Archives of Cardiovascular Disease xxx (xxxx) xxx–xxx

Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Clinical Research

Clinical validation of an artificial intelligence algorithm offering cross-platform detection of atrial fibrillation using smart device electrocardiograms☆

Diego Mannhart[a], Baptiste Lefebvre[b], Christophe Gardella[b], Christine Henry[b], Teodor Serban[a], Sven Knecht[a], Michael Kühne[a], Christian Sticherling[a], Patrick Badertscher[a,*]

[a] University Hospital of Basel, 4031 Basel, Switzerland
[b] Cardiologs, 75002 Paris, France

ARTICLE INFO

Article history:
Received 19 December 2022
Accepted 19 April 2023
Available online xxx

Keywords:
Atrial fibrillation
Smartwatch
Artificial intelligence
Deep neural network
Digital health

ABSTRACT

Background: Several smart devices are able to detect atrial fibrillation automatically by recording a single-lead electrocardiogram, and have created a work overload at the hospital level as a result of the need for over-reads by physicians.
Aim: To compare the atrial fibrillation detection performances of the manufacturers' algorithms of five smart devices and a novel deep neural network-based algorithm.
Methods: We compared the rate of inconclusive tracings and the diagnostic accuracy for the detection of atrial fibrillation between the manufacturers' algorithms and the deep neural network-based algorithm on five smart devices, using a physician-interpreted 12-lead electrocardiogram as the reference standard.
Results: Of the 117 patients (27% female, median age 65 years, atrial fibrillation present at time of recording in 30%) included in the final analysis (resulting in 585 analyzed single-lead electrocardiogram tracings), the deep neural network-based algorithm exhibited a higher conclusive rate relative to the manufacturer algorithm for all five models: 98% vs. 84% for Apple; 99% vs. 81% for Fitbit; 96% vs. 77% for AliveCor; 99% vs. 85% for Samsung; and 97% vs. 74% for Withings ($P < 0.01$, for each model). When applying our deep neural network-based algorithm, sensitivity and specificity to correctly identify atrial fibrillation were not significantly different for all assessed smart devices.
Conclusion: In this clinical validation, the deep neural network-based algorithm significantly reduced the number of tracings labeled inconclusive, while demonstrating similarly high diagnostic accuracy for the detection of atrial fibrillation, thereby providing a possible solution to the data surge created by these smart devices.

1. Background

Atrial fibrillation (AF) represents the most common cardiac arrhythmia. The risk of a stroke associated with AF can be significantly reduced through oral anticoagulation [1]. According to the 2020 European Society of Cardiology guidelines [2], diagnosis of AF is possible by electrocardiographic documentation using a standard 12-lead electrocardiogram or with a single-lead electrocardiogram tracing of at least 30 seconds. Several wearable smart devices capable of detecting AF are presently available, and more are expected to enter the market soon [3]. It is anticipated that the use of smart devices will increase from 325 million connected devices in 2016 to 1.1 billion devices worldwide by 2023 [4]. With public health concerns, such as the coronavirus disease 2019 (COVID-19) pandemic, remote contact between patients and healthcare providers will further accelerate this transformation [5]. Despite this increasing adoption of smart devices, multiple uncertainties remain.

Abbreviations: AF, atrial fibrillation; AI, artificial intelligence; CI, confidence interval; DNN, deep neural network; PDF, portable document format; SR, sinus rhythm.

☆ Tweet: The inconclusive rate of commercially available smartwatches is as high as 25%. Is AI able to improve AF detection? Check out your device agnostic AI algorithm offering cross-platform detection of AF #EPeeps #Smartwatches #AF #Wearables #AI. Twitter handle: @BadertscherPat, @Manndie.

* Corresponding author at: University Hospital of Basel, Petersgraben 4, 4031 Basel, Switzerland.
E-mail address: patrick.badertscher@usb.ch (P. Badertscher).

https://doi.org/10.1016/j.acvd.2023.04.003
1875-2136/© 2023 Published by Elsevier Masson SAS.

A common clinical issue with smart watch technologies is that automated rhythm interpretation is conservative, and therefore many automated diagnoses are deemed inconclusive, despite yielding a readable single-lead electrocardiogram tracing [6]. To overcome this issue, manual clinician interpretation has been suggested. As these smart devices incorporate a paradigm shift in the healthcare system by generating consumer-initiated instead of clinician-initiated diagnostics, data overload is a major concern. Smart devices may substantially increase the workload of a cardiology service already under significant pressure.

There has not, to our knowledge, been any use of a device agnostic artificial intelligence (AI) algorithm to analyze single-lead electrocardiogram tracings recorded from different devices. Cardiologs (Cardiologs Technologies, Paris, France) has developed a proprietary algorithm that is based on deep neural networks (DNNs), and has been trained to detect QRS complexes and ventricular ectopic beats, extract QRS features, measure heart rates and intervals and, most importantly, determine heart rhythms, such as AF. This algorithm was developed for arrhythmia detection in Holter electrocardiograms [7]. An adapted version of this algorithm has also been validated on implantable loop recorder [6] electrocardiograms. So far, no attempt has been made to use the Cardiologs AI algorithm on single-lead electrocardiograms from different smart devices.

The aim of this study was to compare the rate of inconclusive tracings and the diagnostic accuracy for the detection of AF between the manufacturers' algorithms and a DNN-based algorithm on five smart devices, using a physician-interpreted 12-lead electrocardiogram as the reference standard.

2. 3. Methods

2.1. Study design and population

Between April 2021 and February 2022, we enrolled 168 subjects presenting to the University Hospital of Basel in a prospective single-centre diagnostic study. Patients included were scheduled for catheter ablation procedures, electric cardioversions or pacemaker or implantable cardioverter defibrillator implantation. Patients had to be aged > 18 years to participate in this study. Participants with paced ventricular rhythm, missing recordings because of logistical or technical issues and patients with heart rhythm change between the 12-lead electrocardiogram and corresponding single-lead electrocardiograms were excluded. The study was approved by the local ethics committee, pre-registered (ClinicalTrial.gov identifier: NCT04809922) and carried out according to the principles of the Declaration of Helsinki. Informed consent was provided by all patients included in the study. The study was designed by the authors. The design, data collection and analysis were conducted according to the STROBE guidelines [8] for observational studies (Table A.1). All authors vouch for the data and analysis, wrote the paper together and made the decision to submit the manuscript for publication.

