Prediction of incident atrial fibrillation in community-based electronic health records: a systematic review with meta-analysis

Ramesh Nadarajah 1,2, Eman Alsaeed, 1 Ben Hurdus 1,2, Suleman Aktaa, 1,2, David Hogg 3, Matthew G D Bates, 4, Campbell Cowan, 2, Jianhua Wu 1,5, Chris P Gale 1,2

ABSTRACT

Objective Atrial fibrillation (AF) is common and is associated with an increased risk of stroke. We aimed to systematically review and meta-analyse multivariable prediction models derived and/or validated in electronic health records (EHRs) and/or administrative claims databases for the prediction of incident AF in the community.

Methods Ovid Medline and Ovid Embase were searched for records from inception to 23 March 2021. Measures of discrimination were extracted and pooled by Bayesian meta-analysis, with heterogeneity assessed through a 95% prediction interval (PI). Risk of bias was assessed using Prediction model Risk Of Bias ASsessment Tool and certainty in effect estimates by Grading of Recommendations, Assessment, Development and Evaluation.

Results Eleven studies met inclusion criteria, describing nine prediction models, with four eligible for meta-analysis including 9 289 959 patients. The CHADS (Congestive heart failure, Hypertension, Age≥75, Diabetes mellitus, prior Stroke or transient ischemic attack) (summary c-statistic 0.674; 95% CI 0.610 to 0.732; 95% PI 0.513–0.803), CHA2DS2-Vasc (Congestive heart failure, Hypertension, Age≥75 (2 points), Stroke/transient ischemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category) (summary c-statistic 0.679; 95% CI 0.620 to 0.736; 95% PI 0.531–0.811) and HATCH (Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, Heart failure) (summary c-statistic 0.669; 95% CI 0.600 to 0.732; 95% PI 0.513–0.803) models resulted in a c-statistic with a statistically significant 95% PI and moderate discriminative performance. No model met eligibility for inclusion in meta-analysis if studies at high risk of bias were excluded and certainty of effect estimates was ‘low’. Models derived by machine learning demonstrated strong discriminative performance, but lacked rigorous external validation.

Conclusions Models externally validated for prediction of incident AF in community-based EHR demonstrate moderate predictive ability and high risk of bias. Novel methods may provide stronger discriminative performance. We performed this systematic review and meta-analysis with a number of aims. First, to identify prediction models for incident AF derived or validated in community-based EHR. Second, to summarise the performance of individual prediction models to understand if any would be suitable for use in targeted screening. Third, to summarise the methods by which prediction models have been developed in EHR to inform research within the field.

METHODS

We reported this systematic review and meta-analysis in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (online supplemental material).11

Search strategy and inclusion criteria

The research question was framed using the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (online supplemental material).12
through the Ovid platform from inception through 23 March 2021. We used a combination of keywords and subject headings related to AF, prediction models and EHR based on previous literature. The search was limited to the English language and to human studies. The full search strategy is provided in online supplemental material. We manually searched the reference lists of included studies and previous systematic reviews. Duplicates were removed using Endnote’s duplicate identification strategy and then manually.

To be eligible for inclusion a study had to:
1. Be an original study in human adults (≥18 years of age).
2. Develop and/or validate a prediction model(s) for incident AF or atrial flutter (AFl) based on multivariable analysis in a community-based EHR. We included AFl as a co-outcome because it has a similar indication for anticoagulation.
3. Be written in English.

Articles were excluded if they:
1. Included patients with AF or AFl at baseline.
2. Only reported measures of association between risk factors and incident AF rather than a full prediction model.
3. Studied only a subset of the general population, for example, individuals diagnosed with a particular morbidity.
4. Incorporated variables that would not be routinely available in community-based EHR (eg, ECG parameters) (online supplemental material).

We uploaded records to a systematic review web application (Rayyan, Qatar Computing Research Institute). Four investigators (RN, EA, BH and SA) independently screened them for inclusion by title, abstract, full text and supplementary materials. Disagreements were resolved by consultation with a fifth investigator (JW).

