Detecting undiagnosed atrial fibrillation in UK primary care: validation of a machine learning prediction algorithm in a retrospective cohort study

Sara Sekelj[1], Belinda Sandler[2], Ellie Johnston[1], Kevin G Pollock[2], Nathan R Hill[2], Jason Gordon[3], Carmen Tsang[3], Sadia Khan[4], Fu Siong Ng[4,5], Usman Farooqui[2]

[1]Imperial College Health Partners, 30 Euston Square, London NW1 2FB, UK

[2]Uxbridge Business Park, Sanderson Road, Uxbridge, Middlesex UB8 1DH, UK

[3]Health Economics and Outcomes Research Ltd, Rhymney House, Unit A Copse Walk, Cardiff Gate Business Park, Cardiff CF23 8RB, UK

[4]Chelsea & Westminster Hospital NHS Foundation Trust, 369 Fulham Rd, Chelsea, London SW10 9NH, UK

[5]Imperial College London, London W12 0NN, UK

This work was supported by Bristol-Myers Squibb Pharmaceuticals Ltd who provided funding for data collection, data analysis and medical writing for this study.

Corresponding author: Jason Gordon

Health Economics and Outcomes Research Ltd, Rhymney House, Unit A Copse Walk, Cardiff Gate Business Park, Cardiff CF23 8RB

Email: jason.gordon@heor.co.uk

Tel: 02920 399146

Full word count (including 3 tables): 4,319 words

ABSTRACT

Aims: To evaluate the ability of a machine learning (ML) algorithm to identify patients at high risk of atrial fibrillation (AF) in primary care.

Methods: A retrospective cohort study was undertaken using the DISCOVER registry to validate an algorithm developed using a Clinical Practice Research Datalink (CPRD) dataset. The validation dataset included primary care patients in London, England aged ≥30 years from 01 January 2006 to 31 December 2013, without a diagnosis of AF in prior five years. Algorithm performance metrics were sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and number needed to screen (NNS). Subgroup analysis of patients aged ≥65 years also performed.

Results: Of 2,542,732 patients in DISCOVER, the algorithm identified 604,135 patients suitable for risk assessment. Of these, 3.0% (17,880 patients) had a diagnosis of AF recorded before study end. The AUROC was 0.87, compared with 0.83 in algorithm development. The NNS was nine patients, matching the CPRD cohort. In patients aged ≥30 years, the algorithm correctly identified 99.1% of patients who did not have AF (NPV) and 75.0% of true AF cases (sensitivity). Among patients aged ≥65 years (n=117,965), the NPV was 96.7% with 91.8% sensitivity.

Conclusions: This AF risk prediction algorithm, based on ML methods, identified patients at highest risk of AF. It performed comparably in a large, real-world population-based cohort and the developmental registry cohort. If implemented in primary care, the algorithm could be an effective tool for narrowing the population who would benefit from AF screening in the United Kingdom.

Abstract word count: 250 words

Keywords: Atrial Fibrillation; Machine Learning; Models, Statistical; Sensitivity and Specificity; Primary Health Care

INTRODUCTION

Atrial fibrillation (AF), is the most common arrhythmia (irregular heart rhythm disorder), and is associated with disability following AF-related stroke, heart failure and premature death.[1, 2] Patients with AF have an approximately fivefold increase in stroke incidence[3] and an approximately twofold increase in the risk of death within 30 days of an AF-related stroke compared with patients without AF.[4] It is estimated that approximately 1.4 million people in England are living with AF, however AF can be difficult to diagnose because it is often paroxysmal and/or asymptomatic or minimally symptomatic.[5, 6] As a result, an estimated 30% of patients living with AF are undiagnosed.[5] Early detection and effective management of AF are likely to both improve patient outcomes and reduce the economic burden of AF-related morbidity. Detection of undiagnosed AF is a fine balance between the associated patient burden, healthcare resource use and costs on one hand, and diagnostic sensitivity and specificity on the other. Opportunistic testing for AF in symptomatic or high-risk patients (such as those with irregular pulse or aged ≥65 years as risk of AF increases with age) typically requires frequent electrocardiogram (ECG) tests to capture the arrhythmia.[7]

In the absence of a formal screening programme in the UK, prediction models based on risk factors for AF could help to identify patients at highest risk of AF to offer a more targeted approach to screening. Existing risk prediction tools include the Framingham,[8] Atherosclerosis Risk In Communities (ARIC,[9] and Cohorts for Aging and Research in Genomic Epidemiology (CHARGE-AF)[10] models, however all of these were developed in the United States and therefore may not be directly applicable to other populations and health care systems, such as patients in the UK.[11, 12] Furthermore, some of these tools require ECG-derived data which is not available to all patients,[8, 9] and none are automated, meaning they are difficult to implement in routine clinical practice.

Machine learning (ML) is a form of artificial intelligence that is particularly useful for examining non-linear associations and complex interactions between variables without having to specify these relationships *a priori.* Investment in, development, and adoption of artificial intelligence across the NHS is at the forefront of the UK government's healthcare agenda.[13, 14] Novel but clinically useful applications of artificial intelligence, such as ML-based prediction algorithms, may have a role in automated screening of a chosen population (e.g. a General practitioner, GP, practice) to narrow the population who could benefit from screening for AF.