2.2. Study aim

The primary aim of this study was to compare the rate of inconclusive tracings and the diagnostic accuracy for the detection of AF based on the manufacturers' algorithms and a DNN-based algorithm (capable of interpreting any recording based on a portable document format [PDF] file, derived from the different devices used). The devices used for recording were five wearable smart devices (Apple Watch 6 [Apple Inc., Cupertino, CA, USA], AliveCor Kardia Mobile or Kardia Mobile 6L [AliveCor Inc., Mountain View, CA, USA], Fitbit Sense [Fitbit Inc., San Francisco, CA, USA], Samsung Galaxy Watch 3 [Samsung Inc., Seoul, South Korea] and Withings Scanwatch [Withings SA, Issy les Moulineaux, France]), which are readily commercially available, as well as Conformité Européenne (CE) and Food and Drug Administration (FDA) marked [9]. As gold standard, a nearly simultaneously acquired physician-interpreted 12-lead electrocardiogram was used to compare against the tracings recorded by the smart devices.

2.3. Study assessment

Patients eligible for participation obtained a preprocedural 12-lead electrocardiogram (part of routine clinical care), followed sharply by the five 30-second wearable smart device recordings. Twelve-lead electrocardiograms were recorded with a standard electrocardiogram machine (Schiller SDS-200, Touch 4.4.3; Schiller, Baar, Switzerland) with a sweep speed of 25 mm/s and standard augmentation of 10 mm/mV. Single-lead electrocardiograms from the above-mentioned smart devices were recorded according to instructions provided by the manufacturing company. Regardless of handedness, single-lead electrocardiograms were recorded while wearing the watch on the left wrist, when possible. The manufacturer's diagnosis from the single-lead electrocardiogram (sinus rhythm [SR], AF or unclassified) was registered, and a report of the single-lead electrocardiogram waveform was saved as a PDF file. Existing hospital records were used to complete demographic data and medical history.

2.4. Electrocardiogram interpretation

2.4.1. Physicians' interpretation

Smart device single-lead electrocardiogram recordings and corresponding 12-lead electrocardiogram recordings were exported as PDF files and anonymized. Twelve-lead electrocardiogram recordings were distributed to two blinded cardiologists (T. S. and D. M.). Each 12-lead electrocardiogram recording was evaluated independently, and a diagnosis (SR, AF or inconclusive) was determined. The diagnosis of AF was made according to the definition given by the 2020 European Society of Cardiology guidelines [2]. Disagreements between the two cardiologists' diagnoses were reviewed and assessed by a third cardiologist (P. B.). The physician-based 12-lead electrocardiogram interpretation was used as the gold standard for any single-lead electrocardiogram tracing, and regardless of the algorithm used for the interpretation.

2.4.2. AI interpretation

Single-lead electrocardiogram recordings distributed to the cardiologists needed to be reformatted for analysis by the Cardiologs DNN-based algorithm. PDF files were parsed to extract the corresponding single-lead electrocardiogram waveforms, which were transmitted to the Cardiologs platform through a secured/encrypted connection for assessment of the presence of AF. The graphical representations of the single-lead electrocardiogram waveforms provided by the platform were carefully compared with the graphical representations provided by the PDF files. Reference grids with a common scale (large square: 0.2 s × 0.5 mV; small square: 0.04 s × 0.1 mV) were superimposed on the waveforms to support the visual comparison, making this method equivalent to the superposition of the two signals. For all sampling times, voltage differences were always found to be lower than one small square (Central illustration). This method is the best that could be achieved without direct access to the raw data. Note that in the case of badly parsed signals, the DNN-based algorithm is more likely to indicate recordings as uninterpretable, which would be detrimental to its performance. In other words, the DNN-based algorithm is the only algorithm that would see its performance increased if the parsing procedure could be improved.

The Cardiologs platform is a cloud-based platform that automatically analyses the single-lead electrocardiograms it receives. To remove high frequency artifacts and baseline wander, electrocardiograms are first preprocessed using wavelet filtering. Electrocardiograms are also normalized and resampled at 250 Hz to support various devices with different sampling frequencies. Following this necessary preprocessing filtering, the DNNs then perform a beat detection and a rhythm classification, which enable assessment of the presence or absence of AF.

The platform uses two DNNs, one for wave detection and one for rhythm classification. The heart waves are detected using a DNN with a U-net architecture [10], consisting of 11 convolutional layers and six residual blocks (800,000 parameters). Taking the electrocardiogram signal as input, it outputs the onsets and offsets of P waves, QRS complexes and T waves. For the rhythm, a DNN with a VGG-like architecture [11] is used, consisting of 13 convolutional layers followed by three fully connected layers (4 million parameters); it outputs multiple labels, which correspond to different rhythms (e.g. SR, AF, atrial flutter, other atrial tachycardia, atrioventricular block, ventricular tachycardia, noise).

These DNNs were trained and validated using a dataset of more than 1 million electrocardiograms from an anonymized dataset that had previously been adjudicated by physicians or certified electrocardiogram technicians. The dataset contains 12-lead electrocardiograms and Holter electrocardiograms, and the DNN supports, by design, the analysis of a single lead of these electrocardiograms. These electrocardiograms were acquired from North American, European and Asian independent diagnostic testing facilities, hospitals or public datasets. Stochastic descent, early stopping and dropout were used during training to avoid overfitting. The Keras framework with TensorFlow backend (Google; Mountain View, CA, USA) on K-80 (NVIDIA) graphics processing units were used to implement and train the DNNs. The trained DNN reached 96% sensitivity and 99% specificity for the detection of AF or atrial flutter [12]. The Cardiologs DNN-based algorithm analyzed the five single-lead electrocardiogram tracings, analysis was performed individually and separately per tracing (each tracing was analyzed individually, with no grouped analysis, and there was no connection between tracings from different devices recorded from the same patient).