Data extraction and quality assessment
Two investigators (RN and EA) independently extracted the data from the included studies based on CHARMS. This included the following domains: data source, participants, outcome(s),

Figure 1 Flow diagram of literature search. AF, atrial fibrillation; AFl, atrial flutter; EHR, electronic health record.
### Table 1: Characteristics of included studies

| Study            | Cohort (country) | Study aim | EHR description | AF cases (n)/Total patients (n) (%) | Age (mean±SD) | Female (%) | BMI (mean±SD) | Diabetes (%) | Hypertension (%) | Heart failure (%) | Outcome definition | Outcome coding | Enrollment period (mean F/U in years) | Exclusion criteria |
|------------------|------------------|-----------|-----------------|-------------------------------------|----------------|------------|---------------|--------------|-----------------|-------------------|-------------------|-----------------|------------------|-------------------|--------------------|
| Aronson et al    | MHS (IL)         | D         | Ambulatory clinics | 566/206 770 (5.80) | 62.0±9.0 | 53.7      | 28.2±5.1      | 13.5         | 34.3            | 1.00              | AF, AFI           | ICD codes       | 2005–2015 (10.0) | Past history of AF, incomplete follow-up |
| Chao et al       | NHIRD (TW)       | EV        | National health insurance | 918/3702 592 (1.30) | 41.3±16.4 | 49.1      | N/A           | 3.1          | 5.2             | 0.40              | AF               | ICD codes        | 2001–2009 (8.0)   | Age <18 years, past history of cardiac arrhythmia, rheumatic heart disease |
| Hill et al       | CPRD (UK)        | Q, EV     | Nationwide primary care | 95 607/2948 837 (3.19) | 56.0±14.5 | 53.4      | 27.6±6.0      | 6.9          | 25.0            | 0.70              | AF, AFI           | EHR codes | 2006–2016 (NA)  | N/A               |
| Chao et al       | NHIRD (TW)       | EV        | National health insurance | 918/3702 592 (1.30) | 41.3±16.4 | 49.1      | N/A           | 3.1          | 5.2             | 0.40              | AF               | ICD codes        | 2001–2009 (8.0)   | Age <18 years, past history of AF, incomplete data |
| Hill et al       | CPRD (UK)        | Q, EV     | Nationwide primary care | 95 607/2948 837 (3.19) | 56.0±14.5 | 53.4      | 27.6±6.0      | 6.9          | 25.0            | 0.70              | AF, AFI           | EHR codes | 2006–2016 (NA)  | N/A               |
| Hu et al         | NHIRD (TW)       | D         | National health insurance | 918/3702 592 (1.30) | 41.3±16.4 | 49.1      | N/A           | 3.1          | 5.2             | 0.40              | AF               | ICD codes        | 2001–2009 (8.0)   | Age <18 years, past history of AF, incomplete data |
| Hill et al       | CPRD (UK)        | Q, EV     | Nationwide primary care | 95 607/2948 837 (3.19) | 56.0±14.5 | 53.4      | 27.6±6.0      | 6.9          | 25.0            | 0.70              | AF, AFI           | EHR codes | 2006–2016 (NA)  | N/A               |
| Chao et al       | NHIRD (TW)       | EV        | National health insurance | 918/3702 592 (1.30) | 41.3±16.4 | 49.1      | N/A           | 3.1          | 5.2             | 0.40              | AF               | ICD codes        | 2001–2009 (8.0)   | Age <18 years, past history of AF, incomplete data |
| Hill et al       | CPRD (UK)        | Q, EV     | Nationwide primary care | 95 607/2948 837 (3.19) | 56.0±14.5 | 53.4      | 27.6±6.0      | 6.9          | 25.0            | 0.70              | AF, AFI           | EHR codes | 2006–2016 (NA)  | N/A               |
| Chao et al       | NHIRD (TW)       | EV        | National health insurance | 918/3702 592 (1.30) | 41.3±16.4 | 49.1      | N/A           | 3.1          | 5.2             | 0.40              | AF               | ICD codes        | 2001–2009 (8.0)   | Age <18 years, past history of AF, incomplete data |

In Chao et al it is not reported how outcome was defined or measured but given the authors were using the same database as Suenari, we have assumed outcomes were measured in the same way.

*In Kim and Li and Hu W-S the percentage of patients related to sex, diabetes, hypertension, heart failure was calculated from reported values categorized by incident AF or not.

