ORIGINAL RESEARCH ARTICLE

Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms

Ming-Zher Poh,[1] Yukkee Cheung Poh,[1] Pak-Hei Chan,[2] Chun-Ka Wong,[2] Louise Pun,[3] Wangie Wan-Chiu Leung,[3] Yu-Fai Wong,[3] Michelle Man-Ying Wong,[3] Daniel Wai-Sing Chu,[3] Chung-Wah Siu[2]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/heartjnl-2018-313147).

[1]Cardiio, Cambridge, Massachusetts, USA
[2]Division of Cardiology, Department of Medicine, University of Hong Kong, Hong Kong
[3]Department of Family Medicine and Primary Healthcare, Hong Kong East Cluster, Hospital Authority, Hong Kong

Correspondence to
Dr Ming-Zher Poh, Cardiio, Inc., Cambridge, MA 02139, USA; mingzher@cardiio.com

M-ZP and YCP contributed equally.

Received 9 February 2018
Revised 20 March 2018
Accepted 9 April 2018

ABSTRACT

Objective To evaluate the diagnostic performance of a deep learning system for automated detection of atrial fibrillation (AF) in photoplethysmographic (PPG) pulse waveforms.

Methods We trained a deep convolutional neural network (DCNN) to detect AF in 17 s PPG waveforms using a training data set of 149 048 PPG waveforms constructed from several publicly available PPG databases. The DCNN was validated using an independent test data set of 3039 smartphone-acquired PPG waveforms from adults at high risk of AF at a general outpatient clinic against ECG tracings reviewed by two cardiologists. Six established AF detectors based on handcrafted features were evaluated on the same test data set for performance comparison.

Results In the validation data set (3039 PPG waveforms) consisting of three sequential PPG waveforms from 1013 participants (mean (SD) age, 68.4 (12.2) years; 46.8% men), the prevalence of AF was 2.8%. The area under the receiver operating characteristic curve (AUC) of the DCNN for AF detection was 0.997 (95% CI 0.996 to 0.999) and was significantly higher than all the other AF detectors (AUC range: 0.924–0.985). The sensitivity of the DCNN was 95.2% (95% CI 88.3% to 98.7%), specificity was 99.0% (95% CI 98.6% to 99.3%), positive predictive value (PPV) was 72.7% (95% CI 65.1% to 79.3%) and negative predictive value (NPV) was 99.9% (95% CI 99.7% to 100%) using a single 17 s PPG waveform. Using the three sequential PPG waveforms in combination (<1 min in total), the sensitivity was 100.0% (95% CI 87.7% to 100%), specificity was 99.6% (95% CI 99.0% to 99.9%), PPV was 87.5% (95% CI 72.5% to 94.9%) and NPV was 100% (95% CI 99.4% to 100%).

Conclusions In this evaluation of PPG waveforms from adults screened for AF in a real-world primary care setting, the DCNN had high sensitivity, specificity, PPV and NPV for detecting AF, outperforming other state-of-the-art methods based on handcrafted features.

INTRODUCTION

Atrial fibrillation (AF) is associated with a third of all strokes,[1] but is asymptomatic in over one-third of patients[2] and often goes undiagnosed. Although treatment of patients with AF with oral anticoagulants is effective in reducing stroke risk by 60%–70%,[3] nearly 25% of patients with stroke only discover the presence of AF after the potentially preventable stroke event.[4] As the use of smartphone apps, wearable fitness trackers and smartwatches capable of acquiring pulse waveforms via photoplethysmography (PPG) becomes increasingly common, these tools may present a new avenue for early detection of undiagnosed AF and timely anticoagulant treatment to prevent stroke.

Prior work on PPG-based AF detection algorithms relied predominantly on explicit rules and handcrafted features derived from a sequence of interbeat intervals of the PPG waveform aimed at capturing pulse irregularity, the hallmark of AF. Published methods include coefficient of variation (CoV),[5] coefficient of sample entropy (CoSEn),[6] normalised root mean square of successive differences (nRMSSD) + Shannon entropy (ShE),[7] nRMSSD + Poincaré plot geometry (SD1/SD2),[8] Poincaré plot patterns[9] and autocorrelation analysis using a support vector machine (SVM).[10] Thus far, achieving both very high sensitivity and specificity remains challenging because of other arrhythmias such as ectopic beats and the presence of motion or noise artefacts in the PPG signal that can mimic AF.

In this work, we trained a deep convolutional neural network (DCNN) to distinguish between noise, sinus rhythm, ectopic rhythms and AF using a large set of PPG signals. In contrast to handcrafted features, the DCNN automatically learns the most predictive features directly from the raw PPG waveform based on the training examples.

METHODS

Data sets and reference standards

To develop the DCNN, we constructed a data set (PPG-RHYTHM) from several publicly accessible PPG repositories, including the MIMIC-III critical care database,[11] the Vortal data set from healthy volunteers[12] and the IEEE-TBME PPG Respiratory Rate Benchmark data set.[13] All PPG recordings were resampled to 30 Hz and divided into segments of 512 samples (approximately 17 s long). A total of 186 317 PPG segments with concurrent ECG from 3373 unique persons were analysed and assigned to one of four rhythm classes: sinus rhythm (n=81 437 waveforms), noise (n=6561), ectopic rhythm (n=17 257) and AF (n=81 062). The signal quality index (SQI) of each PPG segment was assessed by forming a template beat and quantifying the degree of similarity between a given beat and the running template.[14] PPG segments with an

To cite: Poh M-Z, Poh YC, Chan P-H, *et al*. *Heart* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/heartjnl-2018-313147

