Accuracy of methods for diagnosing atrial fibrillation using 12-lead ECG: A systematic review and meta-analysis

Jaspal S Taggar¹, Tim Coleman², Sarah Lewis³, Carl Heneghan⁴, Matthew Jones⁵

¹ University of Nottingham; This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

² University of Nottingham; This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

³ University of Nottingham; This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

⁴ University of Oxford; This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

⁵ University of Nottingham; This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

Address for correspondence:

Dr Jaspal Taggar
Division of Primary Care,
University of Nottingham,
Room C34, Medical School,
Queen’s Medical Centre,
Nottingham,
NG7 2UH
Email: Jaspal.taggar@nottingham.ac.uk
Tel: +44 115 8230462
Fax: +44 115 82 30214

Funding:

The study was supported by an NIHR/School for Primary Care Research GP Career Progression Fellowship and Research Capability Funding from Nottingham City Clinical Commissioning Group.

The funding bodies had no role in the design, conduct, interpretation and writing of the report for this research.

Views and opinions expressed in this article are those of the authors and do not necessarily reflect those of the National Institute for Health Research (NIHR) or Nottingham City Clinical Commissioning Group (CCG).

Conflicts of interest:

CH receives payment for running educational courses at the University of Oxford and University of Oxford ISIS consulting services for external teaching and training. He also receives royalties for books (Evidence Based Toolkit series by Blackwell BMJ Books). All other authors report no relationships that could be construed as a conflict of interest.

Keywords: Atrial fibrillation; electrocardiogram; diagnostic accuracy

1
ABSTRACT

Background: Screening for Atrial fibrillation (AF) using 12-lead-electrocardiograms (ECG) has been recommended; however, the best method for interpreting ECGs to diagnose AF is not known. We compared accuracy of methods for diagnosing AF from ECGs.

Methods: We searched MEDLINE, EMBASE, CINAHL and LILACS until March 24, 2014. Two reviewers identified eligible studies, extracted data and appraised quality using the QUADAS-2 instrument. Meta-analysis, using the bivariate hierarchical random effects method, determined average operating points for sensitivities, specificities, positive and negative likelihood ratios (PLR, NLR) and enabled construction of Summary Receiver Operating Characteristic (SROC) plots.

Results: 10 studies investigated 16 methods for interpreting ECGs (n=55,376 participant ECGs). The sensitivity and specificity of automated software (8 studies; 9 methods) were 0.89 (95% CI 0.82-0.93) and 0.99 (95% CI 0.99-0.99), respectively; PLR 96.6 (95% C.I 64.2-145.6); NLR 0.11 (95% C.I 0.07-0.18). Indirect comparisons with software found healthcare professionals (5 studies; 7 methods) had similar sensitivity for diagnosing AF but lower specificity [sensitivity 0.92 (95% CI 0.81-0.97), specificity 0.93 (95% CI 0.76-0.98), PLR 13.9 (95% C.I 3.5-55.3), NLR 0.09 (95% C.I 0.03-0.22). Sub-group analyses of primary care professionals found greater specificity for GPs than nurses [GPs: sensitivity 0.91 (95% C.I 0.68-1.00); specificity 0.96 (95% C.I 0.89-1.00). Nurses: sensitivity 0.88 (95% C.I 0.63-1.00); specificity 0.85 (95% C.I 0.83-0.87)].

Conclusions: Automated ECG-interpreting software most accurately excluded AF, although its ability to diagnose this was similar to all healthcare professionals. Within primary care, the specificity of AF diagnosis from ECG was greater for GPs than nurses.
BACKGROUND

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia and has a prevalence that increases with age.\textsuperscript{1} AF is associated with significant morbidity, mortality and impaired quality of life, most notably from its associated four to five fold increased risk of ischaemic stroke, and poses a significant public health burden.\textsuperscript{2, 3}

Screening for AF in primary care has been found to be an effective strategy for detecting AF.\textsuperscript{4, 5} The SAFE trial, set in primary care, found opportunistic pulse palpation and confirmatory 12-lead electrocardiogram (ECG) significantly increased the detection of incident AF cases when compared to routine practice in patients over 65 years old.\textsuperscript{4} Consequently, this approach to AF screening, combined with the appropriate provision of antithrombotic therapy, has been proposed as a population intervention to reduce the burden of thromboembolic complications.\textsuperscript{6}

A sub-study of the SAFE trial investigated the accuracy of 12-lead ECG diagnoses of AF made by GPs, nurses and automated software in primary care.\textsuperscript{7} Mant et al. found, as compared to ECG diagnoses made by cardiologists, interpretive software had a significantly greater specificity than all methods for diagnosing AF. However, the sensitivities for GPs, nurses and software for AF diagnosis were substantially lower and similar across all groups, and suggested the accuracy in primary care using any single method was insufficient for screening.