2.4.3. Performance of AF diagnosis excluding inconclusives

For this analysis, single-lead electrocardiograms that led to an inconclusive diagnosis were excluded from the analysis to get a binary classification (i.e. AF or SR) of these single-lead electrocardiograms for each device and each algorithm. This means that the performance metrics (i.e. accuracy, sensitivity, specificity and positive predictive value) were evaluated on a different subset of subjects for each model.

2.4.4. Performance of AF diagnosis with an intention to diagnose

As a second analysis, single-lead electrocardiograms that led to an inconclusive diagnosis were considered as false positives when the subject was in SR or as false negatives when the subject was in AF, to get binary classifications. This intention-to-diagnose analysis was suggested by Schuetz et al. [13] for the reallocation of diagnostic tests with non-binary results. This means that the performance metrics were evaluated on the exact same set of subjects. Fig. 1 provides an explanatory overview of the reallocation process.

2.5. Statistical analysis

Normality was assessed visually using histograms. Quantitative variables are described by medians (first and third quartiles) because of non-Gaussian distributions. Categorical variables are described by numbers and percentages. For each smart device, conclusive rate, accuracy, sensitivity, specificity and positive predictive value were reported as performance metrics for both algorithms. Non-parametric 95% confidence intervals (CIs; exact method: Clopper-Pearson intervals) were also reported and compared, except for positive predictive value. Differences in conclusive rate, accuracy, sensitivity and specificity (unpaired data) were tested with two-sided proportion Z tests, with a significance level of 5%. The inter-rater correlation for heart rhythm classification between cardiologists was calculated using Cohen's kappa coefficient.

3. Results

3.1. Baseline data

In this prospective single-centre study we enrolled 168 subjects from April 2021 to February 2022. Fifty-one patients were excluded because of paced ventricular rhythm ($n = 12$) or the lack of five available single-lead electrocardiograms ($n = 39$), leaving 117 patients in the final cohort. The median age was 65 years (interquartile range 56–74 years), 27% of patients were female, and in 83% of patients there was a known history of AF. The median CHA_2DS_2-VASc score was 2 (interquartile range 1–3), and 30% of patients presented in AF; among these, six (5%) patients presented in atrial flutter/atrial tachycardia and premature ventricular contractions were present in 12 (10%) patients. Table 1 provides baseline characteristics. The inter-rater correlation (i.e. Cohen's kappa) for heart rhythm classification based on the 12-lead electrocardiogram between cardiologists was 0.94 (95% CI 0.86–1.0).

3.2. Comparison of inconclusive rates between the manufacturers' algorithms and the DNN-based algorithm

Of the 117 patients included in the final analysis, resulting in 585 analyzed single-lead electrocardiogram tracings, the rate of inconclusive tracings was 16% for Apple, 15% for Samsung, 26% for Withings, 19% for Fitbit and 23% for AliveCor. Reasons for inconclusive tracings were: low quality (13%); bradycardia (20%); tachycardia (8%); and no reason stated by the manufacturers' algorithms (59%, (Table A.3). By comparison, the DNN-based algorithm exhibited a higher conclusive rate relative to the manufacturers' algorithm for all five models: 98% vs. 84% for Apple; 99% vs. 81% for Fitbit; 96% vs. 77% for AliveCor; 99% vs. 85% for Samsung; and 97% vs. 74% for Withings ($P < 0.001$ for each model; Fig. 2). Overall, we observed a mean conclusive rate of $98 \pm 1.4\%$ for the DNN-based algorithm and a mean conclusive rate of $80 \pm 4.4\%$ for the manufacturers' algorithms across all five devices (Table 2). Using manual interpretation, the inconclusive rate was 0 (0%) for the 12-lead electrocardiograms.

3.3. Comparison of diagnostic accuracy for the detection of AF between the manufacturers' algorithms and the DNN-based algorithm

Sensitivity and specificity for the manufacturers' algorithms to correctly identify AF were 100% (95% CI 89–100%) and 97% (95% CI 90–99%) for Apple, 96% (95% CI 82–99%) and 100% (95% CI 95–100%) for Fitbit, 100% (95% CI 89–100%) and 98% (95% CI 91–100%) for AliveCor, 100% (95% CI 89–100%) and 93% (95% CI 84–97%) for Samsung and 96% (95% CI 80–99%) and 95% (95% CI 87–98%) for Withings. When applying the DNN-based algorithm, sensitivity and specificity to correctly identify AF were not significantly different from those observed for the manufacturer's algorithms: 100% (95% CI 90–100%) and 98% (95% CI 91–99%) for Apple; 100% (95% CI 90–100%) and 99% (95% CI 93–100%) for Fitbit; 100% (95% CI 90–100%) and 97% (95% CI 91–99%) for AliveCor; 97%

Reallocation of tracings labeled inconclusive, examplary for the Fitbit Sense :

1

Test diagnosis	Gold standard +	Gold standard -
+	True positive **26**	False positive **2**
-	False negative **1**	True negative **68**

1 - Contingency table with inconclusive tracings excluded.

2

Test diagnosis	Gold standard +	Gold standard -
+	True positive **26**	False positive **2**
Inconclusive	Inconclusive (positive) **8**	Inconclusive (negative) **16**
-	False negative **1**	True negative **68**

2 - Inconclusive tracings listed but not relabeled.
Arrows indicate how tracings labeled inconclusive will be reallocated.

3

Test diagnosis	Gold standard +	Gold standard -
+	True positive **26**	False positive + inconclusive (negative) **2 + 16 = 18**
	False negative + inconclusive (positive) **8 + 1 = 9**	
-		True negative **68**

3 - Contingency table with inconclusive tracings reallocated.
Bold number was used for ITD analysis.

Fig. 1. Methodological procedure of the intention-to-diagnose analysis. Reallocation of tracings labeled inconclusive, explanatory for the Fitbit Sense: (1) contingency table with inconclusive tracings excluded; (2) inconclusive tracings listed but not relabeled (arrows indicate how tracings labeled inconclusive will be reallocated); (3) contingency table with inconclusive tracings reallocated. Bold numbers used for intention-to-diagnose analysis.

