# BMJ Open

# Observational study protocol for an arrhythmia notification feature

Sanchit Kumar ![ORCID],[1] Jason Weinstein,[1] Hannah Camille Melchinger,[1] Abigail Smith,[1] Emily Capodilupo,[2] Joseph G Akar,[3] Kanika Garg,[1] Kyle D O'Connor,[1] Mary Kate Staunton,[1] Melissa Martin,[4] Narjes Akhlaghi,[1] Oluoma Edeh,[5] Stephanie Perez,[1] Victoria Lee,[2] Kyoung A V Lee,[1] Francis P Wilson ![ORCID] [1]

¹Clinical and Translational Research Accelerator, Department of Medicine, Yale School of Medicine, New Haven, Connecticut, USA
²WHOOP, Boston, Massachusetts, USA
³Division of Cardiology, Department of Medicine, Yale School of Medicine, New Haven, Connecticut, USA
⁴Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA
⁵University of New Mexico, Albuquerque, New Mexico, USA

**Correspondence to**
Francis P Wilson;
francis.p.wilson@yale.edu

## ABSTRACT

**Introduction** Screening for atrial fibrillation (AF) in the general population may help identify individuals at risk, enabling further assessment of risk factors and institution of appropriate treatment. Algorithms deployed on wearable technologies such as smartwatches and fitness bands may be trained to screen for such arrhythmias. However, their performance needs to be assessed for safety and accuracy prior to wide-scale implementation.

**Methods and analysis** This study will assess the ability of the WHOOP strap to detect AF using its WHOOP Arrhythmia Notification Feature (WARN) algorithm in an enriched cohort with a 2:1 distribution of previously diagnosed AF (persistent and paroxysmal) and healthy controls. Recruited participants will collect data for 7 days with the WHOOP wrist-strap and BioTel ePatch (electrocardiography gold-standard). Primary outcome will be participant level sensitivity and specificity of the WARN algorithm in detecting AF in analysable windows compared with the ECG gold-standard. Similar analyses will be performed on an available epoch-level basis as well as comparison of these findings in important subgroups.

**Ethics and dissemination** The study was approved by the ethics board at the study site. Participants will be enrolled after signing an online informed consent document. Updates will be shared via clinicaltrials.gov. The data obtained from the conclusion of this study will be presented in national and international conferences with publication in clinical research journals.

**Trial registration number** NCT05809362.

## STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ A major strength of the trial is that participants are enrolled based on presence/absence of a historical atrial fibrillation diagnosis. Unlike prior studies, this reduces the selection bias of restricting gold-standard ECG patch monitoring to only those who are first identified as having an abnormality by the test device(s).

⇒ Our evaluation of device performance includes all times when reference data are available, even if adequate quality data are not available from the test device. This offers a better estimate of real-world capabilities.

⇒ Our study design also has certain limitations—using preidentified patients with an electronic health record diagnosis of atrial fibrillation may be prone to misclassification and the workflow is different from the real-world use where an alert is issued first which is followed by an ECG patch recording.

⇒ Our preliminary experience shows good performance of the device algorithm with calculated performance metrics that are important in informing regulatory decisions.

## INTRODUCTION

Atrial fibrillation (AF) is the most common sustained abnormal heart rhythm, with an incidence in the USA expected to increase from 1.2 million cases in 2010 to 2.6 million by 2030.[1] It is associated with factors such as increasing age, smoking, diabetes and hypertension, and can result in important clinical consequences such as worsening heart failure and stroke.[2 3]

The consequences of cerebrovascular events resulting from AF can be devastating and nearly a fourth of patients with acute stroke are found to have previously undiagnosed AF.[4] Due to the serious implications of this arrhythmia, population-based screening may help identify individuals at risk and allow institution of appropriate mitigation strategies based on risk factors. However, the low population prevalence and high cost of implementing broad population level screening using traditional methods has precluded widespread adoption of these strategies.[5]

Increasing use of consumer wearables has resulted in the ability to leverage technology in implementing AF screening strategies.[6] Prior studies have shown the usefulness of smart-watch based detection of abnormal heart rhythms using photoplethysmography (PPG)-based analysis of heart rhythm regularity.[7 8] These rely on proprietary algorithms for analysing data and as such need to be validated in clinical studies prior to clearance for use. The use of technology allows for novel study designs and adjudication mechanisms

**BMJ**

which have their own benefits and limitations compared with traditional study designs.

