicate C-statistics which are statistically significantly different from the stepwise model with red values less than the stepwise model and green values greater than the stepwise model.

Table 4 Net reclassification indices (NRI) for machine learning models compared to stepwise regression in cross-registry models

Outcome	Model	Trained in ORBIT-AF			Trained in GARFIELD-AF		
		Tested in GARFIELD-AF			Tested in ORBIT-AF		
		Event NRI	Non-event NRI	Overall NRI	Event NRI	Non-event NRI	Overall NRI
Death	Random forest	−0.5%	−6.26%	−0.068	3.4%	−18.5%	−0.151
	Gradient boosting	−8.5%	−13.6%	−0.220	16.2%	−30.7%	−0.144
Major bleeding	Random forest	25.6%	−21.7%	0.038	1.2%	2.7%	0.039
	Gradient boosting	34.0%	−32.6%	0.014	−12.2%	−19.4%	−0.315
Stroke	Random forest	16.1%	−13.7%	0.024	−14.1%	−2.9%	−0.171
	Gradient boosting	−6.4%	−17.8%	−0.242	−8.5%	−27.2%	−0.357

Event NRI denotes the percent of additional patients correctly predicted to have an event by the ML model, non-event NRI denotes the percent of additional patients correctly predicted not to have an event by the ML model. Overall NRI is the sum of the event and non-event NRIs and ranges from −2 to 2 with positive values (green) reflecting more accurate classification and negative values (red) reflecting less accurate classification.

All the evaluated models performed best at predicting the endpoint of all-cause mortality. There is a high degree of overlap among the risk factors for stroke, bleeding, and death and thus many of the variables captured in these registries are associated with increased risk of all three outcomes. However, while models of stroke and bleeding must account for the competing risk of death from other causes, models of all-cause mortality do not. Other risk models such as the GARFIELD-AF risk score and CHA_2DS_2-VASC score show higher discriminatory power for all-cause mortality than for embolic or bleeding events for a similar reason.[14]

The primary advantage of ML methods over linear models is their ability to learn complicated relationships and improve out-of-sample predictions.[23] This improvement comes at a cost: ML models are often difficult to interpret and communicate. Given a set of features, it can be difficult to understand the reason behind the prediction of an ML model whereas linear models allow for a decomposition into relevant parts. In this study, the ML methods failed to outperform the stepwise regression model for the assessment of three different outcomes in two independent populations of patients with AF. Other studies have demonstrated mixed results comparing model performance between ML methods and traditional regression modelling.[7–11] Two studies evaluating ML and traditional regression for prediction of HF readmission demonstrated conflicting results with one showing an improved performance with ML and the other no difference.[10,11] While both utilized structured data (clinical trial and registry case report forms) with a large number (>250) of candidate variables, the study showing similar performance between methodologies had substantially more patients (56 477 vs. 1004). Additionally, the discriminatory power of the ML methods in both studies were similar (C-statistic of 0.628 vs. 0.607 for random forest to predict

30 days HF readmission), but there was substantial differences in the performance of the logistic regression models (C-statistic of 0.533 vs. 0.624 in the smaller vs. larger study). This suggests that there may be a benefit to ML over regression in small samples, but these models perform similarly when derived in larger populations. This study of two large registries shows that stepwise models have similar discrimination and better calibration compared to the more difficult to interpret ML techniques. The stepwise model retains its attractiveness because of its ease of interpretation and use without a corresponding loss in predictive power.

Our results highlight the effect of the bias-variance tradeoff when building predictive models. Random forests, GB, and NNs have low bias and high variance on the training set, which can negatively impact their out-of-sample performance. On the other hand, the stepwise model has higher bias and lower variance than ML methods and leads to more consistent out-of-sample predictive performance. This may be particularly important in healthcare and biomedical science as predictive models are often applied in more diverse and heterogeneous populations than those in which they are derived. This challenge will likely become more important as prediction models become easier to embed in electronic medical records. The finding that the stepwise model was competitive, if not better, than all ML models considered (and particularly outperformed the multi-layer NN), suggests that the linear model captures the relationships in clinical variables we considered, and that non-linear classifiers add little, if anything, in this analysis.