A recent narrative literature review of methods for detecting AF reported substantial variation in the accuracy of ECG diagnoses made by primary care practitioners, although, to date, there has been no systematic evaluation of the evidence for the range of different methods for diagnosing AF using 12-lead ECG. This systematic review therefore aimed to identify the range of methods for interpreting whether or not 12-lead ECGs show AF, and to compare their diagnostic accuracies.
METHODS

Search strategy and selection criteria

This study was conducted in accordance with guidelines and methods for systematic reviews and meta-analyses of diagnostic tests.\textsuperscript{8,11} We searched MEDLINE, EMBASE, Cumulative Index to Nursing & Allied Health (CINAHL) and Latin American and Caribbean Health Sciences Information System (LILACS) in all languages published from inception until 24\textsuperscript{th} March 2014 (See appendix for search terms). Additionally, the reference lists of national guidelines, review articles and eligible studies were hand-searched to identify potential studies. We included all randomised trials and observational studies that i) recruited participants ≥18 years of age, ii) investigated any method for interpreting 12-lead ECGs to show AF (the index test) with 12-lead ECG diagnoses of AF made by a trained specialist (the reference standard), iii) involved healthcare professionals in making AF diagnoses and iv) sufficient data available to enable the calculation of diagnostic accuracy. Studies that investigated invasive or echocardiographic methods for diagnosing AF were excluded, as these methods would not be translatable into routine screening practice. After the removal of duplicate citations, two reviewers (JT and MJ) independently screened citations for relevance and reviewed full-text articles using the predetermined eligibility criteria. Any disagreements were resolved by consensus with a third reviewer (TC).

Data extraction

Two reviewers (JT and MJ) independently extracted data using a pre-specified data extraction form. Any disagreements were resolved by consensus with a third reviewer (TC). The lead authors of studies for which reported data were insufficient to calculate diagnostic accuracy were contacted to ascertain missing data.

Study quality was appraised using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) instrument.\textsuperscript{10,12} Additionally, the studies were graded using the quality scale reported by Van den Bruel et al;\textsuperscript{13} studies were rated as grade A if they fulfilled all QUADAS-2 criteria. Studies were graded D if there was no or unclear verification of the index test findings with the reference standard, or if
the index test results were interpreted un-blinded to the results of the reference test. Studies where there was an unduly long time delay between index and reference test, or where the reference test was not independent of the index test, or where the reference test was interpreted un-blinded to the results of the index test were graded C. Remaining studies which did not fall in to these categories were graded B.

Statistical analysis

We constructed 2x2 contingency tables to enable the calculation of sensitivity and specificity for each method of diagnosing AF as measures of diagnostic accuracy, and we used the bivariate hierarchical random effects method to determine the average operating points for sensitivity and specificity, which enabled construction of Summary Receiver Operating Characteristic (SROC) plots with 95% prediction regions. We also calculated positive likelihood ratios (PLR) and negative likelihood ratios (NLR) for each method of diagnosing AF. Unlike sensitivity and specificity, likelihood ratios make explicit the impact of the test result on the probability of the disease. To minimise heterogeneity we analysed the results a priori grouped according to method of diagnosing AF, and diagnostic accuracy was assessed by comparison of sensitivities and specificities with respective 95% confidence intervals. Sub-group analyses were planned according to study quality and groups of healthcare professionals in primary care. We used univariate random effects meta-analysis to derive pooled estimates for sensitivity and specificity when there were less than four studies within sub-groups as the bivariate model is unreliable in this context. Heterogeneity is presumed in meta-analyses of diagnostic test studies and the $I^2$ statistic cannot be reliably used for its assessment. We therefore described heterogeneity by variation in the outcomes from included studies and our pooled estimates by visual inspection of the SROC plots and how close individual studies lie to the predicted ROC curve. An assessment of publication bias was made according to categories of method for detecting AF using Deek’s Funnel plot asymmetry test; a P-value<0.10 was used to signify the presence of publication
bias. Analyses were conducted using Stata Version 11.0 and Review Manager 5.2 for quality assessments.
RESULTS

We identified, after the removal of duplicate records, 4,426 potential citations, of which 62 were identified as relevant for detailed evaluation (figure 1). After full-text review, 10 studies were included in the final analyses (table 1). There was one study that met selection criteria for which there were insufficient data for reported outcomes (table 2).