Table 1 Summary of the MOBILE-SCREEN-AF (clinical validation) data set

Characteristics	n
Number of PPG waveforms	3039
Patient demographics	
Number of unique individuals	1013
Age, mean (SD), year	68.4 (12.2)
Male, n/N (%)	474/1013 (46.8)
Hypertension, n/N (%)	916/1013 (90.4)
Diabetes mellitus	371/1013 (36.6)
Coronary artery disease	164/1013 (16.2)
Heart failure	45/1013 (4.4)
Previous myocardial infarction	33/1013 (3.3)
Previous stroke	106/1013 (10.5)
CHA_2DS_2-VASc score, mean (SD)	3.0 (1.5)
ECG rhythm diagnosis, n/N (%)	
Sinus rhythm	920/1013 (90.8)
Atrial fibrillation	28/1013 (2.8)
Atrial flutter	1/1013 (0.1)
Premature atrial contractions	28/1013 (2.8)
Premature ventricular contractions	28/1013 (2.8)
Sinus arrhythmia	8/1013 (0.8)

PPG, photoplethysmography. n, number of subsamples. N, total sample. CHA_2DS_2-VASc, congestive heart failure: +1, hypertension: +1, age ≥ 75 yrs: +2, diabetes mellitus: +1, stroke or transient ischemic attack: +2, vascular disease: +1, age 65-74: +1, female gender: +2.

average SQI below 0.4 were assigned to the noise class. From the remaining clean PPG segments, those from the MIMIC-III database were labelled based on charted observations entered by care providers and additional review by an experienced researcher; segments from the Vortal (healthy adults) and IEEE-TBME PPG Respiratory Rate Benchmark data set (predominantly children) were labelled as sinus rhythm. The ectopic rhythm class included premature atrial contractions, premature ventricular contractions, and bigeminy, trigeminy and quadrigeminy rhythms. We divided the PPG-RHYTHM data set into training, tuning and test subsets using an 80:10:10 ratio; the distribution of rhythm classes was kept the same.

For clinical validation of the DCNN, we used an independent data set (MOBILE-SCREEN-AF) described in detail by Chan *et al*.[10] Briefly, 3039 PPG waveforms were acquired from 1013 participants (three consecutive PPG waveforms per participant) at high risk of AF (table 1) using a smartphone (iPhone 4S; Apple) at a general outpatient clinic. The PPG waveforms were sampled at 30 Hz, and each measurement lasted 17 s (512 samples). A single-lead I ECG tracing was also recorded using a handheld device with stainless steel electrodes (first-generation AliveCor heart monitor; AliveCor). All ECG tracings were of sufficient signal quality and reviewed by two independent cardiologists blinded to the PPG waveforms and to each other's diagnosis to provide the reference diagnosis using standard criteria.[15] There were no discrepancies in the ECG interpretations. AF was diagnosed in 28 (2.8%) participants and confirmed with a standard 12-lead ECG; 5 of the 28 (17.9%) patients had newly diagnosed AF detected with the screening test.

DCNN architecture and training

Our deep learning system takes as input a PPG waveform of approximately 17 s long (sampled at 30 Hz) and outputs a label prediction of one of the four rhythm classes, along with a probability distribution over the four classes. All PPG waveforms were detrended and filtered by using a bandpass filter (0.48–12 Hz) to remove baseline wander and high frequency. We use a densely connected DCNN architecture[16] with six dense blocks (a total of 201 layers) and a growth rate of 6 (figure 1A). This architecture was selected because it encourages feature reuse and significantly reduces the number of parameters to be learnt. To improve computational efficiency and model compactness, we use bottleneck and compression layers. The DCNN model consists of a total of 445 856 trainable parameters and only requires 3.6 MB of storage space. The workflow for developing and validating the DCNN is shown in figure 1B. We trained our model from scratch on the PPG-RHYTHM training subset (149 048 waveforms) adopting the weight initialisation of He *et al*[17] and using stochastic gradient descent with a Nesterov momentum[18] of 0.9 for a total of 300 epochs. We used a cyclical learning rate schedule[19] and reduced the learning rates by a factor of 10 at 50% and 75% of the total number of training epochs. The best model based on performance on the PPG-RHYTHM tuning subset (18 631 waveforms) was saved and used for subsequent testing. The PPG-RHYTHM test subset (18 638 waveforms) was used to characterise the accuracy of the DCNN for multiclass rhythm classification, and to visualise the last hidden layer representations in the DCNN using t-SNE (t-distributed stochastic neighbour embedding).[20]

Statistical analysis and performance comparison

The accuracy of the DCNN for detecting AF in a primary care setting was evaluated using the MOBILE-SCREEN-AF data set for binary classification. DCNN predictions of noise, sinus rhythm or ectopy were considered as a non-AF label. For comparison, we also evaluated the performance of six state-of-the-art AF detection algorithms (CoV,[5] CoSEn,[6] nRMSSD + ShE,[7] nRMSSD + SD1/SD2,[8] Poincaré plot[9] and SVM[10]) on the same data set. A brief description of the algorithms is available in online supplementary eAppendix. In addition, we constructed an ensemble learner by combining these six AF detectors using a majority voting scheme. With the exception of the SVM, each of the comparison models was retrained on the same PPG-RHYTHM training subset as the DCNN to determine the optimal thresholds for AF detection (performance of the comparison models on the PPG-RHYTHM test subset is shown in online supplementary eTable 1). Each detector's output was compared with the reference diagnosis in the MOBILE-SCREEN-AF data set for each of the three consecutive PPG waveforms individually (single measurement) and combined (triplicate measurements). Combined readings were considered AF if at least two of the three individual PPG waveforms were classified by the detector as AF.

Receiver operating characteristic (ROC) curves were generated by varying the operating threshold for each AF detector. The accuracy of all the AF detectors was compared using sensitivity, specificity, positive predictive value (PPV), negative predictive values (NPV) and area under the ROC curve (AUC), using cardiologists' annotations of corresponding ECGs as the ground truth. The 95% CIs for the AUCs were computed and compared using the DeLong test.[21] The 95% CIs for the sensitivity and specificity were calculated to be 'exact' Clopper-Pearson intervals[22]; CIs for the PPV and NPV were computed as the standard logit CIs given by Mercaldo *et al*.[23] All statistical tests used in this study were two-sided and a p value less than 0.05 was considered significant.