In this paper, we provide the framework for analysing such an algorithm implemented on the WHOOP fitness strap using a novel, prospective, pragmatic study design approach that aims to improve feasibility and minimise cost.

## METHODS AND ANALYSIS
### Participants
Participants in this prospective, open-label, remotely recruiting, pragmatic study will be enrolled from among patients attending Yale associated healthcare sites. The study has begun in April 2023 and is expected to be completed within a year by April 2024. Based on electronic health record (EHR) screening, potential participants will be identified and offered enrolment by the study team. These will include patients ≥22 years of age with persistent AF, paroxysmal AF or no known cardiac rhythm abnormality. Participants providing informed consent will be recruited for the study. This enrichment based on EHR diagnoses is needed to ensure adequate event rates in the enrolled cohorts. Studies of other wearable devices could leverage large, representative user bases to initially screen for events and then target enrolment to those who screen positive.[7 8] This approach may bias the ultimate assessment of device accuracy, however, as enrolment was predicated on at least one episode of device-based rhythm interpretation. Given this issue, and the smaller commercial userbase of the WHOOP strap, cohort enrichment was used as described to ensure that an adequate number of events will be available to verify the accuracy of the algorithm.

### Exclusion
Patients will be excluded from the study if they have an implanted cardiac device (such as a pacemaker, implantable cardioverter defibrillator, left ventricular assist device, etc), use medications for rhythm control, have known sensitivity/allergy to ECG patches/skin glue/fitness-strap material including polyamide, polyester or elastane, have had a myocardial infarction in the 90 days prior to screening, have a significant tremor, are pregnant/planning on being pregnant during the study or have an active skin disease that preclude collection of usable data. As the test device requires pairing with a smartphone for data collection and transfer using an English-language-only app, non-English speaking patients or those without access to a smartphone (iOS V.14 or higher/Android OS V.10.0 or higher) will be excluded from the study.

### Patient experience and follow-up
Participants will be required to setup and wear the WHOOP wrist strap and BioTel ePatch for the duration of the study as per manufacturer instructions. The participants will be asked to go about their daily activities (apart from taking baths/swimming with the ePatch).

During this period, the data from the WHOOP device will be periodically uploaded to WHOOP servers via the WHOOP mobile application. At the end of 7 days, the BioTel ePatch device will be mailed back to the study centre for processing and uploading the data to BioTel servers for analysis.

### WHOOP-Arrhythmia Notification Feature (WARN) algorithm
The WHOOP strap will be tested for its ability to detect AF. The device itself is a wearable, battery-powered wristband equipped with sensors that include accelerometers, diodes and LEDs. The device is without an external display and communications with the user are through the Bluetooth-synced app that is available in English on both Android and iOS platforms. PPG allows detection of the pulse as data that are transmitted and analysed using the WHOOP's proprietary classification algorithm for the detection of AF. When adequate data are available, a machine learning algorithm (XGBoost) analyses beat-to-beat intervals for irregularity suggestive of AF, with aggregated data analysed as overlapping 30 min 'epochs'. Detection of abnormality within a 30 min window/epoch will be used to label the window as having AF. As the study is intended to only evaluate this functionality, no alerts regarding heart rhythms will be issued to the participants during the study. The 7 day ECG recording results from BioTel ePatch will be shared with the participant on completion of the study.

### Safety
The BioTel ePatch reports will be reviewed by a trained cardiologist for verification and sharing with the participants. In case any worrisome findings requiring more urgent medical evaluation are found, the study coordinators will contact the participant via phone and email to apprise them of the same.

### Data adequacy and quality assurance
The study requires participants to concurrently wear the WHOOP wrist-strap and BioTel ePatch to have the necessary data for analysing algorithm performance. In this regard, participant's concurrent use of the two devices would be assessed and at least one instance of 20 out of 24 hours (on a rolling 24 hour window) of concurrent use of both devices during the 7 day study period will be required for the participant to be included in the analysis.

While the WARN algorithm only makes predictions when adequate data quality is available, the ePatch records ECG continuously. Therefore, to address intervals where the ePatch recording is not of adequate quality, these windows will be labelled and removed prior to analysis. The primary outcome will be the sensitivity and specificity of the WARN algorithm-based interpretations for detecting AF in the participants during 7 days of device use. The algorithm will be blind to participant cohort (AF/control) and ePatch result.