Study characteristics

Of the 10 studies included in our review (table 1), there was one randomised trial, two case-control, and seven cross-sectional studies. The 10 studies investigated a total of 16 methods of diagnosing AF (a total of 55,376 participant ECGs), which were categorised into two intervention groups: 1) automated software (eight studies; nine diagnostic methods) and, 2) any healthcare professional (five studies; seven diagnostic methods). Sub-groups of healthcare professional were defined as: secondary care physicians (two studies; two diagnostic methods) and primary care professionals (three studies; five diagnostic methods), the latter comprising GPs (three studies) and practice nurses (two studies). Across studies without a case control design, the prevalence of AF ranged from 6.7% to 18.6%. (See Table 1). The three studies conducted in a primary care setting included participants over 65 years of age and recruited patients eligible for AF screening. However, the remainder of studies were conducted using patients with existing cardiac pathologies in a secondary care setting.

For five studies, the reference standard was 12-lead ECG interpreted by at least two cardiologists. Of the remaining studies, four used ECG interpretation by a single cardiologist as the reference standard and one study used two trained secondary care clinicians.

Quality assessment

Figure 2 shows the methodological quality of included studies according to QUADAS-2 criteria was generally low. Five studies with the lowest methodological quality (D-grade) were due to the methodological interpretation of the reference standard being unclear or at high risk of bias. One
The study was graded as category C because it was unclear whether the reference standard was interpreted without knowledge of the index test.

**Data synthesis**

Automated software was found to have a pooled sensitivity of 0.89 (95% CI 0.82-0.93) and specificity of 0.99 (95% CI 0.99-0.99) for diagnosing AF using 12-lead ECG. (Figure 3) This corresponded with a PLR of 96.6 (95% CI 64.2-145.6) and NLR of 0.11 (95% CI 0.07-0.18). In contrast, the pooled specificity for the accuracy of any healthcare professional diagnosing AF (Figure 4) was lower than automated software although there was a similar sensitivity of this method for interpreting ECGs; sensitivity 0.92 (95% CI 0.81-0.97), specificity 0.93 (95% CI 0.76-0.98), PLR 13.9 (95% CI 3.5-55.3), NLR 0.09 (95% CI 0.03-0.22).

Figure 5 shows the sensitivity and specificity for diagnosing AF by primary care professionals was relatively high [sensitivity 0.96 (95% CI 0.66-1.00), specificity 0.94 (95% CI 0.85-0.98), PLR 15.4 (95% CI 5.9-40.3), NLR 0.05 (95% CI 0.00 to 0.49)]. The sub-group analyses for categories of GPs and nurses (figure 6) suggest this may be driven by a greater specificity of GPs’ AF diagnoses [GPs: sensitivity 0.91 (95% CI 0.68-1.00); specificity 0.96 (95% CI 0.89-1.00) when compared to nurses: sensitivity 0.88 (95% CI 0.63-1.00); specificity 0.85 (95% CI 0.83-0.87)].

Visual inspection of the SROC plots (figure 7) confirms there was substantial variation in the outcomes from studies investigating the accuracy of clinicians’ 12-lead ECG diagnosis and suggests heterogeneity amongst these studies was greater than the automated software studies.

There was no evidence of publication bias for studies of any clinician (p=0.29) or any primary care clinician diagnosis (p=0.19). However, studies of software ECG interpretation suggested the presence of publication bias, (p=0.02), with the possible underrepresentation of smaller studies with a lower accuracy of diagnosing AF. Bivariate sub-group analyses were similar after exclusion of studies with the lowest quality [Software: sensitivity 0.82 (95% CI 0.73-0.88), specificity 0.99 (95% CI 0.98-0.99); any healthcare professionals: sensitivity 0.92 (95% CI 0.81-0.97), specificity 0.91 (95% CI 0.70-0.98)].
any primary care professionals: sensitivity 0.93 (95% C.I 0.67-0.99), specificity 0.92 (95% C.I 0.85-0.96)].
DISCUSSION

This review of 10 studies found automated software analysis had a borderline greater specificity for AF diagnosis than healthcare professional interpretation of 12-lead ECGs. The sensitivities of automated software, any healthcare professionals and primary care professionals for interpreting 12-lead ECGs to diagnose AF were similar.