As AF has a complex aetiology, models developed using these methods may offer improved predictive performance compared with models built with classical statistical methods to estimate AF-risk. Indeed, a recently published AF risk prediction algorithm, developed using routinely collected UK primary care data from the Clinical Practice Research Datalink (CPRD) was better able to identify patients at highest risk of AF compared with existing models.[15] Compared to the CHARGE-AF model, the AF-risk prediction algorithm was able to reduce the number of high-risk patients needed to be screened to identify one case of AF by 31%, from 13 to 9.[15] However, whilst results are promising based on CPRD data, the algorithm has not yet been applied to other data sources and it is unknown how well the model will perform with different population-based data. Therefore, the aim of this study was to externally validate the ML AF risk prediction algorithm in a large, independent dataset.

METHODS

Study design and data source

For this external validation, a retrospective cohort study was undertaken using coded primary care data from the Whole Systems Integrated Care (WSIC) dataset, which is one of Europe's largest patient-level datasets, containing data from approximately 2.5 million patients across North West London (NWL) at any given time. Study data were obtained through the DISCOVER secure environment, which was developed by Imperial College Health Partners, the Academic Health Science Network for NWL. Unlike other datasets such as CPRD which include data extracted from only a proportion of the primary care population in the UK, DISCOVER contains data for 95% of the population in NWL.[16]

Primary care data in WSIC are extracted directly from clinical systems and sensitive data, such as abortions, and patient opt-outs are purged. Invalid data are either removed, redirected or logged and reported to clinical users as part of data quality checks. The completeness of data in WSIC on six key risk factors (alcohol intake, blood pressure, BMI, cholesterol, ethnicity and smoking) has been previously investigated.[16] The completeness of recorded data in DISCOVER for each of these factors has increased over time. In 2017, completeness was over 70% for smoking, blood pressure, ethnicity, alcohol and BMI, while cholesterol was at 50% completeness.[16]

Favourable ethical opinion was secured in October 2018 to use the Discover Research Platform for research purposes for a period of five years. The Research Ethics Committee reference is 18/WM/0323 and the Integrated Research Application System project identifier is 253449. The opinion clearly stated that there is no requirement for each application to request ethical approval. This research did successfully secure local Research and Development Department

approval to proceed from the NWL Data Research Access Group on 18th October 2018. Patient consent was not required because the study was retrospective study using anonymised data.

Eligibility criteria

The eligible cohort included patients registered with DISCOVER who were aged ≥30 years between 01 January 2006 and 31 December 2016 and who had no history of AF recorded in the preceding five years. Only patients with a complete set of height, weight, body mass index (BMI; three measurements or two measurements and one calculated from standard formulae), systolic blood pressure and diastolic blood pressure within a 12-month period were eligible for inclusion in the study. The index date was defined as the date when the required complete set of measurements was recorded.

Observation period

De-identified primary care data for all patients were extracted via DISCOVER for the period 01 January 2001 to 31 December 2016. Patients were followed up until AF diagnosis or death, transfer out of practice, or study end date (31 December 2016), whichever occurred earliest. Diagnoses of AF, and patient factors included in the model, were identified using relevant Read codes, the coded clinical term system used in UK primary care, that were used for algorithm development (see full code list provided in Hill et al).[15] A five-year look-back period from the index date was used to detect recorded comorbidities.

Sample size considerations

DISCOVER contains records for approximately 2.5 million patients in the general population of NWL. In the algorithm development study using CPRD data, 43.2% of the overall patient sample met all of the study eligibility criteria and were included in the study. It was assumed

that a comparable proportion of patients in the DISCOVER dataset would meet the eligibility criteria (i.e. up to approximately 1 million patients).

Statistical analyses to assess model performance

In order to assess the model's predictive ability to distinguish between patients at high and relatively lower risk of AF, a risk threshold for AF was generated among eligible patients in the Discover dataset. These thresholds were derived from baseline risk factors (age, previous cardiovascular disease, antihypertensive medication usage) and additional time-varying predictors (proximity of cardiovascular events, body mass index [levels and changes], pulse pressure, and frequency of blood pressure measurements). Ethnicity was included in algorithm development but was poorly recorded in DISCOVER at the time of data extraction in this study and was therefore unavailable for analyses. The predictive performance of the model was then assessed using the following metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), number needed to screen (NNS) and area under the curve of the receiver operating characteristic (AUROC) for discrimination between patients with and patients without AF.

Analyses were undertaken with model sensitivity set at 50% and 75%, with corresponding risk thresholds based on the baseline data of eligible patients in DISCOVER of 5.5% and 2.3%, respectively. As the model was originally developed for patients aged ≥30 years, sub-analyses were undertaken in patients aged ≥65 years as AF is more prevalent with increasing age. Reporting and presentation of model validation results was guided by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (see supplementary file).

Descriptive results were reported using summary statistics. Categorical data were summarised with counts and percentages, while continuous data were summarised using mean with standard deviation (SD), or median with interquartile range (IQR) or range, where appropriate. Normality was assessed using skewness/kurtosis tests. Wilcoxon signed rank tests and test of proportions were used to determine statistically significant differences between patients with AF (AF patients) and patients without AF (non-AF patients). Results with p-values <0.05 were considered statistically significant. All analyses were performed using Microsoft Excel 2013, Stata version 15.0 and R version 3.6.0.