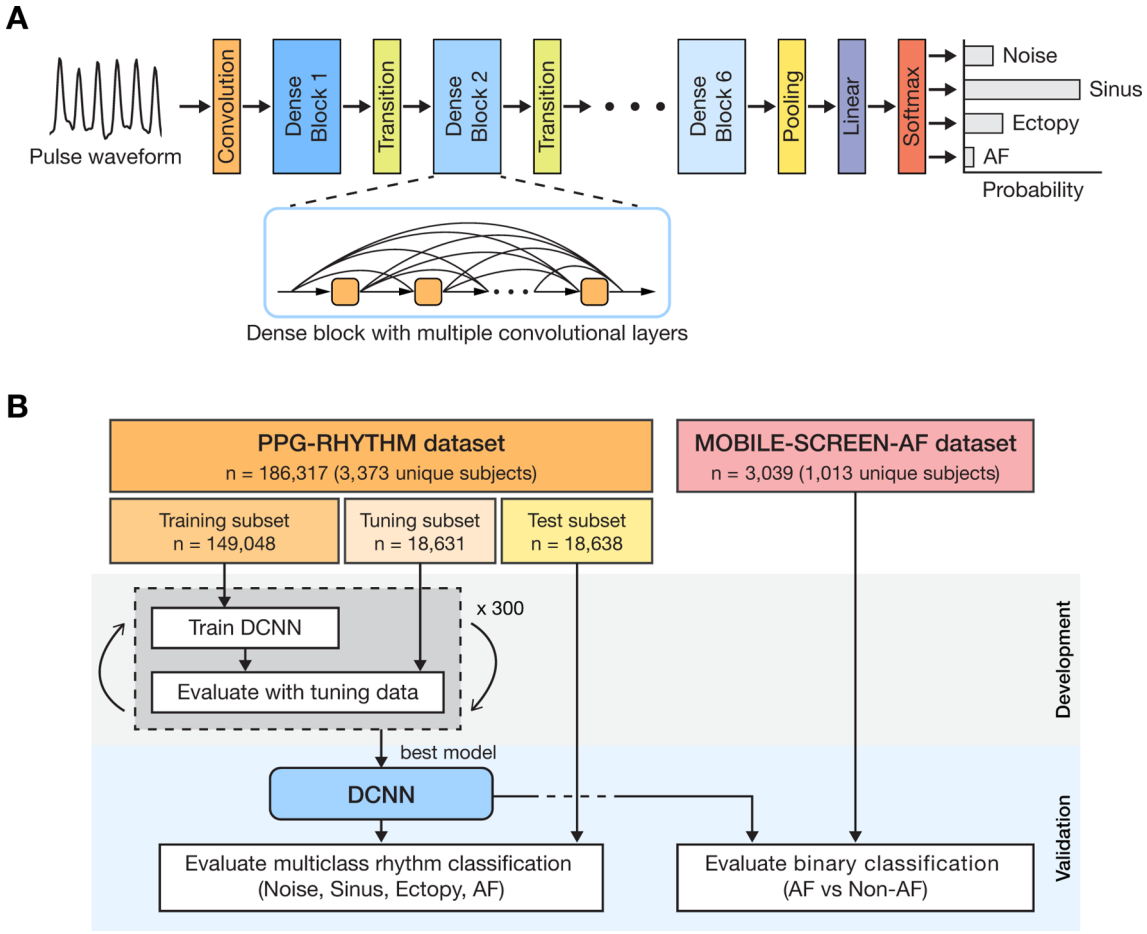

Figure 1 Architecture, development and validation of the DCNN. (A) The DCNN has six dense blocks, each of which consists of multiple densely connected convolutional layers. (B) Workflow diagram showing the data sets used to develop and validate the DCNN. AF, atrial fibrillation; DCNN, deep convolutional neural network; PPG, photoplethysmography.

RESULTS

Multiclass rhythm classification

The DCNN model learnt to distinguish between four rhythm classes of noise, sinus rhythm, ectopy and AF in PPG waveforms with an overall accuracy of 96.1% (95% CI 95.8% to 96.3%). The sensitivity for noise, sinus rhythm, ectopy and AF was 97.0% (95% CI 95.4% to 98.2%), 99.1% (95% CI 98.8% to 99.3%), 72.2% (95% CI 69.9% to 74.4%) and 97.6% (95% CI 97.2% to 97.9%), respectively; the corresponding specificity was 100% (95% CI 99.9% to 100%), 98.5% (95% CI 98.2% to 98.7%), 98.8% (95% CI 98.6% to 99.0%) and 96.5% (95% CI 96.1% to 96.8%), respectively (table 2 and online supplementary eFigure 1).

AF detection from a single measurement

The performance of the AF detectors for classifying individual PPG waveforms is presented in table 3. The DCNN achieved a high sensitivity of 95.2% (95% CI 88.3% to 98.7%), and the highest specificity, PPV and NPV of 99.0% (95% CI 98.6% to 99.3%), 72.7% (95% CI 65.1% to 79.3%) and 99.9% (95% CI 99.7% to 100%) among all the AF detectors. The AUC for the DCNN was 0.997 (95% CI 0.996 to 0.999), significantly higher than all other detectors (AUC range: 0.924–0.985, $p<0.001$) (figure 2A,B). Using an ensemble classifier to combine the six other conventional AF detectors improved performance over all of its individual members except the SVM, but did not reach a performance comparable with the DCNN. The effect of input segment length on performance of the DCNN was evaluated by testing segment lengths of 2–17 s. The AUC and specificity of

Table 2 Performance of the DCNN for multiclass rhythm classification on the PPG-RHYTHM test set

	Value, % (95% CI)			
Class	Sensitivity	Specificity	PPV	NPV
Noise	97.0 (95.4 to 98.2)	100 (99.9 to 100)	98.9 (97.8 to 99.5)	99.9 (99.8 to 99.9)
Sinus rhythm	99.1 (98.8 to 99.3)	98.5 (98.2 to 98.7)	98.1 (97.8 to 98.4)	99.2 (99.1 to 99.4)
Ectopy	72.2 (69.9 to 74.4)	98.8 (98.6 to 99.0)	85.0 (83.1 to 86.7)	97.4 (97.2 to 97.6)
Atrial fibrillation	97.6 (97.2 to 97.9)	96.5 (96.1 to 96.8)	95.5 (95.1 to 95.9)	98.1 (97.8 to 98.4)

Overall accuracy: 96.1% (95% CI 95.8% to 96.3%).
DCNN, deep convolutional neural network; NPV, negative predictive value; PPG, photoplethysmography; PPV, positive predictive value.