### Gold standard
This will be based on the 7 day ECG recording obtained from the BioTel ECG patch. Trained cardiologists with

electrophysiology expertise at BioTel will interpret the ECG data while being blind to the WARN algorithm interpretations and cohort assignment (AF vs control). While >30 s of AF is classified on continuous ECG monitoring as AF, the clinical significance of events <5 mins in duration seem to be uncertain.[9] The WARN algorithm is intended as an alert to initiate further investigation rather than being diagnostic of AF. Therefore, for our study, the ECG recording (BioTel) showing AF for ≥5 mins at any time will be used to define the presence of AF in a participant over the 7 days of monitoring.

### WARN algorithm-defined diagnosis
The classification algorithm uses PPG-derived beat-to-beat intervals collected via the wearable device to determine if the pattern is consistent with a diagnosis of AF within a 30 min 'epoch' or window. We define a 'positive' reading from the WARN algorithm as 'atrial fibrillation' being detected using the proprietary classification algorithm.

### True positive
Defined as individuals who have at least one 30 min window labelled as having AF by the WARN algorithm which has ≥5 min of continuous AF as determined by the BioTel recording and confirmed by tachogram adjudication.

### True negative
Defined as individuals who do not have any 30 min windows labelled by the WARN algorithm as having AF and none of these windows contains ≥5 min of continuous AF as determined by the BioTel recording.

### False positive
Defined as individuals who have 30 min windows labelled as AF by the WARN algorithm but either none of these windows contains ≥5 min of continuous AF as determined by the BioTel recording OR contain ≥5 min of continuous AF on the BioTel recording but none of these can be confirmed by tachogram adjudication.

### False negative
Defined as individuals who do not have any 30 min windows labelled by the WARN algorithm as having AF but at least one such window has ≥5 min of continuous AF as determined by the BioTel recording.

This is summarised in figure 1.

### Mechanism to adjudicate BioTel findings
All reports generated by BioTel from the chest patch ECG recording will be reviewed by a trained BioTel cardiologist with electrophysiology expertise. A board-certified cardiologist at Yale will review and adjudicate the report. If there is a disagreement about the interpretation, a third adjudicator, who is an independent, board-certified Yale cardiologist, will serve as tie breaker.

For the primary analysis, all potential true positives will be adjudicated by a Yale cardiologist to verify overlap between WARN algorithm labelled AF windows and BioTel ECG patch determined periods of AF.

First, BioTel tachograms corresponding to the time windows labelled AF by WARN algorithm will be requested. To avoid bias, 1/2 the number of tachograms from periods NOT labelled as AF by WHOOP will also be requested. The cardiologist will then review this mixture of normal and abnormal tachograms for interpretation.

If ≥5 min of continuous AF is verified on any of these tachograms, then the participant will be labelled a TRUE POSITIVE.

More tachograms will be reviewed till ≥5 continuous minutes of AF is found and if NO such tachograms are found within ANY of the windows labelled as AF by the WARN algorithm, then the participant will be considered a false positive. As the secondary analyses are for technical performance evaluation only, no additional adjudication of BioTel based interpretations will be made.

CIs around the sensitivity and specificity will be calculated using the Clopper-Pearson exact method.

### Secondary analyses
Secondary analyses will include evaluation of all time frames during BioTel recording, irrespective of adequate data and interpretations being provided by the WARN algorithm. While the Gold-standard for AF diagnosis (BioTel) and WARN algorithm defined AF window labels would be the same as primary analysis, the following will be defined differently.

### True positive
Defined as the count of individuals who have AF (on ECG patch as defined above) with AT LEAST one 30 min window being classified by WARN algorithm as AF, that overlaps with ATLEAST 5 min of AF as classified by the ECG patch.

### True negative
Defined as the count of individuals who do not have evidence of AF (on ECG patch as defined above) and have no 30 min windows labelled as AF by the WARN algorithm.

### False positive
Defined as individuals who do NOT have AF (as defined above) on ECG patch but have at least one 30 min window classified as AF by the WARN algorithm OR have AF (as defined above) on ECG patch and have 30 min windows classified as AF by the WARN algorithm BUT without any such window having ≥5 mins of continuous AF as detected by the ECG patch

### False negative
Defined as individuals who have 'true' AF as defined above but without any 30 min window labelled as AF by the WARN algorithm.

Based on these definitions, sensitivity and specificity will be calculated.