RESULTS

Patient characteristics

Out of 2,542,732 patients in the DISCOVER database, 604,135 patients (23.8%) met the eligibility criteria and were included in the study. Patients were followed up for a median of 8.0 years (AF patients: 5.1 years, IQR 2.9 to 11.0 years and non-AF patients: 8.1 years, IQR 5.1 to 10.2 years), with total follow-up of 4,464,687 person-years (AF patients: 94,716 person-years and non-AF patients: 4,369,971 person-years). During follow-up, 17,880 (3.0%) patients had a recorded diagnosis of AF.

Patients with AF were significantly older than patients without AF ($p<0.001$). The mean age of patients with AF was 69 years (SD 11.3) compared with 52 years (SD 13.0) for those without AF. Table 1 also shows that a greater proportion of AF patients were male than non-AF patients (54.3% compared with 48.8%, $p<0.001$) and the mean BMI in AF patients was higher than in non-AF patients, 28.4 kg/m^2 (SD 6.6) compared with 27.0 kg/m^2 (SD 6.1), $p<0.001$. A smaller proportion of patients with AF had type 1 diabetes (10.5%) compared with non-AF patients (15.6%), $p<0.001$ but a greater proportion had type 2 diabetes (15.0% versus 7.5% respectively, $p<0.001$).

Hypertension was the most common condition among patients with and patients without AF (45.5%, n=8,137/17,880 and 17.1%, n=100,159/586,255 respectively). There were significant differences in the clinical histories of patients with and without AF in all conditions of interest ($p\leq0.002$ for each condition) except for congenital heart disease.

Comparison with algorithm development population

Similar rates of AF were identified in the model development and validation datasets (3.2% [n=95,607/2,994,837] and 3.0% [n=17,880/604,135], respectively). Table 2 shows that there were significant differences between the datasets used for developing and validating the algorithm in most baseline patient demographic and clinical characteristics.

For example, 46.6% (n=1,395,397/2,994,837) of all patients in the development study were male compared with 49.0% (n=295,861/604,135) in this study (p<0.0001). The mean age of AF patients was 70.2 years (SD 11.1) in the CPRD study compared with 68.6 years (SD 11.3) in this study (p<0.0001). Fewer patients in this study, overall and in the AF group, were former smokers compared with patients in the development dataset. Greater proportions of patients had been diagnosed with type 1 and type 2 diabetes in this study, p<0.0001 for type 2 diabetes in the overall sample and also the AF group. Overall, 23.4% (n=701,966/2,994,837) in CPRD versus 15.9% (n=96,327/604,135) in DISCOVER, (p<0.0001). AF group – 33.7% (n=32,198/95,607) in CPRD versus 26.3% (n=4,697/17,880), (p<0.001).

Overall algorithm performance

The algorithm's AUROC was 0.87 for patients aged ≥30 years in this study (see supplementary material). At 75% sensitivity and with risk threshold of 2.3%, the algorithm achieved 82.0% specificity, 11.3% PPV and 99.1% NPV, indicating an NNS of nine patients. Table 3 also shows that at reduced sensitivity of 50% and with a risk threshold of 5.5%, specificity, and PPV increased to 92.6% and 16.9%, respectively, while the NPV was similar (98.4%) but the NNS was reduced to six patients.

Performance in patients aged 65 years or older

Approximately a fifth of the study population was aged ≥65 years (19.5%, n=117,965/ 604,135). Among these patients, 10.3% were diagnosed with AF during follow-up (n=12,124/117,965). The algorithm achieved an AUROC of 0.71 in patients aged ≥65 years (see supplementary material). At a risk threshold of 2.3%, algorithm sensitivity was 91.8% with 27.4% specificity, 12.6% PPV, 96.7% NPV and an NNS of eight patients. At a higher risk threshold of 5.5%, sensitivity was 64.8% with 65.9% specificity, 17.9% PPV, 94.2% NPV and an NNS of six patients.

Comparison with model development

Compared with the development study, among patients aged ≥30 years the algorithm displayed better discriminative performance in the validation population (AUROC 0.83 versus 0.87). Table 3 shows that at 50% and 75% sensitivities, the algorithm correctly identified more patients aged ≥30 years without AF during model validation than in development. PPV and NPV were similar using CPRD and DISCOVER datasets at 50% and 75% sensitivities (Table 3). In algorithm development, the risk threshold was set at 7.4% to reach 50% sensitivity. During validation, a lower threshold was required to reach the same sensitivity, with comparable specificity, PPV and NNS to the development study.

DISCUSSION

Among the 604,135 patients in the DISCOVER database who met the eligibility criteria, 3.0% had been diagnosed with AF by the end of the follow-up period. The prediction algorithm displayed good discriminatory power (AUROC 0.87) in distinguishing between patients with AF and those without AF. In the subgroup of patients aged ≥65 years, the discriminatory power was slightly weaker, but was nevertheless acceptable (AUROC 0.71).