Table 3 DCNN performance for detection of AF versus several state-of-the-art AF detectors on the MOBILE-SCREEN-AF (clinical validation) data set

	Single measurement, % (95% CI)				Triplicate measurements, % (95% CI)			
Method	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
CoSEn[6]	88.1 (79.2 to 94.1)	81.9 (80.5 to 83.3)	12.2 (11.0 to 13.4)	99.6 (99.3 to 99.8)	**100.0** (87.7 to 100)	84.7 (82.3 to 86.9)	15.6 (13.8 to 17.7)	**100.0** (99.4 to 100)
Poincaré plot[9]	92.9 (85.1 to 97.3)	82.8 (81.4 to 84.2)	13.3 (12.2 to 14.5)	99.8 (99.5 to 99.9)	**100.0** (87.7 to 100)	85.8 (83.5 to 87.9)	16.7 (14.6 to 18.9)	**100.0** (99.4 to 100)
nRMSSD + ShEn[7]	84.5 (75.0 to 91.5)	88.4 (87.2 to 89.5)	17.1 (15.3 to 19.1)	99.5 (99.2 to 99.7)	92.9 (76.5 to 99.1)	91.0 (89.0 to 92.7)	22.6 (18.9 to 26.8)	99.8 (99.2 to 99.9)
nRMSSD + SD1/SD2[8]	90.5 (82.1 to 95.8)	88.9 (87.7 to 90.0)	18.8 (17.0 to 20.8)	99.7 (99.4 to 99.8)	96.4 (81.7 to 99.9)	92.2 (90.3 to 93.8)	26.0 (21.9 to 30.5)	99.9 (99.3 to 100)
CoV[5]	**96.4** (89.9 to 99.3)	88.5 (87.3 to 89.6)	19.2 (17.6 to 21.0)	**99.9** (99.7 to 100)	**100.0** (87.7 to 100)	91.1 (89.1 to 92.8)	24.1 (20.7 to 28.0)	**100.0** (99.4 to 100)
Ensemble	90.5 (82.1 to 95.8)	90.3 (89.2 to 91.3)	20.9 (18.9 to 23.2)	99.7 (99.4 to 99.9)	**100.0** (87.7 to 100)	92.5 (90.7 to 94.1)	27.5 (23.3 to 32.0)	**100.0** (99.4 to 100)
SVM[10]	90.5 (82.1 to 95.8)	96.1 (95.3 to 96.8)	39.6 (35.1 to 44.2)	99.7 (99.5 to 99.9)	92.9 (76.5 to 99.1)	97.7 (96.5 to 98.5)	53.1 (42.7 to 63.2)	99.8 (99.2 to 100)
DCNN	95.2 (88.3 to 98.7)	**99.0** (98.6 to 99.3)	**72.7** (65.1 to 79.3)	**99.9** (99.7 to 100)	**100.0** (87.7 to 100)	**99.6** (99.0 to 99.9)	**87.5** (72.5 to 94.9)	**100.0** (99.4 to 100)

AF, atrial fibrillation; CoSEn, coefficient of sample entropy; CoV, coefficient of variation; DCNN, deep convolutional neural network; NPV, negative predictive value; nRMSSD, normalised root mean square of successive differences; PPV, positive predictive value; SD1/SD2, Poincaré plot geometry; ShEn, Shannon entropy; SVM, support vector machine. The highest score among all detectors is indicated in bold.

the DCNN decreased slightly as the input segment length was reduced from 17 s to 5 s (figure 2C); the corresponding sensitivity of the DCNN decreased steadily. The performance of the DCNN deteriorated rapidly for segment lengths shorter than 5 s.

The contingency table for each AF detector is shown in figure 3. Examples of the pulse waveforms correctly and incorrectly classified by the DCNN are shown in figure 4. Among the four false negatives produced by the DCNN, one was classified as an ectopic rhythm and the other three as noisy rhythms (figure 4C). In all four cases, the second highest class probability produced by the DCNN corresponded to AF.

AF detection from triplicate measurements

When all three individual PPG recordings for each patient were combined, the performance of all AF detectors improved (table 3). The DCNN outperformed all other methods across all metrics, achieving a sensitivity of 100% (95% CI 87.7% to 100%), specificity of 99.6% (95% CI 99.0% to 99.9%), PPV of 87.5% (95% CI 72.5% to 94.9%) and NPV of 100% (95% CI 99.4% to 100%). Only four mistakes (false positives) were made by the DCNN; three were deemed to be in sinus rhythm and one had premature atrial contractions based on the corresponding ECG tracings.

Visualising the DCNN

To gain insight on what the DCNN learnt, we visualised the first-layer weights that represent the learnt convolutional filters (figure 5A). The learnt filters appear to be suitable for detecting features such as peaks, troughs, and upward and downward slopes. Indeed, visualisation of the first layer activation maps (figure 5B) revealed strong activations at positions coinciding with the peaks, troughs, and upward and downward slopes of an input pulse waveform, providing further confirmation. The internal features automatically learnt by the DCNN are visualised using t-SNE in figure 5C. Each point represents a pulse waveform projected from the output of the DCNN's last hidden layer into two dimensions. Points belonging to the same rhythm class clustered together. AF clustered opposite to sinus rhythm while ectopic rhythms clustered in between them. Figure 5C also shows examples of pulse waveform for each rhythm class, illustrating how certain ectopic rhythms are hard to distinguish from AF.