This study reiterates the value of simple stepwise logistic models in determining which patients are at risk for death, stroke, and major bleeding. These models allow for a simple interpretation of the risk factors, can provide greater stability in out-of-sample predictions, and are easy to monitor over time. While ML methods have shown significant progress in incorporating complex data sources with large feature sets where appropriate data representations must be learned (e.g. image analysis), in the present analysis, we included a relatively small set of features with a historical literature showing their clinical relevance. Therefore, our results reiterate that while ML methods exert impressive utility in some clinical tasks, the first step in finding a robust predictive model is building an effective linear model.

Limitations

The stepwise regression and ML models were evaluated on a heterogeneous population including both patients on and off anticoagulation which may have confounded the models, particularly for prediction of major bleeding. Additionally, a higher proportion of patients in ORBIT-AF were treated with direct-acting oral anticoagulants (DOACs) compared to GARFIELD-AF. While both registries showed an increase in DOAC use over their enrolment period, this increase was more substantial in the ORBIT-AF group, reflecting the heterogeneity in treatment patterns across countries.[24] Subgroup analysis was performed on patients by anticoagulation status and showed similar results to the overall model. Predictor variables were obtained from case report forms with fixed options for responses. Using a more unstructured data collection tool may have revealed non-linear relationships that would be better assessed using ML techniques; however, the goal of the present study was to compare modelling techniques in a structured database rather than develop clinically useful prediction models. These databases may not have captured all relevant risk predictors; however, the goal of the present study was to compare the performance of different analytic techniques. All analytic techniques would be equally disadvantaged by missing important risk predictors, thus this should not impact the overall results. In order to maintain stability in the ML models, outcomes were resampled with replacement to increase the event rate. This does not influence the C-statistics for the models, but likely was the cause for the systemic overestimation of event rates in the calibration plots for the ML models. The two registries evaluated in this study evaluated distinct populations and had different assessments of baseline risk as well as different lengths of follow-up and event rates which may limit their comparability. However, the consistency of results in both populations as well as the patterns seen in the cross-registry analysis using a common data model highlight the generalizability of the studies main findings.

Conclusions

In conclusion, stepwise regression models performed as well or better than ML models for predicting clinical outcomes in large national and global AF cohorts. This suggests that traditional regression models may be better suited for developing prediction models in structured databases as they provide insight into the drivers of risk without compromising predictive capabilities. Machine learning methods have yielded impressive predictive performance when applied to semi-structured data such as electrocardiograms and chest radiographs.[25,26] Future work will compare the performance of ML models based on raw patient data, to existing clinical models. Ultimately, we hypothesize that ML will allow integration of non-structured data to existing data repositories to further improve future predictive models.

Supplementary material

Supplementary material is available at *Europace* online.

Funding

This work was supported by the Ortho-McNeil Janssen Scientific Affairs, LLC (primary funding for ORBIT-AF), and the Thrombosis Research Institute (primary funding for GARFIELD-AF).

Conflict of interest: Z.L. is supported in part by an NIH T32 training grant (#5T32HL069749) and receives grant support for clinical research from Boston Scientific. S.M. is supported in part by an NIH T32 training grant (#T32HL079896). J.P.P. receives grants for clinical research from Abbott, American Heart Association, Boston Scientific, Gilead, Janssen Pharmaceuticals, and the NHLBI and serves as a consultant to Abbott, Allergan, ARCA Biopharma, Biotronik, Johnson and Johnson, LivaNova, Medtronic, Milestone, Oliver Wyman Health, Sanofi, Philips, and Up-to-Date. J.A.C. reported consultancies with honoraria from Actelion Pharmaceuticals, Daiichi-Sankyo, Eli Lilly, GileadSciences, Inc., Heart Metabolics, InCardaTherapeutics, InfoBionic, Johnson and Johnson, Medtronic, Milestone, Pfizer, Boehringer Ingelheim, Boston Scientific, Novartis, Bayer, speaker's fees from Daiichi-Sankyo, Servier, Bayer/ScheringPharma, Boehringer Ingelheim, and research grants from Boehringer Ingelheim, Daiichi-Sankyo, and Pfizer. D.C. reports grants from the National Institutes of Health, Stylli Translational Neuroscience