For a risk prediction algorithm to have clinical validity and utility, it needs to be accurate at ensuring patients classified as low risk are free from AF. At 75% sensitivity, the corresponding risk threshold was 2.3% and the NPV was 99.1%, indicating that >99% of patients classified as low risk did not go on to develop AF during the follow-up period. Similarly, a PPV of 11.3% indicated that only nine higher-risk patients would need to be screened to identify one case of AF. In the subgroup of patients aged ≥65 years, the NPV still remained high at 96.7%, and only eight higher-risk patients would be required screening to identify one case of AF.

A key strength of the study is that it is one of the first to demonstrate the validity of a ML-based prediction model to assess the risk of AF using patient records from a large research database covering approximately 2.5 million patients in one geographical area of England. The study benefitted from using data covering a different population of patients and GP practices to the CPRD model development dataset. While the CPRD dataset included records from practices that use the Vision clinical computer system, this system is used in less than 15% of GP practices across London.[17] There was no overlap between the datasets used in the two studies as DISCOVER contains records from two clinical computer systems other than Vision.[16]

Nevertheless, there are some methodological limitations. For example, results from this study are based on data from a geographically localised area in England and, therefore, the findings may not be generalisable to patient populations in other regions of the UK or other countries. Related to this, a large proportion of AF diagnoses are made opportunistically in secondary care.[18] Therefore it is important to further assess model performance using linked primary care and secondary care data. Due to the use of existing, routinely collected data, analyses and interpretation of results were limited by the accuracy and completeness of original data entry. For example, ethnicity was poorly recorded in DISCOVER at the time of data extraction (but was included in algorithm development) and, therefore, this predictor variable was missing from analyses. However, in contrast to accessible and explicit classical statistical methods, prediction algorithms built using ML methods, lack transparency.[19, 20] As such, the true impact of completely missing ethnicity data on the algorithm's performance is unknown.

While the rate of AF among patients eligible for screening in this study (3.0%) was comparable to the estimated national prevalence of AF (2.5%)[5] and the development study (3.2%),[15] there were notable differences between the DISCOVER and CPRD datasets in patients' demographic and clinical characteristics. These sampling variations were reflected in the different risk thresholds applied in the two studies to reach 50% sensitivity (5.5% and 7.4%, respectively), which is typical in studies such as this one where the risk threshold is data-driven.[21] Additionally, the relative importance of individual risk factors and the relationships between them (i.e. risk profiles), are likely to be different in patients aged ≥30 years compared with patients aged ≥65 years. Consequently, the algorithm would need to be recalibrated for a new population (patients aged ≥65 years).

In both the development study and this validation study, the model displayed improved discriminatory power (AUROC 0.83 and 0.87 respectively) compared with the best performing existing risk prediction model, CHARGE-AF (AUROC 0.73).[10, 15] Other AF prediction tools have been developed using large datasets routinely collected for clinical and administrative purposes, such as the HAVOC (abbreviation for hypertension, age, valvular heart disease, peripheral vascular disease, obesity, congestive heart failure and coronary artery disease) score for detecting AF after cryptogenic stroke and transient ischemic attack using data from the United States.[22, 23] The algorithm showed superior power to the HAVOC score, which had AUROCs of 0.77 and 0.69 at development and validation, respectively.

In the algorithm development study, it was estimated that nine patients identified as at higher risk of AF by the algorithm needed to be screened to diagnose one case of AF.[15] This finding was confirmed in this validation study and indicates the algorithm is more effective than the CHARGE-AF model, which required 13 higher risk patients to be screened to diagnose one case of AF.[10] Furthermore, a NNS of nine is far superior to that reported in a recent systematic review and meta-analysis by Lowres et al (2019) that included studies using AF screening methods (including a mix of opportunistic and more systematic approaches) accepted by the European Society of Cardiology.[24] Lowres et al estimated from meta-regression results that 294 patients aged <60 years are required to be screened to diagnose one case of AF. Even among patients aged ≥65 years, the NNS was 69, significantly greater than the NNS of nine reported in this study.[24]

ML techniques have already been used to aid in the identification of AF via ECG and pulse waveforms.[25-28] Such techniques can detect subtle changes in the ECG waveform that are invisible to the human eye even when a patient is in sinus rhythm,[29] and increase the ability of

clinicians to identify paroxysmal and/or asymptomatic AF. However, these techniques are limited to patients receiving an ECG, or pulse waveform analysis. A key advantage of the AF risk prediction algorithm evaluated in this study is that it can be applied at the population-level and only requires readily available, routinely collected healthcare data with no requirement for ECG analysis. Furthermore, as the AF risk prediction algorithm correctly assigned a low risk of AF to 99% of patients, and only required nine higher risk patients to be screened to detect one case of AF, its routine use in clinical practice is unlikely to result in unnecessary burden on patients or healthcare services.

Current AF screening approaches either lack cost-effectiveness or diagnostic precision. European guidelines recommend the diagnosis of AF via ECG,[30] yet this approach is not always considered cost-effective, regardless of whether screening is targeted at higher-risk patients only, or systematic (e.g. including all patients aged >65 years).[31] Therefore, opportunistic screening by way of a pulse check is favoured in primary care because it is cost-effective but lacks diagnostic precision.[31, 32] There is evidence that systematic screening approaches are more effective than opportunistic activities, especially when GP-led,[33] and, given that an estimated 30% of patients living with AF are undiagnosed,[5] there is significant value in a systematic, accurate, automated risk prediction algorithm that could be applied to medical records of patients in primary care to identify patients at highest risk of undiagnosed AF who should be invited for further screening.