DISCUSSION

To our knowledge, this is the first study to validate the use of a deep learning system to detect AF from a raw PPG waveform. These results demonstrate that the DCNN achieved generalisable detection of AF from short pulse waveforms without having to specify explicit rules or features. The DCNN achieved a very high sensitivity and specificity for AF detection that exceeded other state-of-the-art methods using handcrafted features, and is comparable with automated AF detectors using single-lead ECGs (sensitivity 94%–99% and specificity 92%–97%).[24] Repeated measurements of the PPG waveform improved all metrics of

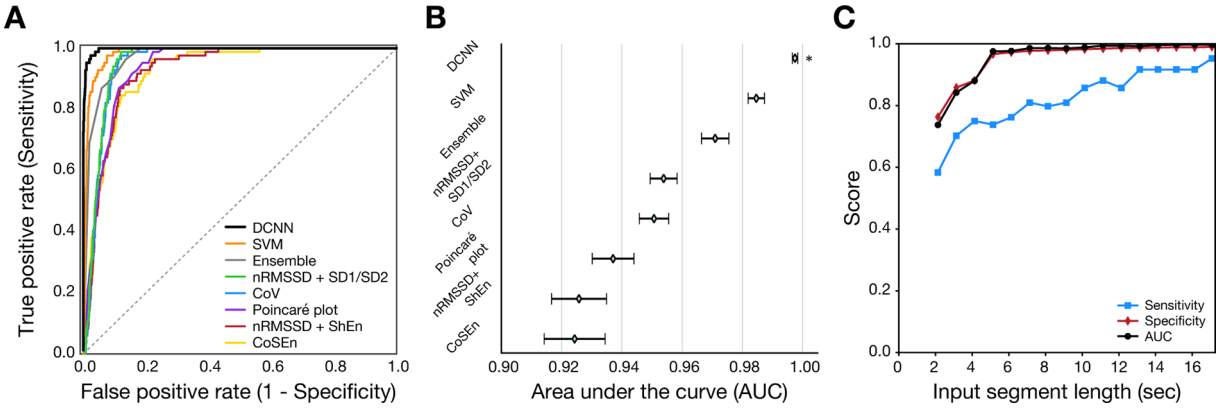

Figure 2 Receiver operating characteristic curves and area under the curves of the DCNN versus other state-of-the-art AF detectors on the MOBILE-SCREEN-AF (clinical validation) data set. (A) Receiver operating curves of several validated AF detectors and (B) corresponding area under the curve values on the MOBILE-SCREEN-AF test set for single measurements. (C) Effect of input segment length on the DCNN performance. *Indicates statistical significance (p<0.001). AF, atrial fibrillation; CoSEn, coefficient of sample entropy; CoV, coefficient of variation; DCNN, deep convolutional neural network; nRMSSD, normalised root mean square of successive differences; ShEn, Shannon entropy; SD1/SD2, Poincaré plot geometry; SVM, support vector machine.