AF, atrial fibrillation; AFI, atrial flutter; BMI, body mass index; BP, blood pressure; ClalitHS, Clalit Health Services; CPRD, Clinical Practice Research Datalink; D, derivation; EHR, electronic health records; EV, external validation; F/U, follow-up; ICD, International Classification of Diseases; ICPC-1, International classification of Primary care version 1 diagnostic codes; IL, Israel; KR, Republic of Korea; MHS, Maccabi Healthcare Services; N/A, not available; NHIRD, National Health Insurance Research Database; NHIS-HEALS, National Health Insurance Service - Health screening Cohort; NHIS-NSC, National Health Insurance Service-based National Sample Cohort; Nivel-PCD, Netherlands Institute for Health Services Research Primary Care Database; NL, Netherlands; TW, Taiwan; YMID, Yunnan Medical Insurance Database.
candidate predictors, sample size, missing data, and model development, performance and evaluation. Discrepancies were resolved with a third investigator (JW). All data came from the primary reference, unless otherwise stated.

To allow quantitative synthesis of the predictive performance of the models we extracted measures of discrimination and calibration.\textsuperscript{13} Discrimination quantifies the model’s ability to distinguish between individuals developing or not developing the outcome. We extracted data on the c-statistic (c-statistic=1 if the model discriminates perfectly, c-statistic=0.5 if discrimination no better than chance) or area under the receiver operating characteristic (AUROC) and corresponding 95% confidence interval (95% CI).

We reported continuous variables as means±SD and categorical variables as percentages. Calibration was infrequently reported, so we restricted meta-analysis to discrimination through a summary measure of c-statistic and corresponding 95% CI. In our primary analysis we assessed overall discrimination for models that had ≥3 EHR cohorts with c-statistic data. When multiple c-statistic data for a model were reported in a single cohort by different studies we only included the first published study.

We calculated the 95% prediction interval (PI) to depict the extent of between-study heterogeneity and to indicate a possible range for prediction model performance in a new validation.\textsuperscript{18} When the 95% CI or PI of the summary c-statistic included 0.5 we concluded that there was insufficient evidence that the prediction model has statistically significant discriminatory ability.\textsuperscript{13,20} We used a Bayesian approach throughout as frequentist methods, where there are fewer studies or a mixture of study sizes, have produced PIs with poor coverage.\textsuperscript{19} The prior distributions specified are summarised in online supplemental material.

Data synthesis and statistical analysis
We reported continuous variables as means±SD and categorical variables as percentages. Calibration was infrequently reported, so we restricted meta-analysis to discrimination through a summary measure of c-statistic and corresponding 95% CI. In our primary analysis we assessed overall discrimination for models that had ≥3 EHR cohorts with c-statistic data. When multiple c-statistic data for a model were reported in a single cohort by different studies we only included the first published study.

We calculated the 95% prediction interval (PI) to depict the extent of between-study heterogeneity and to indicate a possible range for prediction model performance in a new validation. When the 95% CI or PI of the summary c-statistic included 0.5 we concluded that there was insufficient evidence that the prediction model has statistically significant discriminatory ability.\textsuperscript{13,20} We used a Bayesian approach throughout as frequentist methods, where there are fewer studies or a mixture of study sizes, have produced PIs with poor coverage. The prior distributions specified are summarised in online supplemental material. A logit transformation was applied to the c-statistic prior to meta-analysis, as the between-study distribution of the c-statistic is often skewed.\textsuperscript{21} We conducted meta-analyses in R using the metafor and metan package (R foundation for Statistical Computing V.3.6.3).\textsuperscript{22–24} We performed a number of sensitivity analyses:

<table>
<thead>
<tr>
<th>Table 2 Characteristics of included prediction models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>Models originally derived for another purpose but tested for prediction of incident AF</td>
</tr>
<tr>
<td>CHADS\textsubscript{2}</td>
</tr>
<tr>
<td>CHA\textsubscript{DS}-VASC</td>
</tr>
<tr>
<td>HATCH</td>
</tr>
<tr>
<td>Regression model derived in a prospective cohort design</td>
</tr>
<tr>
<td>CHARGE-AF</td>
</tr>
<tr>
<td>Regression models derived in EHR</td>
</tr>
<tr>
<td>C\textsubscript{2}HEST</td>
</tr>
<tr>
<td>MHS</td>
</tr>
<tr>
<td>Machine learning models derived in EHR</td>
</tr>
<tr>
<td>CPRD</td>
</tr>
<tr>
<td>NHIRD</td>
</tr>
<tr>
<td>NHIS-NSC</td>
</tr>
</tbody>
</table>