One in five strokes are linked to AF and many patients are only diagnosed with AF following a stroke event.[2] The burden of stroke on healthcare systems, and patients and their families is substantial and is set to rise alongside the ageing population. However, up to two thirds of AF-related strokes can be prevented with anticoagulation therapy.[34] Interventions such as the AF

risk prediction algorithm evaluated in this study can narrow the population that should be considered for screening in a cost-effective manner,[35] and may potentially enable earlier detection of the condition. There is evidence that some limitations of ECG-based screening may be overcome by the use of portable, hand-held ECG machines by patients at home.[36, 37] It is possible that combined use of different interventions, such as the risk prediction algorithm to identify patients at high risk of AF along with portable ECGs to support diagnosis, may facilitate the timely management of AF in resource and budget-constrained healthcare environments.

CONCLUSION

The aim of this study was to externally validate the performance of a previously developed AF risk prediction algorithm. In this dataset using data from the DISCOVER research database, the algorithm performed similarly to its development dataset. Current screening approaches for AF tend to lack either diagnostic precision and/or cost-effectiveness. Conversely, this AF risk prediction algorithm, which identifies patients at highest risk of AF based on routinely collected patient data, may be useful to narrow the population to detect those at highest risk of AF who should undergo further screening. However, the performance of the algorithm in the wider real-world clinical setting is unknown. Therefore, it will be important to assess the clinical, and also economic, impact of implementing the risk prediction algorithm in routine clinical practice and its clinical value in supporting timely diagnosis of AF. Further research would benefit from patient representative input to ensure the needs of patients are fully considered.

ACKNOWLEDGEMENTS

Medical writing and editorial assistance in the preparation of this article was provided by Carissa Dickerson and Angharad Morgan of Health Economics and Outcomes Research Ltd. Support for this assistance was funded by Bristol-Myers Squibb Company. With thanks to Rachel Ashton, Imperial College Health Partners, for help with medical writing and to Trudie C A Lobban, Founder and Chief Executive Officer of the Atrial Fibrillation (AF) Association, for reviewing the draft manuscript.

FUNDING

This work was supported by Bristol-Myers Squibb Pharmaceuticals Ltd who provided funding for data collection, data analysis and medical writing for this study.

CONFLICTS OF INTEREST

BS, KGP, NRH and UF are employees of BMS. JG and CT are employees of HEOR Ltd, which received funding from BMS to undertake this study. SS and EJ are employees of Imperial College Health Partners, which received funding from BMS to undertake this study. FSN acknowledges funding from the British Heart Foundation (RG/16/3/32175) and the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC).

AUTHORS' CONTRIBUTIONS

BS, NRH, JG and UF contributed to the conception or design of the work. All authors contributed to the acquisition, analysis, or interpretation of data for the work. CT drafted the manuscript. All authors critically revised the manuscript, gave final approval and agree to be accountable for all aspects of work ensuring integrity and accuracy.

DATA SHARING STATEMENT

The datasets analysed in this study are available in the DISCOVER database and available on request from Imperial College Health Partners at https://www.registerfordiscover.org.uk/researchers/how-to-access-discover.

REFERENCES

1. Lamassa M, Di Carlo A, Pracucci G, et al. Characteristics, outcome, and care of stroke associated with atrial fibrillation in Europe: data from a multicenter multinational hospital-based registry (The European Community Stroke Project). *Stroke* 2001; 32: 392-398. 2001/02/07. DOI: 10.1161/01.str.32.2.392.
2. Marini C, De Santis F, Sacco S, et al. Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study. *Stroke* 2005; 36: 1115-1119. 2005/05/10. DOI: 10.1161/01.STR.0000166053.83476.4a.
3. Wolf PA, Abbott RD and Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke* 1991; 22: 983-988. 1991/08/01. DOI: 10.1161/01.str.22.8.983.
4. Lin HJ, Wolf PA, Kelly-Hayes M, et al. Stroke severity in atrial fibrillation. The Framingham Study. *Stroke* 1996; 27: 1760-1764. 1996/10/01. DOI: 10.1161/01.str.27.10.1760.
5. Public Health England. Technical document for sub-national English atrial fibrillation prevalence estimates: Application of Age-sex rates in a Swedish region to the English population, https://www.gov.uk/government/publications/atrial-fibrillation-prevalence-estimates-for-local-populations (2017).
6. Savelieva I and Camm AJ. Clinical relevance of silent atrial fibrillation: prevalence, prognosis, quality of life, and management. *J Interv Card Electrophysiol* 2000; 4: 369-382. DOI: 10.1023/a:1009823001707.
7. National Institute for Health and Care Excellence. Clinical guideline [CG180]. Atrial fibrillation: management. https://www.nice.org.uk/guidance/cg180/documents/thousands-of-strokes-in-people-with-common-heart-rhythm-disorder-are-avoidable-says-nice (2014).
8. Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet (London, England)* 2009; 373: 739-745. 2009/03/03. DOI: 10.1016/s0140-6736(09)60443-8.
9. Chamberlain AM, Agarwal SK, Folsom AR, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *The American journal of cardiology* 2011; 107: 85-91. 2010/12/15. DOI: 10.1016/j.amjcard.2010.08.049.
10. Alonso A, Krijthe BP, Aspelund T, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *Journal of the American Heart Association* 2013; 2: e000102. 2013/03/30. DOI: 10.1161/jaha.112.000102.
11. Pfister R, Bragelmann J, Michels G, et al. Performance of the CHARGE-AF risk model for incident atrial fibrillation in the EPIC Norfolk cohort. *Eur J Prev Cardiol* 2015; 22: 932-939. 2014/07/26. DOI: 10.1177/2047487314544045.
12. Kolek MJ, Graves AJ, Xu M, et al. Evaluation of a Prediction Model for the Development of Atrial Fibrillation in a Repository of Electronic Medical Records. *JAMA Cardiol* 2016; 1: 1007-1013. DOI: 10.1001/jamacardio.2016.3366.
13. Department of Health. *NHS Long Term Plan*. 2019. Department of Health,;.
14. O'Dowd A. Government pins hopes on £250m AI centre for faster diagnosis and treatment. *BMJ (Clinical research ed)* 2019; 366: l5106. DOI: 10.1136/bmj.l5106.
15. Hill NR, Ayoubkhani D, McEwan P, et al. Predicting atrial fibrillation in primary care using machine learning *Plos One* 2019; 14: e0224582-e0224582. DOI: 10.1371/journal.pone.0224582.
16. Bottle A, Cohen C, Lucas A, et al. How an electronic health record became a real-world research resource: comparison between London's Whole Systems Integrated Care database and the Clinical Practice Research Datalink. *BMC Med Inform Decis Mak* 2020; 20: 71-71. DOI: 10.1186/s12911-020-1082-7.
17. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018; 8: e020738. DOI: 10.1136/bmjopen-2017-020738.
18. Morley KI, Wallace J, Denaxas SC, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLOS ONE* 2014; 9: e110900. DOI: 10.1371/journal.pone.0110900.

19. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ (Clinical research ed)* 2019; 364: l886. DOI: 10.1136/bmj.l886.
20. Char DS, Shah NH and Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med* 2018; 378: 981-983. 2018/03/15. DOI: 10.1056/NEJMp1714229.
21. Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. *BMC Medicine* 2019; 17: 192. DOI: 10.1186/s12916-019-1425-3.
22. Kwong C, Ling AY, Crawford MH, et al. A Clinical Score for Predicting Atrial Fibrillation in Patients with Cryptogenic Stroke or Transient Ischemic Attack. *Cardiology* 2017; 138: 133-140. 2017/06/28. DOI: 10.1159/000476030.
23. Ntaios G, Perlepe K, Lambrou D, et al. External Performance of the HAVOC Score for the Prediction of New Incident Atrial Fibrillation. *Stroke*; 0: STROKEAHA.119.027990. DOI: 10.1161/STROKEAHA.119.027990.
24. Lowres N, Olivier J, Chao T-F, et al. Estimated stroke risk, yield, and number needed to screen for atrial fibrillation detected through single time screening: a multicountry patient-level meta-analysis of 141,220 screened individuals. *PLOS Medicine* 2019; 16: e1002903. DOI: 10.1371/journal.pmed.1002903.
25. Poh MZ, Poh YC, Chan PH, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart* 2018; 104: 1921-1928. 2018/06/02. DOI: 10.1136/heartjnl-2018-313147.
26. Kamaleswaran R, Mahajan R and Akbilgic O. A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. *Physiol Meas* 2018; 39: 035006. 2018/01/26. DOI: 10.1088/1361-6579/aaaa9d.
27. Kwon S, Hong J, Choi EK, et al. Deep Learning Approaches to Detect Atrial Fibrillation Using Photoplethysmographic Signals: Algorithms Development Study. *JMIR Mhealth Uhealth* 2019; 7: e12770. 2019/06/15. DOI: 10.2196/12770.
28. Ebrahimzadeh E, Kalantari M, Joulani M, et al. Prediction of paroxysmal Atrial Fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on HRV signal. *Comput Methods Programs Biomed* 2018; 165: 53-67. 2018/10/20. DOI: 10.1016/j.cmpb.2018.07.014.
29. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet (London, England)* 2019; 394: 861-867. 2019/08/01. DOI: 10.1016/S0140-6736(19)31721-0.
30. Kirchhof P, Benussi S, Kotecha D, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2016; 37: 2893-2962. 2016/08/27. DOI: 10.1093/eurheartj/ehw210.
31. Hobbs FDR, Fitzmaurice DA, Mant J, et al. A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study. *Health Technol Assess* 2005; 9: iii-74. DOI: 10.3310/hta9400.
32. Taggar JS, Coleman T, Lewis S, et al. Accuracy of methods for detecting an irregular pulse and suspected atrial fibrillation: A systematic review and meta-analysis. *European journal of preventive cardiology* 2016; 23: 1330-1338. 2015/10/13. DOI: 10.1177/2047487315611347.
33. Petryszyn P, Niewinski P, Staniak A, et al. Effectiveness of screening for atrial fibrillation and its determinants. A meta-analysis. *PLOS ONE* 2019; 14: e0213198. DOI: 10.1371/journal.pone.0213198.
34. Hart RG, Pearce LA and Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Annals of internal medicine* 2007; 146: 857-867. DOI: 10.7326/0003-4819-146-12-200706190-00007.
35. Hill NR, Sandler B, Mokgokong R, et al. Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm. *J Med Econ* 2020; 23: 386-393. 2019/12/20. DOI: 10.1080/13696998.2019.1706543.

