Using Apple Watch for Arrhythmia Detection

December 2020

Contents

Overview

Apple Watch customers have access to two software as medical device features to detect heart arrhythmias such as atrial fibrillation (AFib): the Irregular Rhythm Notification Feature (IRNF) and the ECG app.

On Apple Watch Series 1 or later, the IRNF analyzes pulse rate data collected by the photoplethysmograph (PPG) sensor to identify episodes of irregular heart rhythms suggestive of AFib, and it provides a notification to the user when it detects an episode. On Apple Watch Series 4, Series 5, and Series 6, the ECG app can generate an electrocardiogram (ECG) waveform similar to a Lead I electrocardiogram, then provide a classification of that waveform as sinus rhythm (SR), atrial fibrillation (AF), high or low heart rate, or inconclusive; with ECG 2.0, additional classifications of AFib with high heart rate and poor recording are available.

This paper provides a detailed understanding of the capabilities of these features, including testing and validation.

Introduction

AFib, a type of irregular heart rhythm in which the atria of the heart beat irregularly and sometimes rapidly, is a leading cause of stroke. But because AFib is often asymptomatic, many individuals with AFib may be unaware they have this condition. The combination of stroke risk, asymptomatic presentation, effective pharmacologic treatments minimizing stroke risk, and the increasing market penetration of consumer devices with the potential to detect AFib have increased interest in the early identification of AFib outside the clinical setting.

With watchOS 5.1.2 or later, Apple Watch Series 1 and later are able to use PPG signals combined with an algorithm to identify periods of irregular pulse suggestive of AFib. In addition to using this PPG-based identification algorithm, Apple Watch Series 4, Series 5, and Series 6 also have an electrical heart sensor that, when using the ECG app, enables the generation and analysis of an ECG similar to a Lead I ECG.

PPG-Based Arrhythmia Detection

Technical and Feature Description

Apple Watch has an optical heart sensor that uses green LED lights paired with light-sensitive photodiodes to detect blood volume pulses in a user's wrist using photoplethysmography. These sensors and underlying algorithms are the basis for the heart rate (HR) and heart rate variability (HRV) detection enabled on Apple Watch Series 1 and later. To determine HRV, Apple Watch captures a tachogram—a plot of the time between heartbeats—every two to four hours. Beginning with watchOS 5.1.2, users may also choose to enable an arrhythmia detection feature that uses these tachograms. To use the Irregular Rhythm Notification Feature on Apple Watch, users must first complete onboarding in the Health app on their paired iPhone to learn how to use the feature and receive education regarding AFib. For more information about the user experience, visit support.apple.com/en-hk/HT208931.

If the PPG-based arrhythmia detection is enabled, Apple Watch classifies each tachogram using a proprietary algorithm to determine if an irregular rhythm might be present. An irregular tachogram initiates a cascade of more frequent tachogram collection—as frequently as possible, subject to a minimum spacing of 15 minutes—and analysis. Apple Watch collects and analyzes tachograms only if the user remains still enough to obtain a reading. Because of this, the algorithm isn't always monitoring the user, but rather is doing so opportunistically when adequate signal is available for collection and analysis. If five out of six sequential tachograms—including the initial one—are classified as irregular within a 48-hour period, the user is notified of the potential arrhythmia. In addition to receiving the notification, the user can access more information related to these irregular tachograms in the Health app (Figure 1). If two tachograms are classified as not irregular before the threshold is reached, the cycle is reset and tachogram collection returns to the baseline rate every two hours.