AF, atrial fibrillation; AFI, atrial flutter; CHADS\textsubscript{2}, Congestive heart failure, Hypertension, Age>75, Diabetes mellitus, prior Stroke or transient ischaemic attack [two points]; CHA\textsubscript{DS}-VASC, Congestive heart failure, Hypertension, Age>75 [2 points], Stroke/transient ischaemic attack/thromboembolism [two points], Vascular disease, Age 65–74, Sex Category; CHARGE-AF; Cohorts for Heart and Ageing Research in Genomic Epidemiology; C\textsubscript{2}HEST, Coronary artery disease/Chronic obstructive pulmonary disease [one point each], Hypertension, Elderly (Age>75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); ClalitHS, Clalit Health Services; CN, China; CPRD, Clinical Practice Research Datalink; EHR, electronic health records; HATCH, Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, Heart failure; IL, Israel; KR, Republic of Korea; MHS, Maccabi Healthcare Services; N/A, not available; NHIRD, National Health Insurance Research Database; NHIS-HEALS, National Health Insurance Service - Health screening Cohort; NHIS-NSC, National Health Insurance Service-based National Sample Cohort; Nivel-PCD, Netherlands Institute for Health Services Research Primary Care Database; NL, Netherlands; TW, Taiwan; UK, United Kingdom; YIMID, Yunnan Medical Insurance Database.
1. To only include studies where the participants’ domain in PROBAST assessment was ‘low’ or ‘unclear’ risk of bias.
2. To only include studies where the overall PROBAST assessment was ‘low’ or ‘unclear’ risk of bias.
3. Where a cohort had been reported multiple times we replaced the meta-analysis data with the data on the same cohort from any later study.
4. We excluded data from one of the Korean National Health Insurance Service Health screening cohort (NHIS-HEALS) and Korean National Health Insurance Service-based National Sample cohort (NHIS-NSC) because they originated from the same EHR database.

The Grading of Recommendations, Assessment, Development and Evaluation approach was used to assess the certainty of the evidence. The certainty of the evidence was graded as ‘high’, ‘moderate’, ‘low’ or ‘very low’. One investigator (RN) rated the certainty of the evidence for the primary outcome and this was checked by a second investigator (JW). The criteria used are summarised in online supplemental material.

Patient and public involvement
Patients or the public were not involved in the design, conduct, reporting or dissemination plans of our research.

RESULTS
Study selection
The study selection process is described in figure 1. We identified 3949 unique records, reviewed 102 full-text reports and included 11 studies. A list of excluded studies that met a number of the inclusion criteria is available in online supplemental material.

Characteristics of included studies
The 11 included studies were based on nine cohorts from eight EHR databases, located in Asia Pacific (n=3), Europe (n=3) and the Middle East (n=2) (table 1). The number of times a prediction model had been derived or validated in EHR was skewed to Asia Pacific (n=17) compared with Europe (n=5) and the Middle East (n=3) (table 2).

The total number of participants in the included studies was 17,889,536. Cohort size ranged from 96,778 to 2,994,837. The mean age varied from 41.3 years to 65.7 years and the proportion of female participants ranged from 47.3% to 54.7%. The mean follow-up ranged from 2.9 years to 10.9 years. The incidence of AF during follow-up ranged from 0.2% to 5.8%.

Characteristics of included prediction models
The included studies reported data on nine multivariable prediction models (table 2). Three models had originally been derived for a purpose other than incident AF prediction. Five models had been derived in community-based EHR; three using machine learning techniques. In two of these studies, a range of machine learning techniques had been investigated with the optimum technique chosen by discriminative performance (online supplemental table S2). Among machine learning techniques, random forests were investigated in all three
Arrhythmias and sudden death

Studies and neural networks were considered in two. All studies reported a measure of discrimination (either c-statistic or AUROC), but only two studies provided a measure of calibration. Three prediction models—CPRD (Clinical Practice Research Datalink), CHEST (Coronary artery disease/Chronic obstructive pulmonary disease (one point each), Hypertension, Elderly (Age ≥75, two points), Systolic heart failure, Thyroid disease (hyperthyroidism)) and HATCH (Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, Heart failure)—showed a c-statistic greater than 0.75 in an external validation study (online supplemental table S1). Online supplemental table 3a,b summarises the variables used. The 10 most frequently included variables are summarised in figure 2. Age and chronic heart failure were the only variables included in every model. The number of variables incorporated into machine learning models was far greater than traditional regression models (table 2). The CPRD model was unique in incorporating time-varying variables (eg, change in body mass index (BMI) between the last two quarters of the year). Online supplemental figure S1 plots the performance of traditional regression and machine learning models in the development population of each study. Online supplemental table S2 summarises the performance of traditional regression and machine learning techniques during model development in the CPRD and NHIS-NSC data sets. In each case, machine learning produced stronger discriminative performance in the development population.