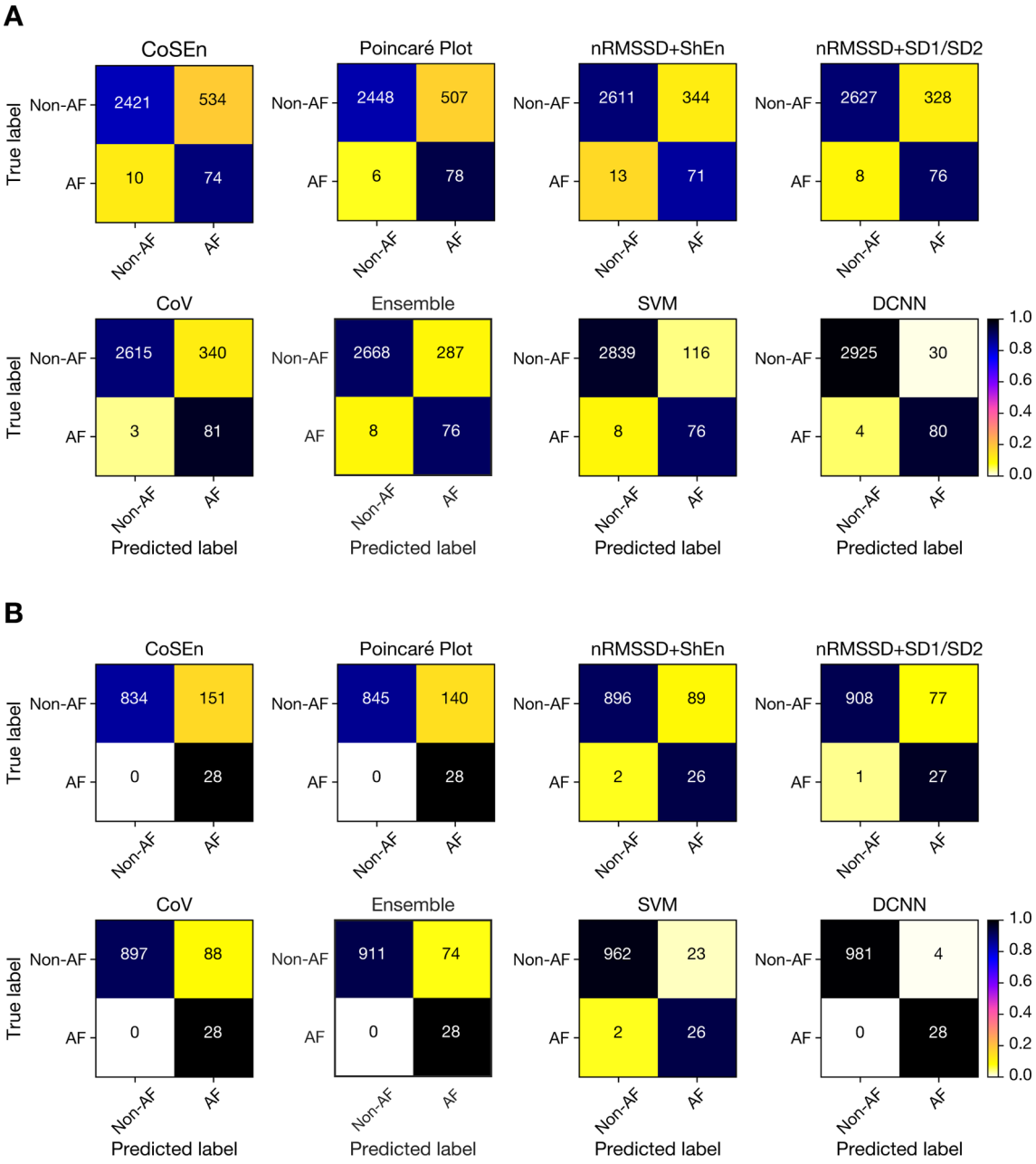

Figure 3 Comparison of contingency tables between the DCNN versus other state-of-the-art AF detectors. Contingency tables for the DCNN and other state-of-the-art AF detectors for the binary classification task of detecting AF based on (A) a single pulse waveform measurement and (B) triplicate measurements. Each column *x* of the contingency table represents the instances in a predicted rhythm, while each row *y* represents the instances in an actual rhythm. The colour of each cell (*x*, *y*) in each contingency table represents the empirical probability of a given AF detector predicting rhythm *x* given that the ground truth was rhythm *y*. For example, the colour of the cell in the first row, first column of the first table in (A) represents the probability of the CoSEn-based AF detector predicting non-AF when the actual rhythm is indeed non-AF (ie, specificity). It is coloured blue because 2421/(2421+534)=0.82. The contingency table of a perfect classifier would have diagonals in black and all other cells in white. Performance of all AF detectors improved when triplicate measurements were used for classification. The DCNN achieved the highest accuracy among all AF detectors. AF, atrial fibrillation; CoSEn, coefficient of sample entropy; CoV, coefficient of variation; DCNN, deep convolutional neural network; nRMSSD, normalised root mean square of successive differences; ShEn, Shannon entropy; SD1/SD2, Poincaré plot geometry; SVM, support vector machine.

diagnostic performance, consistent with previous findings.[25] The ability of the DCNN to learn directly from PPG waveforms and outperform AF detectors based on explicit features underlines the value of information captured by raw data that may be discarded when using handcrafted features. Additionally, the learnt DCNN model showed strong generalisation to pulse waveforms acquired using smartphones despite being trained on examples collected using conventional pulse oximeters.

Previously, Shashikumar *et al*[26] reported a convolutional neural network (CNN) with a lower AUC of 0.92 and accuracy of 85.8% for detecting AF using spectrogram images derived from PPG signals instead of using the PPG waveforms directly as input to the CNN. The discriminative power of the prior CNN

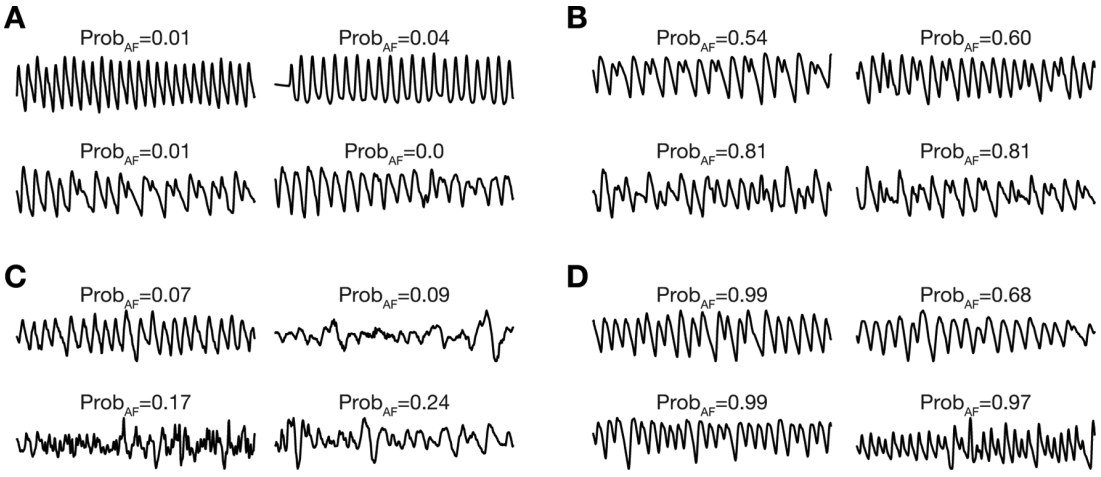

Figure 4 Example of pulse waveforms correctly and incorrectly classified by the DCNN. Examples of (A) true negatives, (B) false positives, (C) false negatives and (D) true positives, along with the probability of AF being present in the pulse waveform produced by the deep learning model (Prob$_{AF}$). AF, atrial fibrillation; DCNN, deep convolutional neural network.