Strengths and limitations

To our knowledge, this study is the first systematic review and meta-analysis of different methods for interpreting 12-lead ECGs to diagnose AF. A strength of our study was the use of a standardised protocol that is consistent with published guidelines for systematic reviews of diagnostic test studies. Moreover, we used a comprehensive search strategy that included contacting authors of potentially relevant studies. Our findings indicated a probable lack of publication bias for studies of clinicians’ 12-lead ECG diagnoses of AF. However, there was the possibility of publication bias for studies investigating automated software and this may limit the validity of the findings for this diagnostic modality. One study was excluded due to the insufficient reporting of outcome data to enable meta-analysis and this could have influenced our findings. However, the number of misdiagnoses of AF was similar to that of other studies investigating the accuracy of automated software for making ECG diagnoses of AF and the impact of excluding this study is likely to be minimal. Only one of the nine included studies adopted a prospective design and there were a number of inherent methodological weaknesses in other studies as reflected by our appraisal of study quality. No studies were judged to have met all QUADAS-2 criteria. However, our bivariate sub-group analyses that excluded studies judged to have the lowest (grade D) methodological quality found similar outcomes to our primary analyses, and strengthens the validity of our findings. Most studies were conducted in a secondary care setting and there was substantial variation in the proportion of patients with AF. This limits the generalisability of our findings to unselected primary care populations that AF screening is intended for. However, the measures used to determine diagnostic accuracy in our study are prevalence independent and, consequently, our findings could be translatable to populations in different
healthcare settings. There was heterogeneity amongst the studies within all categories of methods for diagnosing AF and this is likely to be attributable to differences in study population and design. This variation was least for studies of automated software and strengthens the validity of findings for this approach to AF diagnosis. Heterogeneity was greatest for the category of any healthcare professionals’ interpretation of 12-lead ECGs and is likely to arise from differences in professional groups and clinical expertise.

Findings in context of previous research

Our review identified automated software and healthcare professional interpretation of 12-lead ECGs as methods for diagnosing AF. Furthermore, we also analysed the interpretation of ECGs in a restricted group of primary care professionals. Our findings for automated software, using sensitivity and specificity as measures of diagnostic accuracy, are consistent with those from the SAFE study. Due to the significantly higher specificity of this diagnostic modality, our findings suggest software is the best method for correctly identifying patients with normal 12-lead ECGs and minimising the risk of false positive diagnoses of AF.

The sensitivities of all methods for diagnosing AF were similar, although these were sufficiently low to give rise to false negative AF diagnoses. As compared to any healthcare professionals’ ECG interpretation, the point estimates for sensitivity and specificity were reassuringly high for AF diagnoses made by primary care professionals. However, our sub-group analyses suggest this may be attributable to better 12-lead ECG interpretation by GPs; in comparison to GPs, nurses were found to have a significantly lower specificity for diagnosing AF. The point estimates for the accuracy of primary care clinicians’ AF diagnoses contrast with outcomes from the SAFE study. The SAFE trial was the largest, pragmatic study of AF screening in primary care and secondary analyses of the trial data suggest the accuracy of GP and nurse AF diagnoses was substantially lower than our findings. Although data from SAFE were included in our meta-analyses, the difference in outcomes between
this study and our pooled results may arise from the statistical approach used for data synthesis; random effects meta-analysis would have provided greater weighting to the smaller, low quality studies that reported a higher diagnostic accuracy arising from primary care professionals, thus inflating the pooled estimates in our review. Moreover, the primary care studies investigating the accuracy of diagnosing AF are likely to have included self-selecting practices with an interest in AF and it is possible that the accuracy of diagnosing arrhythmia by primary care professionals in routine practice could be even lower than that found in the SAFE trial and our review.

The current gold-standard test for diagnosing AF is 12-lead ECG\textsuperscript{6,25} and consensus recommends this should be interpreted by competent healthcare professionals as part of AF screening.\textsuperscript{26} Both systematic and opportunistic screening for AF using 12-lead ECG in patients over 65 years was found to be an effective approach for improving the detection of this arrhythmia.\textsuperscript{4}

