Prioritising electrocardiograms for manual review to improve the efficiency of atrial fibrillation screening*

Mary Adeniji, James Brimicombe, Martin R. Cowie, Andrew Dymond, Hannah Clair Lindén, Gregory Y. H. Lip, Jonathan Mant, Madhumitha Pandiaraja, Kate Williams and Peter H. Charlton, on behalf of the SAFER Investigators

Abstract— **Screening for atrial fibrillation (AF) could reduce the incidence of stroke by identifying undiagnosed AF and prompting anticoagulation. However, screening may involve recording many electrocardiograms (ECGs) from each participant, several of which require manual review which is costly and time-consuming. The aim of this study was to investigate whether the number of ECG reviews could be reduced by using a model to prioritise ECGs for review, whilst still accurately diagnosing AF. A multiple logistic regression model was created to estimate the likelihood of an ECG exhibiting AF based on the mean RR-interval and variability in RR-intervals. It was trained on 1,428 manually labelled ECGs from 144 AF screening programme participants, and evaluated using 11,443 ECGs from 1,521 participants. When using the model to order ECGs for review, the number of reviews for AF participants was reduced by 76% since no further reviews are required after an AF ECG is identified; however, it did not impact the number of reviews in non-AF participants (the vast majority of participants), so the overall number of reviews was reduced by 3% with no missed AF diagnoses. When using the model to also exclude ECGs from review, the overall number of reviews was reduced by 27% with no missed AF diagnoses, and by 52% with only 2% of AF diagnoses missed. In conclusion, the workload can be reduced by using a model to prioritise ECGs for review. Ordering ECGs alone only provides only a moderate reduction in workload. The additional use of a threshold to exclude ECGs from review provides a much greater reduction in workload at the expense of some missed AF diagnoses.**

Clinical Relevance—**This shows the potential benefit of using a model to prioritise electrocardiograms for review in order to reduce the manual workload of AF screening.**

I. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia. It is associated with a fivefold increase in stroke risk, and is associated with over a quarter of ischemic strokes [1], and increasing healthcare costs [2]. Fortunately, the risk of stroke can be reduced through anticoagulation if AF is recognised. However, AF is often unrecognised, leaving patients unnecessarily exposed to an increased stroke risk [3]. Consequently, screening for AF is being investigated as an approach to identify AF at scale.

Current approaches to AF screening often involve taking several short electrocardiogram (ECG) recordings, which must be manually reviewed to diagnose AF [4]. This approach allows even infrequent AF episodes to be identified, which is important as AF can occur only intermittently. However, it results in a large manual review workload, as each individual records about 20-100 ECGs, depending on the screening programme design [4], [5]. An automated algorithm can be used to identify abnormal ECGs which require manual review, and exclude the remainder. However, even when using an automated algorithm, approximately 35 abnormal ECGs had to be reviewed to identify each pathological ECG in a recent study [6]. Therefore, the manual review workload associated with AF screening remains high. Strategies to reduce this workload could reduce the cost of AF screening and therefore make it more cost-effective.

A potential approach to reduce the manual review workload is to use a model to order an individual's ECGs for review according to their likelihood of exhibiting AF. This could reduce the number of ECGs reviewed, since no further reviews are required for an individual when an ECG exhibiting AF is identified, as an AF diagnosis can be made on the basis of a single ECG. This approach ensures that all individuals who have an AF ECG sent for review would be identified. The approach could be extended by excluding ECGs from review which do not meet a threshold likelihood of AF, although this could result in missing AF diagnoses.

The aim of this study was to investigate whether a model could be used to reduce the number of manual ECG reviews whilst still accurately diagnosing AF. A model was designed to quantify the likelihood of an ECG exhibiting AF based on the ECG's characteristics. The potential benefits of the model were assessed when using the model to simply order ECGs, and also when using it with a threshold to exclude ECGs from review. The number of reviews and the accuracy of AF

*This study is funded by the National Institute for Health Research (NIHR), grant number RP-PG-0217-20007; and the British Heart Foundation (BHF), grant number FS/20/20/34626. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

M. Adeniji, M. Pandiaraja, J. Brimicombe, A. Dymond, J. Mant, K. Williams, and P.H. Charlton are with the Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK (e-mail: pc657@medschl.cam.ac.uk).

M. R. Cowie, is with the Royal Brompton Hospital (Guy's and St Thomas' NHS Foundation Trust), Sydney Street, London, SW3 6NP, UK.

H.C. Lindén is with Zenicor Medical Systems AB, 113 59 Stockholm, Sweden.