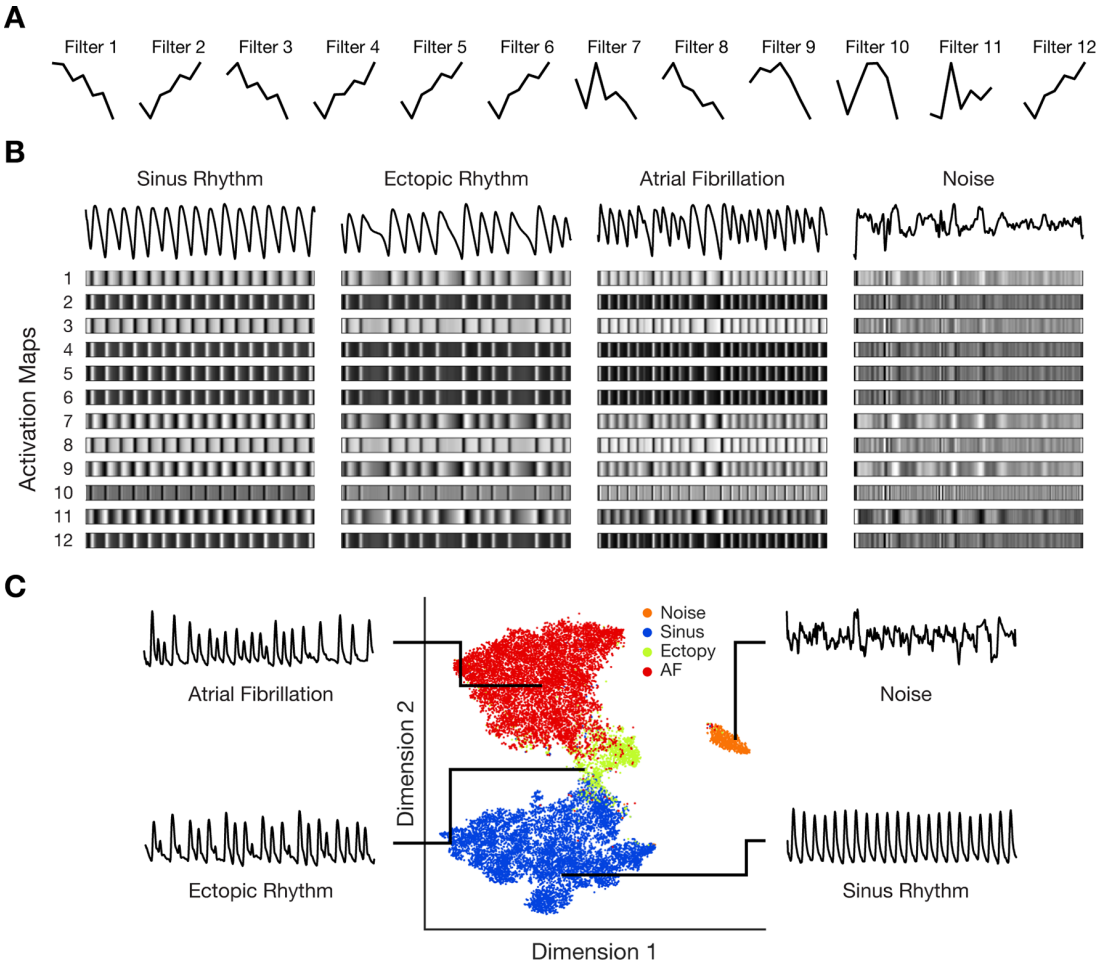

Figure 5 Visualising what the DCNN learns. (A) Learnt filters (first-layer weights) of the DCNN. (B) Layer activations. Examples of how pulse waveforms from the four different rhythm classes activate the neurons of the first convolutional layer of the DCNN. The activation maps represent the result of applying the learnt filters to the input pulse waveform. The position of a pixel in the activation map corresponds to the same position in the corresponding pulse waveform. White pixels represent strong positive activations, while black pixels represent strong negative activations at that position. (C) t-SNE visualisation. Each point in the t-SNE map represents an individual pulse waveform projected from the output of the DCNN's last hidden layer into two dimensions (of arbitrary units). The coloured clusters represent the different rhythm classes: sinus rhythm (blue), ectopy (green), noise (orange) and AF (red). Insets show examples of pulse waveforms from the different rhythm classes. AF, atrial fibrillation; DCNN, deep convolutional neural network; t-SNE, t-distributed stochastic neighbour embedding.

may have been limited by the potential loss of information when converting a PPG waveform into a spectrogram image, the small training set (98 patients) and a comparatively shallow network architecture (six layers).

There are areas where the DCNN can improve. For example, the sensitivity for ectopy detection is lower as the two most difficult rhythm classes for the DCNN to distinguish between were ectopy and AF. This could be due in part to the imbalance in rhythm class representations during training (the ratio of ectopic to AF pulse waveforms was around 1:5). Using a more balanced data set, increasing the total number of training examples or increasing the input segment length (eg, 1–5 min) may lead to performance gains.

Considering the significantly elevated stroke risk associated with AF, a high sensitivity is the primary requisite of a screening tool. At the same time, a high specificity and PPV is particularly desirable for mass screening programmes to avoid triggering unnecessary anxiety in people and prevent avoidable costs of follow-up investigations. The ability of the DCNN to achieve both high sensitivity and specificity is promising for precise screening of AF in a real-world primary care setting. Although the European Society of Cardiology AF guidelines recommend opportunistic pulse palpation in all patients ≥65 years of age (or in high-risk subgroups) followed by an ECG if irregular,[27] pulse taking is not common practice in routine primary care and has a lower sensitivity of 87.2% and specificity of 81.3%.[28] Using the DCNN to screen for AF from smartphone-acquired or pulse oximeter-acquired PPG may be an attractive replacement for pulse palpation given its ease of use and superior accuracy. Pairing AF screening programmes with an existing workflow in primary care and community pharmacies such as influenza vaccination is preferred for scalability, sustainability and cost savings.[24 29] Beyond this, we anticipate that the DCNN may be built into various consumer devices with PPG capabilities including smartphone apps, wearable fitness trackers and smartwatches.

Key messages

What is already known on this subject?

- Photoplethysmography (PPG) offers an attractive method for detecting atrial fibrillation (AF) from pulse waveforms given the rising popularity of smartphone applications and wearable fitness trackers that use it to measure heart rate.
- Prior methods for automated detection of AF in PPG pulse waveforms are predominantly based on explicit rules and handcrafted features derived from beat-to-beat intervals.

What might this study add?

- In this evaluation of pulse waveforms from adults screened for AF in a real-world primary care setting, we found that a deep learning system that automatically learns the most predictive features directly from the pulse waveform based on the training examples outperformed six other state-of-the-art methods based on handcrafted features for AF detection.