Secondary analyses will also include epoch level (30 min block) concordance between the ECG patch and WARN algorithm interpretation. On an epoch level basis, EACH
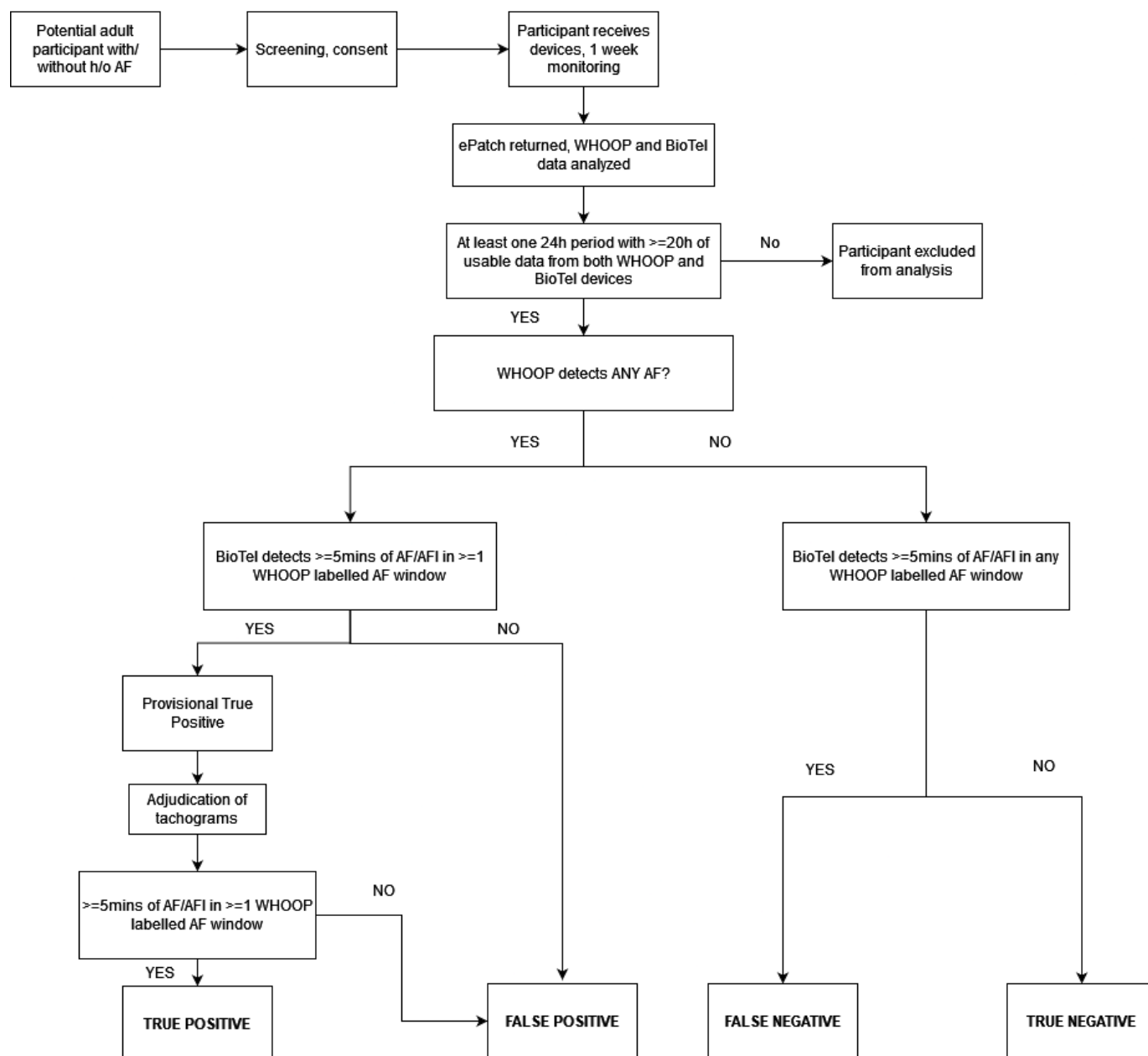
**Figure 1** Adjudication of participant level events. AF, Atrial fibrillation.

30 min window that is recorded by the WHOOP device will be compared (via the WARN algorithm) to the ECG patch reading for concordance where ANY ECG patch recording with POOR quality will be removed prior to analysis. The following definitions will be used.