G.Y.H. Lip, is with the Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart and Chest Hospital, Liverpool L69 7TX, UK.

diagnoses when using each of these approaches were compared to the current approach of reviewing ECGs chronologically.

II. Methods

A. Dataset

The data used in this study were collected in the SAFER Feasibility Study (ISRCTN 16939438), a study to assess the feasibility of delivering an AF screening programme in Primary Care, approved by the London - Central Research Ethics Committee (REC ref: 18/LO/2066). Briefly, 2,141 participants aged 65 and over were screened for AF. Participants were asked to record 30-second ECGs four times per day, for a period of 1-4 weeks. ECGs were acquired between two thumbs using the Zenicor EKG-2 device (Zenicor Medical Systems AB), as shown in Fig. 1. Each participant recorded a median (lower-upper quartiles) of 61 (53-111) ECGs, resulting in a total of 162,515 ECGs. Participants were then allocated a diagnosis of either AF or non-AF by using an automated algorithm to identify ECGs with abnormalities for review, and then clinicians manually reviewing these ECGs [5].

The ECGs used for this study were those which would be sent for manual review in an AF screening programme. The ECGs were identified using the same approach as in the SAFER Trial of AF screening (ISRCTN 72104369): the Cardiolund ECG Parser algorithm (Cardiolund AB) [7] was used to identify ECGs which exhibited either an 'irregular rhythm' or a 'fast, regular rhythm'. This approach has been found to identify most ECGs exhibiting AF: the algorithm's 'irregular rhythm' classification identifies approximately 90% of AF ECGs, and its 'fast, regular rhythm' classification identifies many of the remainder [6]. This resulted in 11,975 ECGs from 1,538 participants (including 65 participants in whom AF was identified). ECGs recorded from any participants diagnosed with AF for whom there wasn't at least one ECG labelled as AF by clinicians were excluded. This resulted in 11,443 ECGs for model evaluation consisting of 1,613 ECGs from 48 AF participants and 9,830 ECGs from 1,473 non-AF participants.

A subset of the ECGs were used to train and validate the model, consisting of 1,428 ECGs from 144 participants. The AF ECGs in this subset were those which were labelled as AF during manual review, and recorded by AF participants. This resulted in 687 AF ECGs from 48 AF participants. The non-AF ECGs in this subset were those recorded from non-AF participants in which the algorithm found no abnormalities, and those for which two independent Cardiologists provided a non-AF label. This resulted in 741 non-AF ECGs from 111 participants.

B. Model training and validation

First, ECG characteristics associated with AF were identified as candidate inputs to the model. The Cardiolund algorithm extracted characteristics from the RR intervals in each ECG – the time intervals between consecutive R waves, indicating heartbeats. It extracted heart rate (HR, in beats per minute, bpm) and RR interval variability (RRvar = (RRstd/RRmean) × 100%, %). From these, the mean RR interval (RRmean, ms) and standard deviation of RR intervals (RRstd, ms) were calculated. The values of each characteristic in AF and non-AF ECG groups were expressed as median (lower-upper quartiles). The values were compared between groups using the Mann-Whitney test.

Figure 1. The Zenicor EKG-2 device used to acquire single-lead, 30-second ECG recordings in the SAFER Feasibility Study.

Second, a multiple logistic regression model was developed to quantify the likelihood of each ECG exhibiting AF. Stepwise regression was used to select which candidate inputs to include in the model. The performance of the model for identifying AF ECGs was assessed on the subset of 1,428 ECGs, using 5-fold cross-validation, with data split at the participant level. Performance was reported as the area under the receiver-operating curve (AUROC).

C. Model evaluation

First, the potential utility of the model for ordering ECGs was evaluated. The number of reviews for each participant was calculated as the number of reviews required when reviewing ECGs ordered from most to least likely to exhibit AF, until either an ECG exhibiting AF was reviewed, or all ECGs for that participant had been reviewed.

Second, the potential utility of the model for excluding ECGs from review was assessed. Similarly, the number of reviews for each participant was calculated when reviewing ECGs in descending order of likelihood of AF. However, with this approach, reviewing stopped when either an ECG exhibiting AF was reviewed, or when there were no more ECGs with a likelihood above a selected threshold. The number of participants who would have been diagnosed with AF was also calculated: participants who would have had at least one AF ECG reviewed were deemed to have been diagnosed with AF.

Model evaluation was performed using the same 5-fold cross-validation as used in training and testing.