(95% CI 85–99%) and 98% (95% CI 91–99%) for Samsung; and 91% (95% CI 77–97%) and 98% (95% CI 91–99%) for Withings (Table 2 and TableA.3, Fig. 2).

3.4. Intention-to-diagnose analysis

When performing an intention-to-diagnose analysis to assess the actual diagnostic accuracy in everyday use by reallocating inconclusive results as interpreted by the manufacturer algorithms as false positive and false negative (see Methods section), the results were as follows: for manufacturers' algorithms, diagnostic accuracy ranged from 71% (Withings) to 82% (Apple) for everyday performance based on the intention-to-diagnose analysis, whereas the diagnostic accuracy achieved by the DNN-algorithm ranged from 93% (Withings) to 98% (Fitbit). The DNN-based algorithm outperformed all of the five manufacturers' algorithms in each of the assessed variables (accuracy, sensitivity, specificity, and positive predictive value), as shown in Fig. 3 and Table A.2.

4. Discussion

In this large prospective diagnostic study, we set out to compare the rate of inconclusive single lead electrocardiogram tracings and the diagnostic accuracy for the detection of AF between the

Table 1
Baseline characteristics of the study population.

	Not paced & five ECGs ($n = 117$)	AA ($n = 35$)	No AA ($n = 82$)
Male sex	86 (73.5)	26 (74.3)	60 (73.2)
Age (years)	65 (56–74)	72 (63–76)	63 (55–73)
BMI (kg/m^2)	26.8 (24.1–30.1)	26.9 (24.8–30.0)	26.1 (23.9–30.2)
Procedure			
PVI	80 (68.4)	28 (80.0)	52 (63.4)
EPS/RFA SVT	15 (12.8)	3 (8.6)	12 (14.6)
EPS/RFA VT	6 (5.1)	0 (0.0)	6 (7.3)
Pacemaker	3 (2.6)	0 (0.0)	3 (3.7)
EPS/RFA CTI	7 (6.0)	1 (2.9)	6 (7.3)
ICD	4 (3.4)	1 (2.9)	3 (3.7)
Cardioversion	2 (1.7)	2 (5.7)	0 (0.0)
Known AF	97 (82.9)	34 (97.1)	63 (76.8)
Previous MI/PCI	12 (10.3)	5 (14.3)	7 (8.5)
CHA_2DS_2–VASc	2.1 ± 1.6	3.0 ± 1.8	1.7 ± 1.3
Vascular disease	13 (11.1)	6 (17.1)	7 (8.5)
Diabetes	15 (12.8)	8 (22.9)	7 (8.5)
Hypertension	60 (51.3)	21 (60.0)	39 (47.6)
Stroke	9 (7.7)	6 (17.1)	3 (3.7)
CHF	22 (18.8)	10 (28.6)	12 (14.6)
LVEF (%)	57 (50–61)	52 (46–57)	59 (54–62)
LA diameter (mm)	42 (36–45)	43 (37–46)	42 (35–45)
LAVI (%)	39 (30–46)	43 (39–47)	34 (24–44)
12-lead diagnosis			
AF	29 (24.8)	29 (82.9)	0 (0.0)
AFL/AT	6 (5.1)	6 (17.1)	0 (0.0)
SR	82 (70.1)	0 (0.0)	82 (100.0)

Data are expressed as number (%), median (interquartile range) or mean ± standard deviation. AA: atrial arrhythmia; AF: atrial fibrillation; AFL: atrial flutter; AT: atrial tachycardia; BMI: body mass index; CHF: congestive heart failure; CTI: cavotricuspid isthmus; ECG: electrocardiogram; EPS: electrophysiology studies; ICD: implantable cardiac device; LA: left atrial; LAVI: left atrial volume index; LVEF: left ventricular ejection fraction; MI: myocardial infarction; PCI: percutaneous coronary intervention; PVI: pulmonary vein isolation; RFA: radiofrequency ablation; SR: sinus rhythm; SVT: supraventricular tachycardia; VT: ventricular tachycardia.