Screening for AF is likely to be implemented in primary care and, consequently, healthcare professionals in this setting would be expected to undertake screening activities. It is unclear what the optimal service configurations are for delivering such a screening programme and both GPs and nurses may be expected to undertake the role of diagnosing AF. Our findings suggest there is potential for combining mixed modalities of ECG interpretation for the diagnosis of AF. Certainly, automated software has potential utility for the triage of ECGs and exclusion of patients with normal ECG findings. However, correctly diagnosing AF using software interpretation alone has a limited sensitivity resulting in the potential for incorrect exclusion of AF, and interpreting ECGs to verify the presence of AF in this circumstance is likely to require additional interpretation from a competent healthcare professional.\textsuperscript{26} The findings of this review suggest the accuracy of correctly making AF diagnoses by primary care professionals, in particular nurses, could be improved. It is therefore conceivable that the skills of healthcare professionals in this setting would need improving to ensure the effectiveness of screening is not undermined.
The opinions of primary care professionals about AF screening are not known and a greater understanding of the barriers, facilitators and learning needs of these important stakeholder groups, in particular nurses, is required before screening can be implemented optimally. Furthermore, studies that investigate optimal service configurations for diagnosing AF in primary care and how these translates into better detection rates would help plan the delivery of an effective national AF screening programme.

Conclusions

Automated software had the greatest specificity for AF diagnosis using 12-lead ECG than healthcare professional diagnosis of this arrhythmia. Although the accuracy of diagnosing AF in primary care is reassuring, this is driven by GP’s diagnosis of AF. If a national AF screening programme is introduced into primary care it is possible that the skills of GPs and nurses for making 12-lead ECG diagnoses of AF could be improved to ensure the effectiveness of screening is not undermined.
REFERENCES


# TABLES

## Table 1: Characteristics of included studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting, population &amp; sample size</th>
<th>AF Prevalence/ proportion (%)</th>
<th>Study Design†</th>
<th>Index test(s)</th>
<th>Reference test</th>
<th>Outcomes</th>
<th>Quality grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bourdillon 1978¹⁴</td>
<td>UK; secondary care; 221 ECGs of adult subjects</td>
<td>18.6</td>
<td>CS</td>
<td>Software interpretation (Mount Sinai)</td>
<td>2 clinicians, independent interpretation</td>
<td>Test 1: Sensitivity 0.85; specificity 0.98</td>
<td>C</td>
</tr>
<tr>
<td>Davidenko 2007¹⁵</td>
<td>USA; secondary care; 35,508 consecutive ECGs were reviewed</td>
<td>7.9</td>
<td>CS</td>
<td>Software diagnosis (Marquettes)</td>
<td>Interpretation by several cardiologists with a group consensus</td>
<td>Sensitivity 0.97; specificity 1.00</td>
<td>D</td>
</tr>
<tr>
<td>Gregg 2008¹⁶</td>
<td>UK; secondary care; database of 50,000 hospital ECGs; 1,785 randomly selected</td>
<td>6.1</td>
<td>CS</td>
<td>Software interpretation</td>
<td>Interpreted by 2 cardiologists</td>
<td>Sensitivity 0.89; specificity 0.99</td>
<td>D</td>
</tr>
<tr>
<td>Hakacova 2012¹⁷</td>
<td>Sweden; secondary care; total of 576 ECGs from 503 participants with a mean age of 64 years</td>
<td>10.4</td>
<td>CS</td>
<td>Test 1: Non expert secondary care clinian</td>
<td>Interpreted by 2 expert cardiologists</td>
<td>Test 1: Sensitivity 0.86; specificity 0.99</td>
<td>B</td>
</tr>
<tr>
<td>Hobbs 2005¹⁸</td>
<td>UK; primary care; 9,866 patients aged 65 years, 2595 ECGs were reviewed</td>
<td>6.8</td>
<td>RCT</td>
<td>Test 1: General practitioner interpretation</td>
<td>Test 1: Sensitivity 0.80; specificity 0.92</td>
<td>Test 2: Sensitivity 0.77; specificity 0.85</td>
<td>B</td>
</tr>
<tr>
<td>Author</td>
<td>Country/Setting</td>
<td>N</td>
<td>Study Type</td>
<td>Examination</td>
<td>Interpretation</td>
<td>Diagnosis Method</td>
<td>Sensitivity/Specificity</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------</td>
<td>----------------------------------------</td>
<td>-------------------------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Poon 2005</td>
<td>USA; secondary care; 4,297 consecutive ECGs were reviewed</td>
<td>6.3</td>
<td>CS</td>
<td>Software</td>
<td>Cardiologist interpretation</td>
<td>Sensitivity 0.91; specificity 0.99</td>
<td>D</td>
</tr>
<tr>
<td>Reddy 1998</td>
<td>USA; secondary care; 10,352 ECGs were reviewed</td>
<td>8</td>
<td>CS</td>
<td>Mac-rhythm</td>
<td>Cardiologist interpretation</td>
<td>Sensitivity 0.88; specificity 0.99</td>
<td>D</td>
</tr>
<tr>
<td>Rhys 2013</td>
<td>UK; primary care; patients ≥65 years recruited from flu clinics; 32 ECGs reviewed</td>
<td>6.3</td>
<td>CS</td>
<td>Test 1: Software interpretation; Test 2: General practitioner interpretation</td>
<td>ECG interpreted by cardiologist</td>
<td>Test 1: Sensitivity 1; specificity 1; Test 2: Sensitivity 1; specificity 1</td>
<td>D</td>
</tr>
<tr>
<td>Shiyovich 2010</td>
<td>Israel; secondary care; 268 patient’s ECGs</td>
<td>81.7</td>
<td>CC</td>
<td>Secondary care clinician interpretation</td>
<td>Interpretation by 2 senior cardiologists</td>
<td>Sensitivity 0.97; specificity 0.31</td>
<td>B</td>
</tr>
<tr>
<td>Somerville 2000</td>
<td>UK; Primary care; 86 patients recruited from one general practice, 86 ECGs reviewed</td>
<td>31.5</td>
<td>CC</td>
<td>Test 1: Practice nurse interpretation; Test 2: General practitioner interpretation</td>
<td>Interpreted by consultant cardiologist</td>
<td>Test 1: Sensitivity 0.97; specificity 0.88; Test 2: Sensitivity 1; specificity 0.98</td>
<td>B</td>
</tr>
</tbody>
</table>