Figure 1: Health App View of Irregular Rhythm Measurements
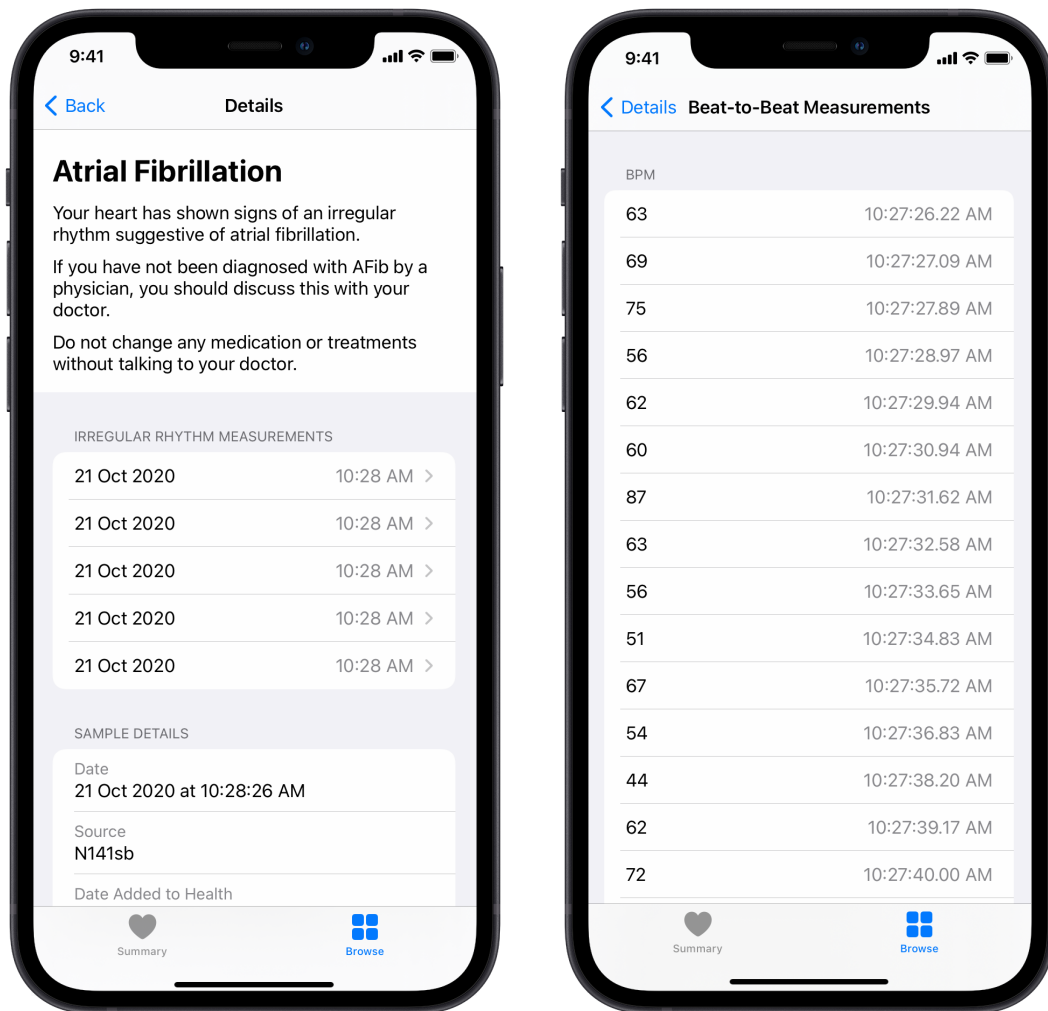
In the Health app, users can see the times when the algorithm identified an irregular tachogram that contributed to a notification (left). Tapping a date and time allows a user to visualize the beat-to-beat measurements calculated from each tachogram.

Preclinical Development

Prior to clinical testing, studies were conducted to develop the PPG-based detection algorithm and to evaluate algorithm performance across a variety of conditions and user behaviors. Among these were deep breathing, riding in a car, hand tremors and motion, reduced hand or wrist perfusion, overnight wear, rapid ventricular response in individuals with AFib, and other arrhythmias. These studies were performed in 2300 control subjects and more than 500 subjects with AFib.

Because PPG relies on light absorptivity, the arrhythmia detection algorithm was tested across a variety of skin types and tones to ensure that sensor platform adjustments for skin tone were sufficient in the context of the algorithms used to detect arrhythmias. Melanin has high absorptivity at the wavelength used by the green LED on Apple Watch, making PPG heart rate measurement potentially more difficult in darker skin tones. To account for this, the Apple Watch sensing platform adjusts LED current (and hence light output), photodiode gain (sensitivity to light), and sampling rate to ensure adequate signal amplitude across the full range of human skin tone.

For validation purposes, 1.3 million tachograms from 1124 subjects (51 percent female) with varying skin types and tones were analyzed (based on Fitzpatrick skin types and spectrophotometer measurements of skin lightness at the wrist). As the primary engineering concerns focused on signal amplitudes in individuals with dark skin, nearly 5 percent of enrolled subjects had Fitzpatrick type VI skin. Validation efforts demonstrated no significant difference in algorithm sensitivity or specificity across skin types or tones.