36. Proietti M, Farcomeni A, Goethals P, et al. Cost-effectiveness and screening performance of ECG handheld machine in a population screening programme: The Belgian Heart Rhythm Week screening programme. *Eur J Prev Cardiol* 2019; 26: 964-972. 2019/04/03. DOI: 10.1177/2047487319839184.
37. Halcox JPJ, Wareham K, Cardew A, et al. Assessment of Remote Heart Rhythm Sampling Using the AliveCor Heart Monitor to Screen for Atrial Fibrillation. *Circulation* 2017; 136: 1784-1794. DOI: doi:10.1161/CIRCULATIONAHA.117.030583.

TABLES

Table 1. Patient characteristics

		All patients	AF cohort	Non-AF cohort	p-value
		N=604,135	N=17,880	N=586,255	
Baseline demographic characteristics					
Age (years), mean (SD)		52.2 (13.3)	68.6 (11.3)	51.7 (13.0)	<0.001
Sex, n (%)	Male	295,861 (49.0)	9,717 (54.3)	286,144 (48.8)	<0.001
	Female	308,268 (51.0)	8136 (45.7)	300,105 (51.2)	<0.001
	Unknown†	<5 (<0.1)	0 (0)	<5 (<0.1)	NA
Smoking status, n (%)	Current	113,364 (18.8)	2,230 (12.5)	111,134 (19.0)	<0.001
	Former	96,327 (15.9)	4,697 (26.3)	91,630 (15.6)	<0.001
	Passive†	173 (<0.1)	<5 (<0.1)	167 (<0.1)	NA
	Non-smoker	369,342 (61.1)	10398 (58.2)	358,944 (61.2)	<0.001
	Unknown	2,492 (4.1)	549 (3.1)	24,380 (4.2)	<0.001
Clinical histories* n (%)					
Hypertension		108,296 (17.9)	8,137 (45.5)	100,159 (17.1)	<0.001
Heart failure		2,909 (0.5)	608 (3.4)	2,301 (0.4)	<0.001
Left ventricular hypertrophy		1,595 (0.3)	183 (1.0)	1,412 (0.2)	0.0024
Myocardial infraction		5,555 (0.9)	549 (3.1)	5,006 (0.9)	<0.001
Coronary heart disease		29,589 (4.9)	3,196 (17.9)	26,393 (4.5)	<0.001
Congenital heart disease		63 (<0.1)	9 (0.1)	54 (0.0)	0.7163
Type 1 diabetes		93,459 (15.5)	1,878 (10.5)	91,581 (15.6)	<0.001

	All patients	AF cohort	Non-AF cohort	p-value
	N=604,135	N=17,880	N=586,255	
Type 2 diabetes	46,392 (7.7)	2,688 (15.0)	43,704 (7.5)	<0.001
Clinical measurements, mean (SD)				
Weight (kg)	75.7 (17.1)	80.4 (19.0)	75.6 (17.0)	<0.001
Height (m)	1.7 (0.1)	1.7 (0.1)	1.7 (0.1)	1.000
BMI (kg/m^2)	27.0 (6.1)	28.4 (6.6)	27.0 (6.1)	<0.001
SBP (mmHg)	130.1 (18.3)	139.0 (18.9)	129.8 (18.2)	<0.001
DBP (mmHg)	79.0 (10.8)	79.1 (11.1)	79.0 (10.8)	0.3392

*Clinical histories up to 5 years before index date

†p-values not shown due to small numbers

AF: atrial fibrillation; BMI: Body mass index; CPRD: Clinical Practice Research Datalink; DBP: Diastolic blood pressure; NA: Not applicable; SBP: Systolic blood pressure.

Table 2. Comparison of patients in development (CPRD) and validation (DISCOVER) datasets

		All patients			AF cohort		
		DISCOVER	**CPRD**	**p-value**	**DISCOVER**	**CPRD**	**p-value**
Baseline demographic characteristics							
Age (years), mean (SD)		52.2 (13.3)	55.98 (14.46)	<0.0001	68.6 (11.3)	70.23 (11.07)	<0.001
Sex, n (%)	Male	295,861 (49.0)	1,395,397 (46.6)	<0.0001	9,717 (54.3)	51,738 (54.1)	0.6223
Smoking status, n (%)	Current	11,3364 (18.8)	555,074 (18.5)	<0.0001	2,230 (12.5)	10,571 (11.1)	<0.001
	Former	96,327 (15.9)	701,966 (23.4)	<0.0001	4,697 (26.3)	32,198 (33.7)	<0.001
	Passive	173 (<0.1)	7,876 (0.3)	NA	<5 (<0.1)	279 (<0.5)	NA
	Non-smoker	369,342 (61.1)	1,269,538 (42.4)	<0.0001	10,398 (58.2)	37,384 (39.1)	<0.001
	Unknown	24,929 (4.1)	460,383 (15.4)	<0.0001	549 (3.1)	15,175 (15.9)	<0.001
Clinical histories* n (%)							
Hypertension		108,296 (17.9)	748,849 (25.0)	<0.0001	8,137 (45.5)	50,501 (52.8)	<0.0001
Heart failure		2,909 (0.5)	22,054 (0.7)	<0.0001	608 (3.4)	2,805 (2.9)	0.0003
Left ventricular hypertrophy		1,595 (0.3)	4,727 (0.2)	<0.0001	183 (1.0)	502 (0.5)	<0.001
Myocardial infraction		5,555 (0.9)	42,830 (1.4)	<0.0001	549 (3.1)	3,009 (3.1)	1.0000
Coronary heart disease		29,589 (4.9)	154,029 (5.1)	<0.0001	3196 (17.9)	13,703 (14.3)	<0.001
Congenital heart disease†		63 (<0.5)	501 (<0.1)	NA	9 (0.1)	58 (<0.5)	NA