Risk of bias assessment

Online supplemental table S4 shows the results of the risk of bias and applicability assessment for each PROBAST domain for each model in the included studies. Figure 3 gives an overall summary of PROBAST domain assessments across all included studies. Overall, 96% of model results were at high risk of bias predominantly driven by high risk of bias in the analysis domain (88%). This resulted from exclusion of participants with missing data from analysis (72%) or not mentioning missing data (16%).
Figure 4  Forest plot of primary analysis of c-statistics. C2HEST, Coronary artery disease/chronic obstructive pulmonary disease (one point each), Hypertension, Elderly (Age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); CHADS2, Congestive Heart failure, hypertension, Age >75, Diabetes mellitus, prior Stroke or transient ischaemic attack (two points); CHADDS2-VASc, Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (two points), Vascular disease, Age 65–74, sex category; ClalitHS, Clalit health services; HATCH, Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, and Heart failure; NHIRD, National Health Insurance Research Database; NHIS-HEALS, National Health Insurance Service - Health screening Cohort; NHIS-NSC, National Health Insurance Service-based National Sample Cohort; Nivel-PCD, Netherlands Institute for Health Services Research Primary Care Database; YMID, Yunnan Medical Insurance Database.

Meta-analysis

Four models were eligible for the primary meta-analysis, incorporating 9 289 959 patients (figure 4). Only C2HEST was derived specifically for the purpose of predicting incident AF.31 There were three models that resulted in a summary c-statistic with statistically significant 95% PI in our primary meta-analysis: CHADS2 (Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, prior Stroke or transient ischemic attack) (summary c-statistic 0.674; 95% CI 0.610 to 0.732; 95% PI 0.526–0.815; n=5 studies; n=3 119 807), CHADDS2-VASc (Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex Category) (summary c-statistic 0.679; 95% CI 0.620 to 0.736; 95% PI 0.531–0.811; n=5 studies; n=3 119 807), and ClalitHS (Congestive heart failure, Hypertension, Age >75, 2 points), Stroke/transient ischemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex Category) (summary c-statistic 0.679; 95% CI 0.620 to 0.736; 95% PI 0.531–0.811; n=5 studies; n=3 119 807).

Online supplemental table S5 shows the results of the sensitivity analyses. Only CHADDS2-VASc maintained a summary c-statistic with statistically significant 95% PI when either restricting the primary analysis to studies with ‘low’ or ‘unclear’ risk of bias for the participants domain of PROBAST, or using later data when a cohort had been analysed multiple times, or excluding data from either of the NHIS-HEALS or NHIS-NSC cohorts. However, when restricting primary analysis to models with ‘low’ or ‘unclear’ risk of bias for overall PROBAST assessment, no models met eligibility for inclusion.

Certainty of evidence

The initial certainty level of the included prediction modelling studies was set at ‘high’ because the association between the predictors and outcomes was considered irrespective of any causal connection.39 The overall certainty level was, however, downgraded to ‘moderate’ and then ‘low’ because of inconsistent results (high heterogeneity) and the large proportion of high overall risk of bias amongst studies. The final overall certainty of ‘low’ implies that our confidence in the effect estimates is limited and further research is very likely to change the effect estimate.
**DISCUSSION**

This systematic review and meta-analysis identified nine models that have been derived and/or validated in community-based EHR for incident AF. Five had been derived in EHR for this purpose; three by machine learning methods. Three models (CHADS\textsubscript{2}, CHA\textsubscript{2}DS\textsubscript{2}-VASc and HATCH) produced a summary c-statistic with statistically significant 95% PI for prediction of incident AF despite high heterogeneity. However the summary c-statistics were only 0.669–0.679. For an outcome such as AF that is considered difficult to predict, a c-statistic of 0.75 may be adequate for the models to be useful.\textsuperscript{40} This threshold has been achieved by prediction models for incident AF in the community in non-EHR-based external validation studies,\textsuperscript{41–43} as well as in EHR by the machine learning CPRD model.\textsuperscript{33} Furthermore, in sensitivity analyses no model met eligibility for inclusion in meta-analysis if studies at overall high risk of bias were excluded.