Clinical Validation

Apple Heart Study
The Apple Heart Study (AHS) is a prospective, single-arm pragmatic study conducted virtually to evaluate the ability of the Apple Watch–based irregular pulse notification algorithm to identify arrhythmias suggestive of AFib. In the study, if a user met the five-out-of-six threshold, the user received an iPhone and Apple Watch notification and had the option of contacting a telehealth study physician and receiving an ambulatory ECG patch, the ePatch (from BioTelemetry, Inc. in Conshohocken, Pennsylvania). Participants were instructed to wear the ePatch for up to seven days, but data collected from a participant were considered adequate with a minimum analyzable time of one hour.

The detailed results of AHS were published in November 2019 in the *New England Journal of Medicine* (Perez, Marco V., et al. "Large-scale assessment of a smartwatch to identify atrial fibrillation." *New England Journal of Medicine* 381.20 (2019): 1909–1917).

AHS Sub-Study Experiment Design
A sub-study of data collected in AHS was conducted to determine if the tachogram classification algorithm (individual or spot tachogram) and the confirmation cycle algorithm (alert-level, five out of six tachograms) have acceptable positive predictive value (PPV) compared with the ePatch monitoring in identifying irregular rhythms consistent with AFib in a subset of AHS participants. AHS investigators were aware of the sub-study, subsequent analyses, and data submission to the FDA, but they were were blinded to the sub-study results while AHS was ongoing. The institutional review board (IRB) that approved AHS determined that this sub-study was exempt from IRB oversight. All AHS participants provided informed consent, which included the use of their study data for the purposes of the sub-study.

Sub-study data were collected from AHS participants enrolled between November 30, 2017, and June 22, 2018. The subjects in this sub-study received an irregular rhythm notification from the AHS app and subsequently received and wore the ePatch for interpretation of the ambulatory ECG findings. The initial irregular tachograms leading to the first notification and potentially launching the first telehealth encounter weren't analyzed as part of this sub-study; only irregular tachograms and notifications that occurred while a user was wearing the study-provided ePatch were analyzed.

Two independent ECG adjudicators with U.S. board certification in cardiology or electrophysiology provided review and adjudication of ECG strips, classifying them as SR, AFib, another irregular rhythm, or unreadable. If the adjudicators didn't agree, a third, similarly qualified adjudicator evaluated the strip. These adjudicators were blinded to the tachogram classification. The adjudicator ECG classification and algorithm-determined tachogram classification were securely sent to the study statistician for data analysis.

Results

Of the 226 sub-study participants who received an initial arrhythmia notification and wore an ePatch for approximately one week, 41.6 percent (94 participants) had AFib detected by the ePatch. During concurrent wear of Apple Watch and an ePatch, 57 out of the 226 participants received an AFib notification—that is, they had five out of six consecutive tachograms classified as irregular. Of those 57, 78.9 percent (45 participants) showed concordant AFib on the ePatch data, and 98.2 percent (56 participants) showed AFib or other clinically relevant arrhythmias. These results demonstrate that, while in the majority of cases the notification will accurately represent the presence of AFib, in some instances a notification may indicate the presence of an arrhythmia other than AFib. No serious device adverse effects were observed.

ECG-Based Detection

Technical and Feature Description

Apple Watch Series 4, Series 5, and Series 6 incorporate a titanium electrode in the Digital Crown and an ultrathin chromium silicon carbon nitride layer applied to the sapphire crystal on the back of Apple Watch. The ECG app reads and records the electrical impulses that control the heart from the user's fingertip (with the Digital Crown) and the wrist (with the back of Apple Watch), which creates a closed circuit. To use the ECG app on Apple Watch, a user must first complete onboarding in the Health app on the user's paired iPhone to learn how to use the feature and receive education regarding AFib. To generate an ECG, a user opens the ECG app installed on Apple Watch, then applies a finger—from the hand contralateral to the wrist with Apple Watch—to the Digital Crown for 30 seconds. Lead polarity is determined by the wrist placement of Apple Watch selected in Settings.