How might this impact on clinical practice?

- Application of a deep learning system may improve diagnostic accuracy for automated screening of AF from pulse waveforms.

Limitations

There are limitations to this system. Currently, there is no mechanism for clinicians to over-read PPG waveforms, and an ECG is still required to confirm AF. On the other hand, pulse-based detection systems are attractive for AF screening given the wide accessibility of smartphones, smartwatches and fitness bands. A 12-lead ECG was not available for every PPG waveform in the clinical validation data set given time and cost constraints. The reliance on a single-lead I ECG to provide a reference diagnosis may have resulted in false negatives. However, all single-lead ECG tracings were reviewed by two cardiologists and all patients identified to have AF received a confirmatory 12-lead ECG. Although the PPG recordings in the clinical validation data set were only collected with an iPhone, it is unlikely that the accuracy is dependent on the hardware given the ability of the DCNN to generalise, provided a PPG recording of sufficient signal quality can be obtained. As with all automated AF detection algorithms, a low-noise, high-quality recording is needed for optimal performance. The PPG recordings in the clinical validation data set were performed under supervision by trained personnel. However, the DCNN only requires a recording as short as 17 s compared with other PPG-based detectors that need 2–5 min recordings,[7 8] making it easier to obtain a noise-free waveform. Future work should test and optimise the DCNN's performance on unsupervised PPG recordings collected in the wild (eg, at home or work).

CONCLUSIONS

In this evaluation of smartphone-acquired pulse waveforms from adults at high risk of AF in a primary care setting, the DCNN achieved better diagnostic performance than six other state-of-the-art AF detectors based on handcrafted features. Further studies are needed to evaluate the DCNN in long-term ambulatory setting and determine its utility for clinical decision making and improving patient outcomes.

Contributors M-ZP and YCP designed the study. LP, C-KW, WW-CL, Y-FW, MM-YW and DW-SC contributed to data acquisition. C-WS, MP-HC, C-KW, M-ZP and YCP contributed to analysis and interpretation of data. M-ZP and YCP drafted the manuscript. C-WS, MP-HC and C-KW contributed to critical revision of the manuscript for important intellectual context. All authors reviewed the manuscript and approved the final version for publication.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests M-ZP and YCP are employees of Cardiio and have an ownership stake in the company, which holds intellectual property rights to the new algorithm tested in this work.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Friberg L, Rosenqvist M, Lindgren A, *et al*. High prevalence of atrial fibrillation among patients with ischemic stroke. *Stroke* 2014;45:2599–605.
2 Healey JS, Connolly SJ, Gold MR, *et al*. Subclinical atrial fibrillation and the risk of stroke. *N Engl J Med* 2012;366:120–9.
3 Hart RG, Benavente O, McBride R, *et al*. Antithrombotic therapy to prevent stroke in patients with atrial fibrillation: a meta-analysis. *Ann Intern Med* 1999;131:492–501.
4 Sposato LA, Cipriano LE, Saposnik G, *et al*. Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis. *Lancet Neurol* 2015;14:377–87.
5 Tateno K, Glass L. Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and deltaRR intervals. *Med Biol Eng Comput* 2001;39:664–71.

6 Lake DE, Moorman JR. Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices. *Am J Physiol Heart Circ Physiol* 2011;300:H319–25.
7 McManus DD, Lee J, Maitas O, *et al*. A novel application for the detection of an irregular pulse using an iPhone 4S in patients with atrial fibrillation. *Heart Rhythm* 2013;10:315–9.
8 Krivoshei L, Weber S, Burkard T, *et al*. Smart detection of atrial fibrillation. *Europace* 2016;19:euw125–757.
9 Sarkar S, Ritscher D, Mehra R. A detector for a chronic implantable atrial tachyarrhythmia monitor. *IEEE Trans Biomed Eng* 2008;55:1219–24.
10 Chan PH, Wong CK, Poh YC, *et al*. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *J Am Heart Assoc* 2016;5:e003428.
11 Johnson AE, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
12 Charlton PH, Bonnici T, Tarassenko L, *et al*. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiol Meas* 2016;37:610–26.
13 Karlen W, Raman S, Ansermino JM, *et al*. Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Trans Biomed Eng* 2013;60:1946–53.
14 Li Q, Clifford GD. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiol Meas* 2012;33:1491–501.
15 January CT, Wann LS, Alpert JS, *et al*. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation* 2014:130:2071–104.
16 Huang G, Liu Z, Weinberger KQ, *et al*. Densely connected convolutional networks. *arXiv preprint arXiv* 2016:160806993.
17 He K, Zhang X, Ren S, *et al*. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. IEEE Int Conf Comput Vis. 2015.
18 Sutskever I, Martens J, Dahl G, *et al*, 2013. On the importance of initialization and momentum in deep learning. Int Conf Mach Learn.
19 Smith LN. Cyclical learning rates for training neural networks. IEEE Winter Conf Appl Comp Vis (WACV). 2017.
20 Lvd M, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579–605.
21 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
22 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934;26:404–13.
23 Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med* 2007;26:2170–83.
24 Freedman B, Camm J, Calkins H, *et al*. Screening for Atrial Fibrillation. *Circulation* 2017;135:1851–67.
25 Wiesel J, Fitzig L, Herschman Y, *et al*. Detection of atrial fibrillation using a modified microlife blood pressure monitor. *Am J Hypertens* 2009;22:848–52.
26 Shashikumar SP, Shah AJ, Li Q, *et al*. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. IEEE EMBS Int Conf Biomed & Health Inf (BHI). 2017.
27 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;37:2893–962.
28 Hobbs FD, 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;9:93.
29 Jacobs MS, Kaasenbrood F, Postma MJ, *et al*. Cost-effectiveness of screening for atrial fibrillation in primary care with a handheld, single-lead electrocardiogram device in the Netherlands. *Europace* 2018;20:euw285.