III. Results

A. Model training and validation

Most of the ECG characteristics assessed (HR, RRmean, and RRvar) differed significantly between AF and non-AF ECGs in all of the cross-validation folds. RRstd only differed significantly in two of the five folds. Table I shows the comparison of ECG characteristics on the entire dataset, in which HR, RRmean and RRvar differed significantly

TABLE I. COMPARISON OF ECG CHARACTERISTICS

ECG Characteristic	Value, median (lower-upper quartiles)	
	AF	***non-AF***
HR (bpm)	83.0 (71.0 - 94.0)	69.0 (59.0 - 78.0)
RRmean (ms)	722.9 (638.3 - 845.1)	869.6 (769.2 – 1016.9)
RRstd (ms)	120.7 (87.5 - 169.5)	131.2 (90.2 - 187.6)
RRvar (%)	16.5 (12.8 - 22.3)	14.3 (10.0 – 20.4)

($p<0.0001$ in all cases), whilst RRstd did not. HR was higher in AF than non-AF, whilst RRmean and RRstd were lower.

Only a subset of ECG characteristics were selected for inclusion in the model generated for each cross-validation fold. RRvar was included in all five models, RRmean in four models, HR in two models, and RRstd in one model. The models achieved a median AUROC for identifying AF ECGs of 67.4% when trained and assessed using cross-validation. The optimal model for the entire dataset was $\text{logit}(AF) = 3.8 + 0.061 \times RRvar - 0.006 \times RRmean$.

B. Model evaluation

Table II shows the results relating to the model evaluation. Without the model, a total of 10,293 reviews would have been required to identify the 48 participants with AF (9,830 reviews for non-AF participants, and 463 reviews for AF participants). This assumes ECGs are reviewed in the order in which they were measured. When using a model to order ECGs (without a threshold), the number of reviews for AF participants was reduced by 76% from 463 to 109, without any reduction in the number of AF diagnoses. This demonstrates the potential utility of the model for reducing the workload associated with AF participants whilst maintaining the accuracy of AF diagnoses. However, since the vast majority of participants were not diagnosed with AF, and the number of reviews for non-AF participants remained the same, the overall number of reviews remained high at 9,939 (a reduction of 3.4%).

The use of a threshold for excluding ECGs from review resulted in a reduction in workload at the expense of potentially missing AF diagnoses. For instance, a 25% threshold (designed to reduce the number of ECGs by 25%) resulted in a 28% reduction in the number of ECG reviews, whilst still identifying all participants with AF. Higher thresholds resulted in much greater reductions in workload, such as the 50% threshold reducing the number of reviews by 53%, whilst still identifying 98% of AF participants.

TABLE II. THE POTENTIAL UTILITY OF THE MODEL

Model configuration	AF diagnoses (%)	Number of reviews (per AF diagnosis)
No model	48 (100)	10,293 (214)
Model + no threshold	48 (100)	9,939 (207)
Model + 25% threshold	48 (100)	7,392 (154)
Model + 50% threshold	47 (98)	4,843 (103)
Model + 75% threshold	43 (90)	2,452 (57)

IV. DISCUSSION

Screening for AF using intermittent ECG recording holds promise for identifying undiagnosed AF and potentially reducing the incidence of stroke through anticoagulation. This study shows that the workload associated with the manual review of ECGs can be reduced by using a model to order and select ECGs for review. This could potentially reduce the costs of screening, making it more cost-effective.

In this study a model was firstly used to reduce the number of ECG reviews in AF participants by ordering ECGs for review according to the likelihood of them exhibiting AF. This approach is safe, resulting in no missed AF diagnoses. However, since the vast majority of individuals screened do not have AF, this approach provided only a moderate reduction in the total number of reviews. Secondly, the model was used to provide far greater reductions in workload by excluding ECGs from review which do not meet a threshold likelihood of exhibiting AF. When this threshold is chosen appropriately, a large reduction in workload (such as 50%) can be achieved whilst still correctly identifying almost all AF participants.

The model used simple ECG characteristics to determine the likelihood of an ECG exhibiting AF. The mean RR-interval and the variability in RR-intervals were the key characteristics used. These were calculated from the timings of heartbeats. There are well-established ECG signal processing techniques for detecting heartbeats, ensuring that this approach could be used in practice. In the future, additional ECG characteristics (such as P-wave features) could be incorporated into more complex models (such as deep learning models) to improve performance further [8].

A. Significance

The approaches presented here could be used to reduce the time required to manually review ECGs in AF screening, and therefore reduce the costs of AF screening. It has been estimated that each ECG review takes approximately 20s [6]. When coupled with the large number of ECGs requiring review, this indicates that approaches to reduce the number of manual reviews could significantly reduce the cost of screening.