After obtaining the ECG, a proprietary algorithm classifies the ECG tracing as SR, AFib, or inconclusive in ECG 1.0. With ECG 2.0 where available, additional classifications such as AFib with high heart rate (HR 100–150) and additional differentiation between poor recording and inconclusive readings will also be available. These rhythm classifications—average HR, user-reported symptoms, and waveform—are added to the Health app, and they're all stored and can be shared by the user as a PDF from the app on the user's paired iPhone. To learn more about the user experience, visit support.apple.com/en-hk/HT208955.

Preclinical Development

The ECG signal detection and classification algorithm were also tested in multiple studies before beginning clinical validation. The sensors and classification algorithm were tested across various ethnicities, wrist circumferences, BMI ranges, ages, non-AFib arrhythmias, degrees of band tightness, postures, and exercise states and sweating. Approximately 2500 subjects were involved in these tests; about 25 percent of them had previously been diagnosed with AFib or other irregular heart rhythms.

Increased frequency of "unreadable" ECG was the primary variation in algorithm performance. The factors leading to this variation were low signal amplitude (as a result of right-axis deviation—particularly noted in those with low BMI—or sweating noted during testing after exercise sessions) and motion artifacts as a result of user behavior. Apple Watch uses dry electrodes designed to be mechanically strong and corrosion resistant as appropriate for a wearable device. But dry electrodes, particularly those placed on extremities, are inherently more prone to introducing noise—like the kind just described—relative to the temporary gel electrodes used in clinical devices.

In addition to the factors mentioned above, the presence of certain non-AFib arrhythmias also resulted in significantly different algorithm performance compared with subjects in SR. These conditions and the results are described in Table 1 below.

Table 1. Non-AFib Arrhythmias Affecting Algorithm Performance (ECG 2.0)

Arrhythmia	Variation
Left or Right bundle branch block	7.9% of trials classified as AFib
First-degree AV block	10.2% of trials unclassified; of classified trials, 2.4% classified as AFib
Bigeminy	92.5% unclassified
Frequent PVC beats	48.8% unclassified; of those classified, 24.1% classified as AFib
Frequent PAC beats	19.8% unclassified; of those classified, 23.5% classified as AFib
Atrial tachycardia	7.7% classified as AFib
Mixed rhythm	46.5% unclassified; of those classified, 29.7% classified as AFib
High/low HR (outside 50–150 bpm)	94.8% unclassified

Clinical Validation

Experiment Design

Apple sponsored two multicenter studies to validate the ECG app's ability to (1) generate an ECG waveform similar to a Lead I ECG from a standard 12-lead ECG and (2) use a rhythm classification algorithm to classify this single lead ECG as either SR or AFib.

ECG 1.0 Clinical Validation Study

A clinical validation study was performed to validate the performance of the ECG app 1.0. The study's primary end points were sensitivity of the rhythm classification algorithm in detecting AFib and specificity in detecting SR. An external IRB approved the protocol, the informed consent form (ICF), and all other relevant materials prior to subject enrollment, and all subjects provided written consent to participate before they enrolled.

Study participants with known AFib and others with no known cardiac rhythm abnormalities were enrolled. They were asked to record three single lead ECGs using the ECG app as study staff simultaneously recorded three 12-lead ECGs using an FDA-cleared clinical device (GE Healthcare CardioSoft ECG device). The first trials were considered for adjudication and analysis. Participants were given assistance with Apple Watch placement, instructed to keep their arms still—potentially by resting their arms on a table or their legs—and allowed to practice sample acquisition before testing.

To test (1), three independent, certified cardiac technicians overlaid the generated rhythm strips from 140 randomly selected subjects (70 with AFib and 70 with SR) onto the corresponding Lead I strip from the clinical device–generated rhythm strips to visually compare morphology of six consecutive PQRST complexes. Technicians assigned each strip a pass or fail designation based on visually assessed morphological similarity. They were also asked to measure the R amplitude from isoelectric baseline to the nearest millimeter for the first two QRS complexes in both the reference and ECG app–generated rhythm strips, then assess the agreement between the two.

For (2), three blinded, independent, U.S. board-certified cardiologists reviewed each 12-lead ECG reference strip and classified the rhythm as SR, AFib, other (anything that wasn't SR or AFib within the HR parameters), or unreadable (a diagnosis couldn't be made, as the strip wasn't adequate for reading). The ECG app algorithm classified the ECG app–generated ECG as SR, AFib, unclassifiable, or unreadable. The sensitivity and specificity of the ECG app classification of SR and AFib (for classifiable ECGs) compared with cardiologist interpretation of the 12-lead ECG was calculated. One blinded, independent, U.S. board-certified cardiologist was then asked to classify the ECG app–generated rhythm strips according to the same categories.