	All patients			AF cohort		
	DISCOVER	**CPRD**	**p-value**	**DISCOVER**	**CPRD**	**p-value**
Type 1 diabetes†	93,459 (15.5)	19,101 (0.6)	NA	1,878 (10.5)	831 (0.9)	NA
Type 2 diabetes	46,392 (7.7)	187,733 (6.3)	<0.0001	2,688 (15.0)	10,727 (11.2)	<0.0001
Clinical measurements, mean (SD)						
Weight (kg)	75.7 (17.1)	78.32 (18.3)	<0.0001	80.4 (19.0)	81.55 (19.5)	<0.0001
Height (m)	1.7 (0.1)	1.68 (0.1)	1.000	1.7 (0.1)	1.69 (0.1)	1.000
BMI (kg/m^2)	27.0 (6.1)	27.59 (6.0)	<0.0001	28.4 (6.6)	28.56 (6.2)	0.0001
SBP (mmHg)	130.1 (18.3)	133.58 (18.9)	<0.0001	139.0 (18.9)	140.97 (19.3)	<0.0001
DBP (mmHg)	79.0 (10.8)	79.40 (10.9)	<0.0001	79.1 (11.1)	79.12 (11.0)	1.000

*Clinical histories up to 5 years before index date

†p-values not shown due to small numbers

AF: atrial fibrillation; BMI: Body mass index; CPRD: Clinical Practice Research Datalink; DBP: Diastolic blood pressure; NA: Not applicable; SBP: Systolic blood pressure.

Table 3. Assessment of model performance at 75% and 50% sensitivities in all patients aged ≥30 years

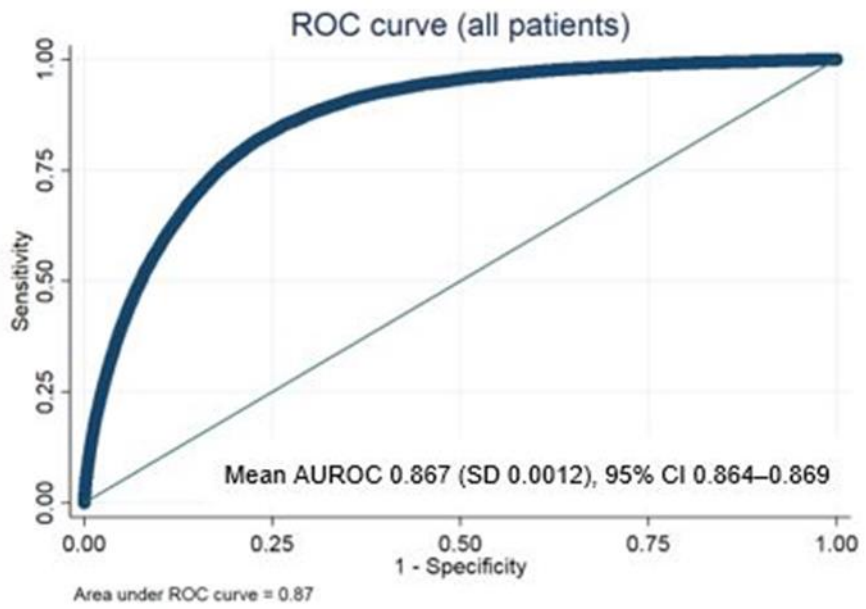
Study	**Sensitivity**	**AF risk threshold**	**Specificity**	**PPV**	**NPV**	**Potential NNS**	**AUROC**
DISCOVER	50%	5.5%	92.6%	16.9%	98.4%	6	0.87
	75%	2.3%	82.0%	11.3%	99.1%	9	
CPRD	50%	7.4%	90.0%	18.3%	97.6%	5	0.83

	75%	Not published	74.9%	11.5%	98.5%	9	

AF: atrial fibrillation; AUROC: area under the receiver operating characteristic curve; CPRD: Clinical Practice Research Datalink; NNS: number needed to screen (to identify one AF case); NPV: negative predictive value (percentage of screened patients not diagnosed with AF); PPV: positive predictive value (percentage of screened patients diagnosed with AF).

Supplementary material for Detecting undiagnosed atrial fibrillation in UK primary care: validation of a machine learning prediction algorithm in a retrospective cohort study (EJPC-D-20-00602) by Sekelj et al in European Journal of Preventive Cardiology

A



B