This study provides evidence to support safely reducing the number of ECGs by ordering them for review (*i.e.* without a threshold). However, it is not yet clear whether the additional use of a threshold to exclude ECGs from review would be acceptable. Further evidence is required on the potential benefits and harms of this approach, since it would inevitably result in missed diagnoses. AF screening programmes which use intermittent ECG recordings already contain compromises on the methodology in order to make screening acceptable to patients and cost-effective. For instance, the duration of screening (typically 2-3 weeks), the number of recordings taken per day (typically 2-4), and the use of an automated algorithm to identify ECGs for review (with approximately 98% sensitivity [6]) could all be expected to influence AF detection. Similarly, it would be helpful to consider whether the use of a threshold to exclude ECGs from review would be cost-effective.

B. Limitations

The key limitation to this study is that not all ECGs recorded in the screening study could be included in the analysis. 10% of ECGs were excluded because they had not been manually labelled as AF or non-AF. Whilst the resulting dataset was still substantial (11,443 ECGs from 1,521 participants), this approach excluded approximately a quarter of AF participants from the analysis. Had these been included, then we expect the reduction in workload observed would have been greater, as the greatest reduction was in AF participants.

C. Future Work

Further work is required to determine whether the approaches presented here could be used beneficially in AF screening. First, the performance of the approaches should be assessed prospectively, potentially in ongoing AF screening trials such as the SAFER Trial. Second, the approaches could be refined to improve performance. Potential refinements include: (i) extending the set of candidate model inputs to include P-wave characteristics; (ii) using machine learning or deep learning methods to estimate the likelihood of AF; and (iii) refining the criteria used to determine whether an ECG is sent for review. Third, the approaches could be combined with other approaches suggested for reducing the number of manual reviews, such as identifying transient noise in ECGs [9].

V. Conclusion

This study demonstrates that approaches to order and select ECGs for review could reduce the manual review workload in AF screening. The use of a model to order ECGs for review according to their likelihood of exhibiting AF could reduce the costs of AF screening whilst ensuring that all AF ECGs sent for review are correctly identified. The use of a model to select ECGs for review would substantially reduce the manual review workload at the expense of missing some AF diagnoses. Further work is required to determine whether this second approach would be cost-effective.

Conflicts of Interest

H.C.L. is employed by Zenicor Medical Systems AB.

References

[1] K. S. Perera *et al.*, "Global survey of the frequency of atrial fibrillation–associated stroke," *Stroke*, vol. 47, no. 9, pp. 2197–2202, 2016, doi: 10.1161/STROKEAHA.116.013378.

[2] P. Burdett and G. Y. H. Lip, "Atrial fibrillation in the UK: predicting costs of an emerging epidemic recognizing and forecasting the cost drivers of atrial fibrillation-related costs," *Eur. Hear. J. - Qual. Care Clin. Outcomes*, vol. [In Press], 2021, doi: 10.1093/ehjqcco/qcaa093.

[3] Public Health England, "Atrial fibrillation prevalence estimates in England: Application of recent population estimates of AF in Sweden," PHE Publications Gateway Number 2014778, 2015.

[4] E. Svennberg, J. Engdahl, F. Al-Khalili, L. Friberg, V. Frykman, and M. Rosenqvist, "Mass screening for untreated atrial fibrillation: the STROKESTOP study," *Circulation*, vol. 131, no. 25, pp. 2176–2184, 2015, doi: 10.1161/CIRCULATIONAHA.114.014343.

[5] M. Pandiaraja *et al.*, "Screening for atrial fibrillation: Improving efficiency of manual review of handheld electrocardiograms," *Eng. Proc.*, vol. 2, no. 1, p. 78, 2020, doi: 10.3390/ecsa-7-08195.

[6] E. Svennberg *et al.*, "Safe automatic one-lead electrocardiogram analysis in screening for atrial fibrillation," *Europace*, vol. 19, no. 9, pp. 1449–1453, 2017, doi: 10.1093/europace/euw286.

[7] M. Stridh and M. Rosenqvist, "Automatic Screening of Atrial Fibrillation in Thumb-ECG Recordings," in *Proc CinC.*, 2012, pp. 193–196, [Online]. Available: http://cinc.mit.edu/archives/2012/pdf/0193.pdf.

[8] Z. I. Attia *et al.*, "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *Lancet*, vol. 394, no. 10201, pp. 861–867, 2019, doi: 10.1016/s0140-6736(19)31721-0.

[9] H. Halvaei, E. Svennberg, L. Sörnmo, and M. Stridh, "Identification of Transient Noise to Reduce False Detections in Screening for Atrial Fibrillation," *Front. Physiol.*, vol. 12, no. June, pp. 1–10, 2021, doi: 10.3389/fphys.2021.672875.