For primary end point analyses, a one-sided exact 97.5 percent lower confidence bound was computed separately for sensitivity and specificity. If the lower bound for sensitivity exceeded 90 percent, the null hypothesis was rejected in favor of the sensitivity exceeding 90 percent. If the lower bound for specificity exceeded 92 percent, the null hypothesis was rejected in favor of the specificity exceeding 92 percent.

Results

The study enrolled 602 subjects, and 588 met eligibility criteria. Of those 588, 301 subjects with self-reported AFib were assigned to the AFib cohort, and 287 subjects without self-reported AFib were assigned to the SR cohort. These cohort assignments were used only to ensure adequate enrollment—evaluators were blinded to cohort, and the presence or absence of AFib was based solely on the ECG obtained during testing. The 14 subjects who completed the study but weren't assigned to an enrollment cohort were ineligible for study participation because of a history of paroxysmal AFib without AFib on ECG at the time of screening. All eligible subjects completed the study (Figure 2). No adverse events were reported during the study.

Figure 2: Flowchart of Subject Disposition
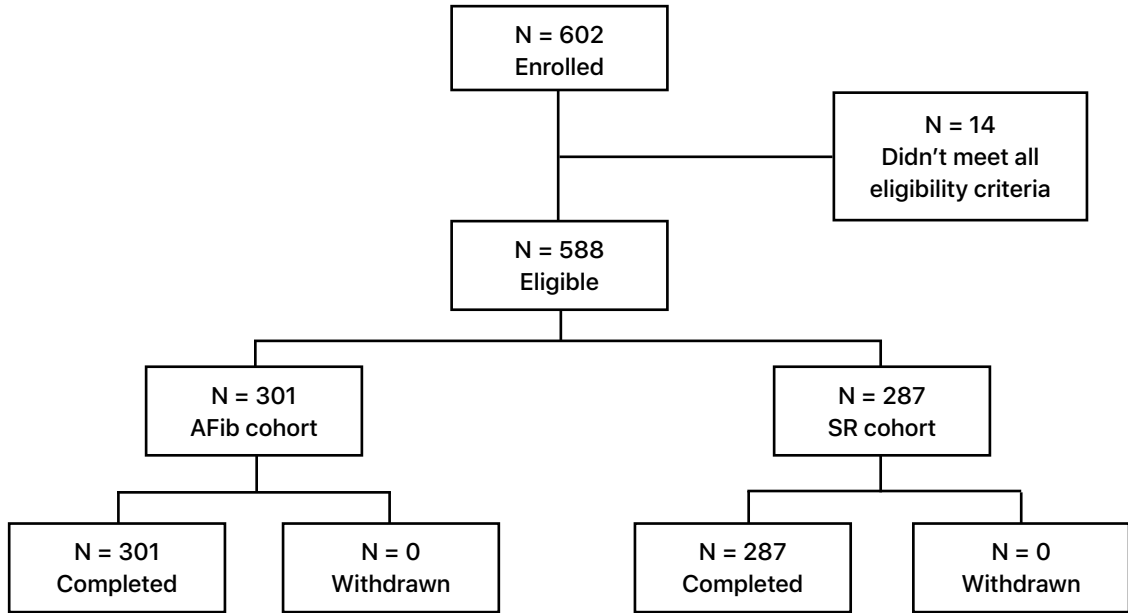
Three independent, certified cardiac technicians found visual morphological equivalence between the ECG app waveform and the reference Lead I ECG generated by the standard clinical device for 98.4 percent of analyzed strips in the AFib cohort and 100 percent in the SR cohort (Table 2). The proportion of overall subjects with a pass rating was 99.2 percent (lower 97.5 percent confidence bound = 95.7 percent). Strips were excluded if six consecutive beats (PQRST complexes) without artifact couldn't be identified in either set of strips (ECG app or reference).