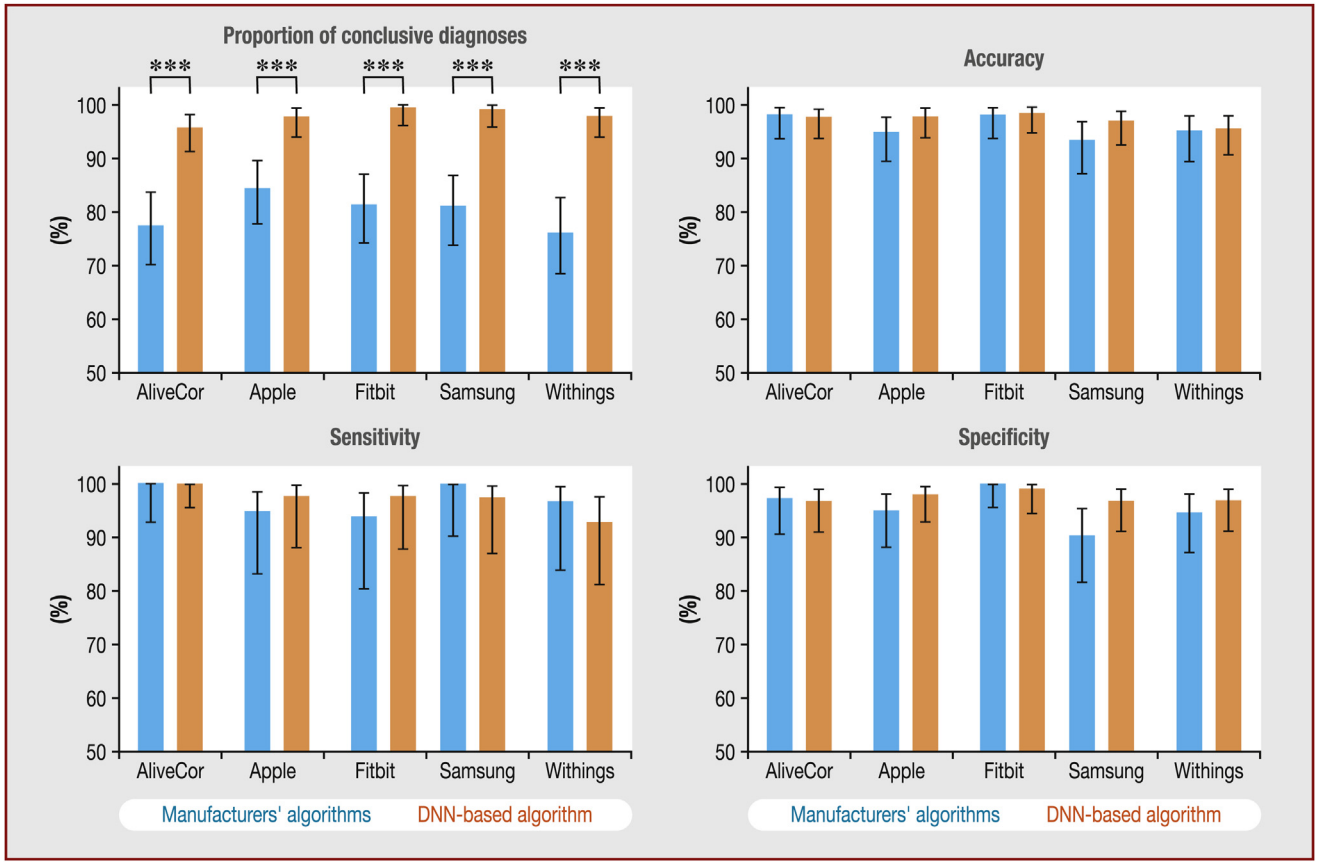


Fig. 2. Performance comparison between deep neural network (DNN)-based algorithm and five manufacturers' algorithms in detecting atrial fibrillation. Top left: comparison of proportion of conclusive tracings of each algorithm. Top right: overview of accuracy between algorithms. Bottom row: sensitivity and specificity of each device. *** $P < 0.001$.

Table 2
Performance evaluation of atrial fibrillation diagnosis by device and algorithm used for analysis.

Device	Algorithm	Number of samples	Inconclusives	CR (%)	True positives	False negatives	False positives	True negatives	Accuracy (%)	Sensitivity (%)	PPV (%)	Specificity (%)
Apple	MFR	117	19	84	32	0	2	64	98	100	94	97
Fitbit	MFR	117	22	81	26	1	0	68	99	96	100	100
AliveCor	MFR	117	27	77	30	0	1	59	99	100	97	98
Samsung	MFR	117	18	85	31	0	5	63	95	100	86	93
Withings	MFR	117	30	74	23	1	3	60	95	96	88	95
Apple	DNN	117	2	98	35	0	2	78	98	100	95	98
Fitbit	DNN	117	1	99	35	0	1	80	99	100	97	99
AliveCor	DNN	117	5	96	35	0	2	75	98	100	95	97
Samsung	DNN	117	1	99	34	1	2	79	97	97	94	98
Withings	DNN	117	3	97	31	3	2	78	96	91	94	98

CR: conclusive rate; DNN: deep neural network; MFR: manufacturer; PPV: positive predictive value.