†CC = Case-control study; CS = cross-sectional study; RCT = Randomised controlled Trial; C = Cohort study
Table 2: Characteristics includable studies with insufficient data.

<table>
<thead>
<tr>
<th>Author/Year</th>
<th>Setting, population &amp; sample size</th>
<th>Study design</th>
<th>Intervention</th>
<th>Comparator</th>
<th>Reported outcomes</th>
<th>Reason for exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bogun 2004\textsuperscript{24}</td>
<td>USA; secondary care; database of 2298 ECGs from 1085 patients</td>
<td>Cross sectional</td>
<td>Software interpretation using GE Marquette 12 SE or MACR programs, overread by cardiologists</td>
<td>Interpretation by 2 electrophysiologists</td>
<td>442 (19%) of the 2298 ECGs had an incorrect computer interpretation of AF in 382 (35%) of patients</td>
<td>Number of true AF, false AF, missed AF, and non AF were not reported</td>
</tr>
</tbody>
</table>
FIGURES

Figure 1: Study selection and stratification

CINAHL 347 citations
EMBASE 2700 citations
LILACS 26 citations
MEDLINE 2982 citations
Reference List 4 citations

6059 titles or abstracts identified and screened for retrieval

5998 excluded:
- 1633 duplicate records
- 4365 not relevant

62 full-text articles assessed

52 excluded:
- 40 not diagnosis studies
- 3 editorials or reviews
- 8 not relevant to study design
- 1 insufficient data

10 studies included in final review
Figure 2: Study quality according to QUADAS-2 criteria

<table>
<thead>
<tr>
<th></th>
<th>Risk of Bias</th>
<th>Applicability Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient Selection</td>
<td>Index Test Software diagnosis</td>
</tr>
<tr>
<td></td>
<td>Flow and Timing</td>
<td>Reference Standard</td>
</tr>
<tr>
<td>Bourdillon 1973</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Davidenko 2007</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>Oregg 2003</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Hakacova 2012</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>Hobbs 2005</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>Poon 2005</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Reddy 1998</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Rhy 2013</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>Shyovitch 2010</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>Somerville 2003</td>
<td>+</td>
<td>?</td>
</tr>
</tbody>
</table>

- **High**
- **Unclear**
- **Low**
Figure 3: Sensitivity and specificity of 12-lead ECG interpretation using automated software
Figure 4: Sensitivity and specificity of 12-lead ECG interpretation by any healthcare professional
Figure 5: Sensitivity and specificity of 12-lead ECG interpretation by primary care professionals
Figure 6: Sub-group analyses of the sensitivity and specificity of 12-lead ECG interpretation by GPs and practice nurses
Figure 7: Summary Receiver Operating Characteristic (SROC) plots for the accuracy of 12-lead ECG interpretation by software, any clinician, and primary care clinician.