### True positive
Defined as an epoch that is detected by the WARN algorithm as 'atrial fibrillation' that has ≥5 mins of continuous AF as detected by the ECG patch.

### True negative
Defined as an epoch that is NOT labelled as 'AF' by the WARN algorithm and has <5 mins of AF as determined by the ECG patch.

### False positive
Defined as an epoch that is detected by the WARN algorithm as 'AF' that has <5 mins of AF as detected by the ECG patch.

### False negative
Defined as an epoch that is NOT detected by the WARN algorithm as 'AF' but that has ≥5 mins of continuous AF as detected by the ECG patch.

We will also examine the performance of the algorithm among in the following prespecified subgroups of CHA2DS2-VASc score of 0–1 versus those with higher scores, BMI>30 and <30 kg/m$^2$, males versus females, above and below median age, Fitzpatrick I–III versus IV–VI.

## Statistical analysis
### Power
Due to the inaccuracy of ICD-10 codes and the changes in status based on the natural history of AF, the actual rate of detected AF in study participants will likely not be 2:1 (paroxysmal and persistent AF: healthy controls). Furthermore, the power of the study would need to be based on the targets for the device performance along with expected performance of the algorithm. Therefore, for determination of the specifications of this study, we leveraged the data from smaller proof-of-concept studies. For this study, we aimed to demonstrate performance characteristics similar to predicate devices, operationalised by a lower bound of the 95% CI for sensitivity>60% and specificity>90%. Based on simulations, a total of 450 participants (350 with AF and 100 with no history of AF) with ~31% actual AF (similar to proof-of-concept study) would provide 98.70% power to detect a lower bound of sensitivity>60% and specificity>90% (assuming actual device specificity of 96% and sensitivity 83%) at the 0.05 significance level. Accounting for ~11% loss to follow-up and bad quality data, we will be aiming to recruit ~500 participants for the study to have the required number of participants with adequate data for evaluation.

### Interim analysis
No interim analyses will be performed for this study.

## Ethics and dissemination
### Ethical considerations
The study will be conducted in accordance with the Declaration of Helsinki, and ICH Good Clinical Practice guidelines and the applicable regulatory requirements (21 CFR Parts 50, 54, 56 and 812). The conduct of the study has been approved by the Institutional Review Board (IRB) of the Investigational site (version 7, dated 21 March 2023). Safety assessments (adverse events) for this study will be conducted in accordance with ISO 14155–1 (2020).

### Data safety monitoring plan
The participants will be able to describe any adverse events during the study to the research team. At the end of their participation, the participants will be contacted by the study team to describe any adverse events. These will be adjudicated by a physician at the investigation site. Any serious adverse events will be reported to the IRB and the study sponsor as per policy.

### Confidentiality and access to data
Data collected during the study will be stored on 21 CFR compliant RedCap and ZS Connected Research Platform servers. Access to this data will be restricted to the study teams. No protected health information will be included as a part of any publication/analysis.

### Dissemination policy
At the conclusion of the study, the data and analysis will be disseminated via peer-reviewed publications and presentations at conferences. The study team will prepare and submit the same for publication without the involvement of the study sponsor in this decision. Professional writers will not be used.

### Data availability
Deidentified patient data will be made available on reasonable request.

### Patient and public involvement
There is no involvement of any patients in the design, conduct or dissemination of this study.

## DISCUSSION
Our study employs a novel study design that aims to improve feasibility and reduce costs associated with testing technologies intended to be deployed in consumer devices. Compared with previous study designs that leverage large user pools to push alerts and recruit based on positive alerts, this design offers certain benefits.[8 10] The inclusion of prerecruitment screening increases the number of observed events during the study period, thereby offering an improved power at a lower sample size. This is beneficial from both a time and financial perspective. Furthermore, the shorter time to availability of results can allow iterative cycles that serve to both test and train prediction algorithms to allow continuous improvements before undertaking more major studies for the purpose of market clearance. The concurrent use of the test device with the standard of care also allows real-time assessment of the device performance, with all interpretations made during the use of the devices being available for adjudication due to the one-step process.