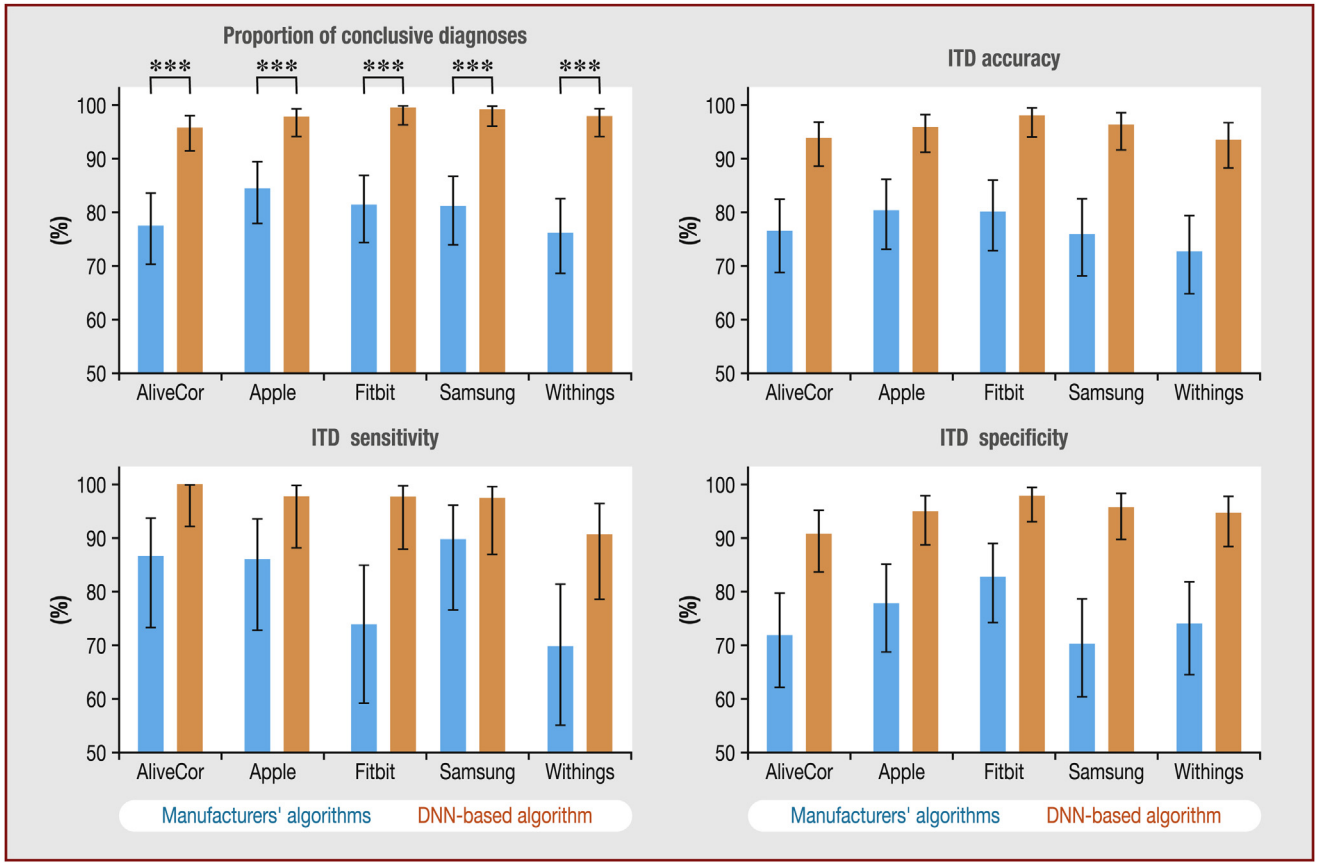


Fig. 3. Performance comparison between deep neural network (DNN)-based algorithm and five manufacturers' algorithms in detecting atrial fibrillation based on an intention-to-diagnose (ITD) analysis. Top left: comparison of proportion of conclusive tracings of each algorithm. Top right: overview of accuracy between algorithms. Bottom row: sensitivity and specificity of each device. *** $P < 0.001$.

manufacturers' algorithms and a DNN-based algorithm on five smart devices, using a simultaneously acquired physician-interpreted 12-lead electrocardiogram as the reference standard.

The main findings were as follows. First, in this real-world cohort of patients, we found a high rate of inconclusive tracings, ranging between 15% and 26%, for all five smart device manufacturers. Second, by applying a DNN-based algorithm, the rate of inconclusive tracings could be lowered to 1–4%; this comes close to the reduction in tracings labeled as inconclusive by physician review, as shown in other work [14,15]. Third, we found high diagnostic accuracy for the detection of AF for all five smart devices, and the DNN-based algorithm performed with a non-significantly different high diagnostic accuracy, while allowing a diagnosis on a larger number of recordings. Fourth, when relabeling inconclusive results as false negative or false positive to assess everyday performance via an intention-to-diagnose analysis, the DNN-based algorithm outperformed each manufacturer's algorithm in terms of diagnostic performance, as defined by sensitivity and specificity.

Our study's findings extend and corroborate previous work with smart devices, and help us to understand the most appropriate clinical use of any smart device and associated single-lead electrocardiogram function for AF management [16–18]. The rates of inconclusive tracings were higher than initially reported [19], but within the range of performances observed lately in other studies [16,20,21]; our results expose the true clinical value of the manufacturers' algorithms in a real-world cohort of patients.

Because of the primary multi-categorical classification of the interpretation, we used the intention-to-diagnose analysis to counter bias overestimating sensitivity and specificity due to results not being reported exactly in a binary test [13] (i.e. positive,

negative, inconclusive, with inconclusive interpretations being neglected, as was often done in previous work on this subject), and provide a realistic picture of the clinical potential of the smart devices.

AI-based algorithms have been used for the detection of AF in Holter and loop recorder electrocardiograms [22,23]. This type of algorithm has shown that the technology can achieve high accuracy [18,24,25]. Similarly, the proposed DNN-based algorithm could be valuable in assisting physicians to manage the large amount of data in the detection of AF using smart devices. One of the key advantages is that it dramatically reduces the number of inconclusive strips, hence the time needed for manual over-reads by a physician is likely to decrease in the future. This is important, as the time needed to analyze a single-lead recording is a concern, and represents a potential burden on limited clinical staff because of the increased amount of electrocardiogram data sent by smart device users. Nevertheless, it is important to be aware to not rely solely on this technology, but instead to consider careful review of tracings when the diagnosis remains inconclusive or when doubt exists.

To the best of our knowledge, this is the first direct comparison of five smart devices with a novel DNN-based device-agnostic algorithm within the same cohort. The results of our study suggest that the current major limitation of smart watch technology use for rhythm determination is based on the devices' automated rhythm detection algorithms, rather than the electrocardiogram acquisition technology. This poses a significant challenge for the medical community, as these algorithms are proprietary, and highly variable among companies. Unfortunately, detailed information on how each of these algorithms adjudicates AF diagnoses is not available, nor is it likely to be shared between companies, because of vested commercial interests. Thus, most manufacturers protect their raw data and their algorithms.

In this context, a device-agnostic DNN-based algorithm provides two key advantages. First, it guarantees an identical adjudication of AF diagnoses across devices; hence the importance of device-specific considerations is significantly reduced when the physician assesses the output of the algorithm. Second, it has the advantage to generalize across smart devices, which reflects that the size and diversity of the training dataset were key and appropriate to develop the algorithm. We hypothesize that the DNN can generalize to the noise and filtering differences resulting from using one of the five different smart devices. For each smart device, the fact that the algorithm generalizes to four other devices shows that device-specific characteristics of the single-lead electrocardiogram tracings are not considered to derive the diagnosis, which contributes to the robustness of the performance results. It is also interesting to note that, if single-lead electrocardiogram recordings of a new smart device should be considered, the device-agnostic algorithm is more likely to give similar performances without the need to retrain the DNN.

The device-agnostic algorithm made its diagnosis based on a PDF export, which posed different problems regarding the basis and resources available for this algorithm. Not all PDFs were vectorized when exported as a PDF from the devices. At the time of the study, the Withings PDF was exported as a rasterized PDF, which made parsing of the PDF even more circumstantial. However, the DNN-based algorithm highlights the potential for such technology. Despite not being able to rely on raw data points generated directly by the smart device sensors, the DNN-based algorithm outperformed the devices' algorithm, relying on possibly filtered data.

4.1. Study limitations

There are several limitations present that should be considered. First, inconclusive electrocardiograms were not repeated. Repeating tracings labeled as inconclusive after a few minutes could lead to a greater number of conclusive tracings, and improve the performances of the manufacturers' algorithms, at the cost of time spent by the patient and the physician.