Award, and Marcus Foundation. G.C.F. serves as a consultant to Abbott, Amgen, Bayer, Janssen, Medtronic, and Novartis. K.A.A.F. has received grants and honoraria from Bayer, Janssen, and AstraZeneca; has served as a consultant for Sanofi/Regeneron and Lilly. E.D.P. receives grants from Janssen Pharmaceuticals and Eli Lilly, discloses consulting for Janssen Pharmaceuticals and Boehringer Ingelheim. K.P. reports no disclosures or conflicts of interest. A.K.K. has received research support from Bayer AG; personal fees from Bayer AG, Boehringer-Ingelheim Pharma, Daiichi Sankyo Europe, Janssen Pharma, Sanofi SA and Verseon.

Data availability

Data availability statement in Methods section.

References

1. O'Brien EC, Kim S, Hess PL, Kowey PR, Fonarow GC, Piccini JP *et al*. Effect of the 2014 atrial fibrillation guideline revisions on the proportion of patients recommended for oral anticoagulation. *JAMA Intern Med* 2015;**175**:848–50.
2. January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC Jr *et al*. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *J Am Coll Cardiol* 2014;**64**:e1–76.
3. Friberg L, Rosenqvist M, Lip GY. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the Swedish Atrial Fibrillation cohort study. *Eur Heart J* 2012;**33**:1500–10.
4. O'Brien EC, Simon DN, Thomas LE, Hylek EM, Gersh BJ, Ansell JE *et al*. The ORBIT bleeding score: a simple bedside score to assess bleeding risk in atrial fibrillation. *Eur Heart J* 2015;**36**:3258–64.
5. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B *et al*. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Europace* 2016;**18**:1609–1678.
6. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;**375**:1216–9.
7. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016;**44**:368–74.
8. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;**12**:e0174944.
9. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R *et al*. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res* 2017;**121**:1092–101.
10. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li S-X *et al*. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes* 2016;**9**:629–40.
11. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF *et al*. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017;**2**:204–9.
12. Piccini JP, Fraulo ES, Ansell JE, Fonarow GC, Gersh BJ, Go AS *et al*. Outcomes registry for better informed treatment of atrial fibrillation: rationale and design of ORBIT-AF. *Am Heart J* 2011;**162**:606–12.e601.
13. Steinberg BA, Blanco RG, Ollis D, Kim S, Holmes DN, Kowey PR *et al*. Outcomes registry for better informed treatment of atrial fibrillation II: rationale and design of the ORBIT-AF II registry. *Am Heart J* 2014;**168**:160–7.
14. Fox KAA, Lucas JE, Pieper KS, Bassand JP, Camm AJ, Fitzmaurice DA *et al*. Improved risk stratification of patients with atrial fibrillation: an integrated GARFIELD-AF tool for the prediction of mortality, stroke and bleed in patients with and without anticoagulation. *BMJ Open* 2017;**7**:e017157.
15. Kakkar AK, Mueller I, Bassand JP, Fitzmaurice DA, Goldhaber SZ, Goto S *et al*. International longitudinal registry of patients with atrial fibrillation at risk of stroke: global Anticoagulant Registry in the FIELD (GARFIELD). *Am Heart J* 2012;**163**:13–9.e11.
16. Schulman S, Kearon C; Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. *J Thromb Haemost* 2005;**3**:692.
17. Enders CK. *Applied Missing Data Analysis*. New York: Guilford Press; 2010.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
19. Géron A. *Hands-on Machine Learning with SciKit-Learn and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems*. Sebastapol: O'Reily Media; 2017.
20. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York: Springer; 2013.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–45.
22. Pencina MJ, D' Agostino RB, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157–72; discussion 207–112.
23. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;**104**:1156–64.
24. Steinberg BA, Gao H, Shrader P, Pieper K, Thomas L, Camm AJ *et al*. International trends in clinical characteristics and oral anticoagulation treatment for patients with atrial fibrillation: results from the GARFIELD-AF, ORBIT-AF I, and ORBIT-AF II registries. *Am Heart J* 2017;**194**:132–40.
25. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G *et al*. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–4.
26. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH *et al*.; for the DLAD Development and Evaluation Group. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;**2**:e191095.