A previous meta-analysis investigated prediction models for incident AF that had been derived or validated in community cohorts.\textsuperscript{13} Nevertheless, this review included predominantly carefully curated prospective cohort designs, the results from which will have limited generalisability. In addition, a number of the included models require variables, such as ECG parameters, that are not routinely available in community-based EHR.\textsuperscript{44} The authors found CHA\textsubscript{2}DS\textsubscript{2}-VASc and CHARGE-AF (Cohorts for Heart and Ageing Research in Genomic Epidemiology) resulted in a summary c-statistic with statistically significant 95% PI on meta-analysis. There is conflicting evidence as to how well CHARGE-AF performs in EHR, especially given the incompleteness of structured EHR fields for height, weight and ethnicity,\textsuperscript{9,10} and for our study it did not meet eligibility for inclusion into meta-analysis. Another systematic review summarised a similar selection of prediction models for the detection of AF in the community and externally validated these models head-to-head in a commercial screening cohort.\textsuperscript{14} However, the outcome was prevalence, rather than future incident AF. Both of these reviews predated the emergence of machine learning models in this field, which are summarised for the first time regarding the prediction of incident AF here.

The use of age alone to target screening strategies for incident AF has yet to show a benefit for systematic versus opportunistic screening, which is reflected in international guidelines.\textsuperscript{2,7} Prediction models could target screening and if implemented through primary care EHR would minimise extra resources. The use of CHA\textsubscript{2}DS\textsubscript{2}-VASc for prediction of incident AF has advantages given it uses variables available with high completeness in primary care EHR and would simultaneously provide an assessment of stroke risk as an indicator of eligibility for anticoagulation. Even so, there are a number of limitations. First, the discriminative performance was only moderate, overall certainty in the estimate effects was ‘low’ and the vast majority of studies were at high risk of bias. Second, it has predominantly been validated in Asia Pacific countries, where cohorts had different baseline characteristics compared with European counterparts. Third, it was outperformed by CHARGE-AF and C\textsubscript{HEST} when compared head-to-head in individual external validation studies.\textsuperscript{9,31}

Efforts may be best served to develop and externally validate novel prediction models for incident AF in community-based EHR. These data sources offer large samples sizes, providing the opportunity to investigate a larger number of predictors and use novel techniques. Machine learning models in this review showed strong discriminative performance in development data sets but were not included in meta-analysis due to a sparsity of external validation.

This study has a number of strengths. We had a comprehensive search strategy and thorough analysis approach. We included any model that had been used to predict the risk of incident AF, which allowed us to include models that were not originally intended for predicting AF but may have merits. We only included models that had been tested in databases relevant to the general population, which ensures the applicability of our results for screening in a primary care setting. We also did not present meta-regression or subgroup meta-analysis to investigate heterogeneity between studies based on study-level characteristics or subgroups in the absence of available individual patient data given that such analyses would be prone to ecological bias.\textsuperscript{45}

There are limitations to our study. Meta-analysis of model calibration performance was prohibited by poor reporting. We did not assess for ‘reporting biases’ visually through a funnel plot for several reasons. First, some studies reported multiple models in the same cohort so incorporating all these data points would skew the plot; second, producing funnel plots for individual models would not be informative as there would be too few data points; third the sample sizes for all included studies was very large making small-study effects less likely. The vast majority of studies was at high risk of bias, which is consistent with previous literature on clinical prediction models due to limitations in conduct and reporting.\textsuperscript{46} We restricted our search to studies written in English, though this has not been found to lead to significant bias.\textsuperscript{47} Finally,
CONCLUSIONS

In this systematic review with meta-analysis, we identified nine multivariable prediction models relevant to screening for incident AF using community-based EHR. On meta-analysis three models produced a summary c-statistic with statistically significant 95% PI, but discriminative performance was only moderate. At present, due to a combination of high risk of bias and inconsistency, there is no high-performing prediction model for incident AF using primary care EHR. Future research could aim to develop models in primary care EHR using machine learning, but must better handle missing data, report calibration and provide external validation.

REFERENCES


Arrhythmias and sudden death