Second, it is nevertheless to be considered that all manufacturers update the hardware and especially software regularly, and we can therefore expect algorithms to further improve. Thus, our analysis is an assessment of the current capacities of the manufacturers' algorithms.

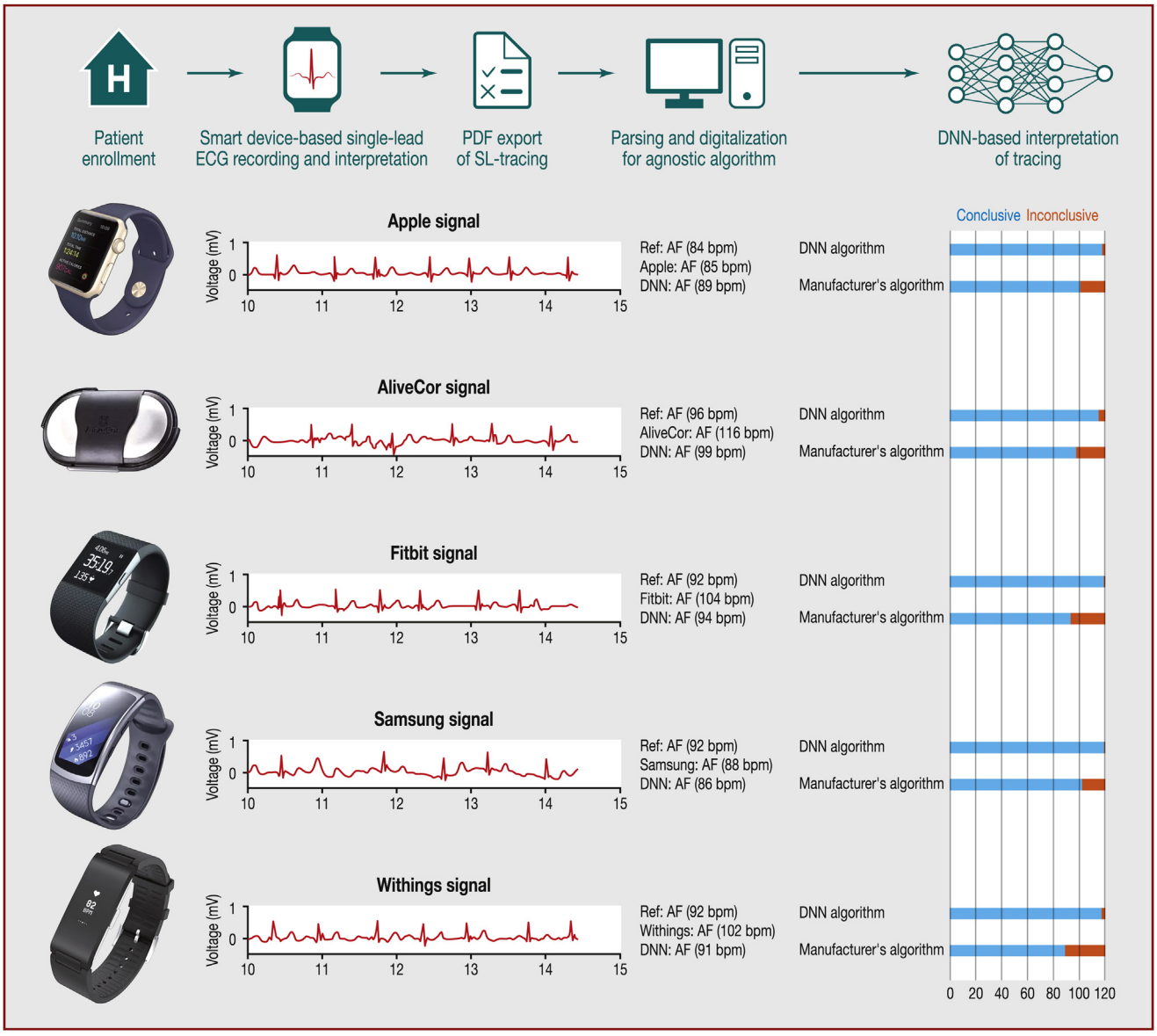
Third, it remains uncertain how tracing quality would have been without the study personnel instructing and assisting patients in recording tracings.

Fourth, the DNN-based algorithm is, like the analyzed manufacturers' algorithms, not yet trained to analyze paced recordings.

Fifth, the graphical representation of the single-lead electrocardiogram waveforms has been carefully compared with the graphical representations provided by the PDF files. This concept is based on human interpretation; however, we used a reference grid.

5. Conclusion

This device-agnostic DNN-based algorithm proved to be capable of reducing the rate of inconclusive tracings recorded with smart devices to a level where the rate of single-lead tracings without a diagnosis was very close to or within the range of tracings interpreted manually by a cardiologist [18,24,25]; this was achieved while maintaining high accuracy for the detection of AF, similar to that reached by manufacturers' algorithms on their conclusive diagnostic domain. The DNN-based algorithm therefore offers a possible solution to the still high number of inconclusive tracings generated by smart devices. This offers the potential to reduce the expected data surge generated through patient-initiated diagnosis related to single-lead recordings from smart devices (Central Illustration).



Central illustration. Top: overview of study design. Left: devices used for this study (top to bottom: Apple Watch 6; AliveCor Kardia Mobile; Fitbit Sense; Samsung Galaxy Watch 3; Withings Scanwatch). Middle: PDF export, as displayed by the corresponding device. Middle: signal pattern used by the Cardiologs deep neural network (DNN)-based algorithm for interpretation. Small box: reference (Ref.) by cardiologist, diagnosis by device algorithm, DNN interpretation. Right: rate of tracings labelled as inconclusive by manufacturers' or DNN-based algorithm. AF: atrial fibrillation; bpm: beats per minute; ECG: electrocardiogram; Inc.: inconclusive; PDF: portable document format; SL: single-lead.

Funding

None.

Acknowledgments

We thank Mirko Lische, Corinne Isenegger, Claudius Vernier and David Vögeli for their contribution to this study.