Comparatively, previous studies used the issuing of the alert by the software to select participants for the main part of the study which consisted of using the test device with the gold standard.[7 8] This may introduce a selection bias—evaluating the positive predictive value of the notifications in this enriched cohort might overestimate performance based on characteristics of the algorithm since it was the algorithm that identified the participants in the first place. Furthermore, as the algorithms also require 'good quality' data (often nocturnal/at rest), analysis is restricted to times where such quality is available from the device being tested. While a large proportion of AF occurs at night, ~1/4th of such episodes occur during the day[11 12] and may also be precipitated by exercise, especially among athletes who maybe more likely to use wearable fitness trackers.[13] This has important implications in reporting results as selectively analysing times with good data availability may not be reflective of the capability of the test device in detection of events during suboptimal recording conditions. Furthermore, reporting high accuracy in such scenarios may falsely reassure users who may not seek timely medical advice.[14] It is therefore imperative that companies disclose this information and clarify that the devices are not intended to be used to assess whether a discussion is warranted with their physician. Reporting performance based on the presence of good data from

reference device (secondary analysis in our study design) irrespective of the same being available from the test device can also help in improving transparency regarding these limitations.

While being more pragmatic and less costly, our study design has certain limitations. First, use of electronic-health record (EHR) based diagnoses is prone to misclassification.[15] This issue is minimised by restricting the use of EHR to screening and this is followed by 7 day ECG monitoring with a sample size that accounts for this inaccuracy. Second, the study process does not reflect real-world usage where an alert is issued that leads to a monitoring phase using ECG monitoring. As the goal of the study is to inform decisions regarding the device characteristics, we believe that our design provides adequate performance metrics that are independent of cohort composition. Undue participant anxiety is also minimised because no alerts are issued as a part of the study.

## CONCLUSION

We here present a novel study design for the evaluation of an SaMD (Software as Medical Device) for detecting AF. Using this approach can allow shorter iterative cycles for refining algorithms and testing performance to allow regulatory clearance.

**X** Sanchit Kumar @sanchitkumarMD and Francis P Wilson @fperrwilson

**ORCID iDs**
Sanchit Kumar http://orcid.org/0000-0003-3994-9793
Francis P Wilson http://orcid.org/0000-0002-2633-2412

## REFERENCES

1 Colilla S, Crow A, Petkun W, et al. Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population. *Am J Cardiol* 2013;112:1142–7.
2 CDC. Atrial fibrillation. Centers for Disease Control and Prevention; 2022. Available: https://www.cdc.gov/heartdisease/atrial_fibrillation.htm [Accessed 28 Oct 2022].
3 Anter E, Jessup M, Callans DJ. Atrial fibrillation and heart failure. *Circulation* 2009;119:2516–25.
4 Sposato LA, Chaturvedi S, Hsieh CY, et al. Atrial fibrillation detected after stroke and transient ischemic attack: a novel clinical concept challenging current views. *Stroke* 2022;53:e94–103.
5 Mandrola J, Foy A. Downsides of detecting atrial fibrillation in asymptomatic patients. *Am Fam Physician* 2019;99:354–5.
6 Kumar S, Victoria-Castro AM, Melchinger H, et al. Wearables in cardiovascular disease. *J Cardiovasc Transl Res* 2023;16:557–68.
7 Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a Smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;381:1909–17.
8 Lubitz SA, Faranesh AZ, Selvaggi C, et al. Detection of atrial fibrillation in a large population using Wearable devices: the fitbit heart study. *Circulation* 2022;146:1415–24.
9 Chen LY, Chung MK, Allen LA, et al. Atrial fibrillation burden: moving beyond atrial fibrillation as a binary entity: a scientific statement from the American heart association. *Circulation* 2018;137:e623–44.
7 Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;381:1909–17.
11 Verrier RL, Josephson ME. Impact of sleep on arrhythmogenesis. *Circ Arrhythm Electrophysiol* 2009;2:450–9.
12 Mitchell ARJ, Spurrell PAR, Sulke N. Circadian variation of arrhythmia onset patterns in patients with persistent atrial fibrillation. *Am Heart J* 2003;146:902–7.
13 Newman W, Parry-Williams G, Wiles J, et al. Risk of atrial fibrillation in athletes: a systematic review and meta-analysis. *Br J Sports Med* 2021;55:1233–8.
14 Predel C, Steger F. Ethical challenges with Smartwatch-based screening for atrial fibrillation: putting users at risk for marketing purposes? *Front Cardiovasc Med* 2020;7:615927.
15 Chamberlain AM, Roger VL, Noseworthy PA, et al. Identification of incident atrial fibrillation from electronic medical records. *J Am Heart Assoc* 2022;11:e023237.