Disclosure of interest

B. L. Employee of the company Cardiologs.
C. G. Employee of the company Cardiologs.
C. H. Employee of the company Cardiologs.
P. B. Received research funding from the "University of Basel", the "Stiftung für Herzschrittmacher und Elektrophysiologie", the "Freiwillige Akademische Gesellschaft Basel" and Johnson & Johnson, all outside the submitted work, and personal fees from Abbott.
S. K. Received funding from the "Stiftung für Kardiovaskuläre Forschung".
C. S. Member of Medtronic Advisory Board Europe and Boston Scientific Advisory Board Europe; received educational grants from Biosense Webster and Biotronik, a research grant from the European Union's FP7 programme and Biosense Webster and lecture and consulting fees from Abbott, Medtronic, Biosense-Webster, Boston Scientific, Microport and Biotronik, all outside the submitted work.
M. K. Received personal fees from Bayer, Boehringer Ingelheim, Pfizer, BMS, Daiichi Sankyo, Medtronic, Biotronik, Boston Scientific, Johnson & Johnson and Roche, and grants from Bayer, Pfizer, Boston Scientific, BMS, Biotronik and Daiichi Sankyo, all outside the submitted work.

The other authors declare that they have no competing interest.

Online Supplement. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.acvd.2023.04.003.

References

[1] Verheugt FW, Granger CB. Oral anticoagulants for stroke prevention in atrial fibrillation: current status, special situations, and unmet needs. Lancet 2015;386:303–10.
[2] Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomstrom-Lundqvist C, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC). Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. Eur Heart J 2021;42:373–498.
[3] Dagher L, Shi H, Zhao Y, Marrouche NF. Wearables in cardiology: here to stay. Heart Rhythm 2020;17:889–95.
[4] Statista. Global connected wearable devices 2016-2021. Available at: https://www.statista.com/statistics/487291/global-connected-wearable-devices/[accessed date: 29th July 2020].
[5] Lakkireddy DR, Chung MK, Gopinathannair R, Patton KK, Gluckman TJ, Turagam M, et al. Guidance for cardiac electrophysiology during the COVID-19 pandemic from the Heart Rhythm Society COVID-19 Task Force; Electrophysiology Section of the American College of Cardiology; and the Electrocardiography and Arrhythmias Committee of the Council on Clinical Cardiology, American Heart Association. Heart Rhythm 2020;17:e233–41.
[6] Mittal S, Oliveros S, Li J, Barroyer T, Henry C, Gardella C. AI filter improves positive predictive value of atrial fibrillation detection by an implantable loop recorder. JACC Clin Electrophysiol 2021;7:965–75.
[7] Fiorina L, Marijon E, Maupain C, Coquard C, Larnier L, Rischard J, et al. AI-based strategy enables faster Holter ECG analysis with equivalent clinical accuracy compared to a classical strategy. EP Europace 2020;22(Suppl 1):i396 [abstract 222].
[8] Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13 [S31-S4].
[9] Bayoumy K, Gaber M, Elshafeey A, Mhaimeed O, Dineen EH, Marvel FA, et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. Nat Rev Cardiol 2021;18:581–99.
[10] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. LNCS 2015;9351:234–41.
[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. Available at: https://arxiv.org/abs/1409.1556.
[12] Smith SW, Rapin J, Li J, Fleureau Y, Fennell W, Walsh BM, et al. A deep neural network for 12-lead electrocardiogram interpretation outperforms a conventional algorithm, and its physician overread, in the diagnosis of atrial fibrillation. Int J Cardiol Heart Vasc 2019;25:100423.
[13] Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ 2012;345:e6717.
[14] Ford C, Xie CX, Low A, Rajakariar K, Koshy AN, Sajeev JK, et al. Comparison of 2 smart watch algorithms for detection of atrial fibrillation and the benefit of clinician interpretation: SMART WARS study. JACC Clin Electrophysiol 2022;8:782–91.
[15] Mannhart D, Lischer M, Knecht S, du Fay de Lavallaz J, Strebel I, Serban T, et al. Clinical validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL wearable study. JACC Clin Electrophysiol 2023;9:232–42.
[16] Bumgarner JM, Lambert CT, Hussein AA, Cantillon DJ, Baranowski B, Wolski K, et al. Smartwatch algorithm for automated detection of atrial fibrillation. J Am Coll Cardiol 2018;71:2381–8.
[17] William AD, Kanbour M, Callahan T, Bhargava M, Varma N, Rickard J, et al. Assessing the accuracy of an automated atrial fibrillation detection algorithm using smartphone technology: the iREAD Study. Heart Rhythm 2018;15:1561–5.
[18] Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. Comput Biol Med 2018;93:84–92.
[19] Apple Inc. Using Apple watch for arrhythmia detection. 2018. Available at: https://www.apple.com/healthcare/docs/site/Apple_Watch_Arrhythmia_Detection.pdf 2020.
[20] Badertscher P, Lischer M, Mannhart D, Knecht S, Isenegger C, Du Fay de Lavallaz J, et al. Clinical validation of a novel smartwatch for automated detection of atrial fibrillation. Heart Rhythm O2 2022;3:208–10.
[21] Seshadri DR, Bittel B, Browsky D, Houghtaling P, Drummond CK, Desai MY, et al. Accuracy of Apple watch for detection of atrial fibrillation. Circulation 2020;141:702–3.
[22] Olier I, Ortega-Martorell S, Pieroni M, Lip GYH. How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. Cardiovasc Res 2021;117:1700–17.
[23] Taniguchi H, Takata T, Takechi M, Furukawa A, Iwasawa J, Kawamura A, et al. Explainable artificial intelligence model for diagnosis of atrial fibrillation using holter electrocardiogram waveforms. Int Heart J 2021;62:534–9.
[24] Fiorina L, Maupain C, Gardella C, Manenti V, Salerno F, Socie P, et al. Evaluation of an ambulatory ECG analysis platform using deep neural networks in routine clinical practice. J Am Heart Assoc 2022;11:e026196.
[25] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–9.