Table 2. Waveform Comparison

Characteristic	AFib subjects (N = 61)	SR subjects (N = 65)	Total (N = 126)	Lower confidence bound*	P value**
Number of paired subject strips (ECG app and reference strips) with a pass rating	60	65	125		
Number of readable paired subject strips (ECG app and reference strips)	61	65	126		
Proportion of subject strips with a pass rating	60/61 (98.4%)	65/65 (100%)	125/126 (99.2%)	95,7%	< 0.0001
Number of paired subject strips excluded	8	5	13		

*Lower exact binomial one-sided 97.5% confidence bound for total
**Test of hypothesis for subject success > 0.8
Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm

For further confirmation that the waveforms generated by the ECG app and the reference device were similar, blinded cardiologist classification of the ECG app strips was compared with cardiologist classification of the reference strips (Table 3). The percent concordance of the device strip classification with the AFib and SR reference results were 100 percent and 99.1 percent, respectively. Unreadable strips weren't included in this analysis.

Table 3. Classifications Between ECG App and Reference Strips

Characteristic	Total (N = 522)
Final ECG reference result = AFib	263
Classification of ECG app strip = AFib	239/263 (90.9%)
Classification of ECG app strip = SR	0/263 (0.0%)
Classification of ECG app strip = Other	0/263 (0.0%)
Classification of ECG app strip = Unreadable	24/263 (9.1%)
% Concordance with AFib reference result*	239/239 (100.0%)
Final ECG Reference Result = SR	244
Classification of ECG app strip = AFib	0/244 (0.0%)
Classification of ECG app strip = SR	232/244 (95.1%)
Classification of ECG app strip = Other	2/244 (0.8%)
Classification of ECG app strip = Unreadable	10/244 (4.1%)
% Concordance with SR reference result*	232/234 (99.1%)
Final ECG reference result = Other	15
Classification of ECG app strip = AFib	0/15 (0.0%)
Classification of ECG app strip = SR	3/15 (20.0%)
Classification of ECG app strip = Other	12/15 (80.0%)
Classification of ECG app strip = Unreadable	0/15 (0.0%)
% Concordance with other reference result*	12/15 (80.0%)

*Unreadable strips were excluded
Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm

A total of 485 out of 602 paired ECG app and reference rhythm strips were deemed classifiable. The remaining pairs had ECG app or reference strips that were deemed unreadable or unclassifiable. Table 4 displays the breakdown among the AFib and SR cohorts.

Table 4. ECG App Algorithm Classification and Reference Strip Final Results

ECG App Algorithm Classification	Reference Strip Classification				
	SR	AFib	Other	Unreadable	Total
SR	238	4	4	1	247
AFib	1	236	2	2	241
Unclassifiable	6	7	6	0	19
Unreadable	18	30	1	0	49
Device result not reported*	32	13	1	0	46
Total	295	290	14	3	602

*Results not reported based on preestablished criteria (such as sync not detected) for all but one subject
Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm

The ECG app algorithm classification achieved a 98.3 percent sensitivity and 99.6 percent specificity (Table 5). Expanding the analysis to include the 2.4 percent (7 out of 290) and 2.0 percent (6 out of 295) of strips categorized as unclassifiable by the device in the AFib and SR reference strip classification categories, respectively, the sensitivity was 95.5 percent (95 percent CI: 92.2 percent, 97.8 percent) and specificity was 97.1 percent (95 percent CI: 94.2 percent, 98.8 percent). These results met the primary end points prespecified in the design of this study. Additionally, 12.2 percent (68 out of 556) of recordings were inconclusive—either unreadable or unclassifiable—and not classifiable as either SR or AFib. When inconclusive recordings were included in the analysis, the ECG app correctly classified SR in 90.5 percent (238 out of 263) of subjects with SR, and AFib in 85.2 percent (236 out of 277) of subjects with AFib. The clinical validation results reflect use in a controlled environment. Real-world use of the ECG app may result in a greater number of strips being deemed inconclusive and not classifiable.

Table 5. Sensitivity and Specificity Analysis—Classifiable Strips

Parameter	Value	Lower confidence bound*	P value**
Final ECG reference result = AFib (n)	240		
ECG app device result = AFib	236/240 (98.3%)		
ECG app device result = SR	4/240 (1.7%)		
Sensitivity	236/240 (98.3%)	95,8%	< 0.0001
Final ECG reference result = SR (n)	239		
ECG app device result = AFib	1/239 (0.4%)		
ECG app device result = SR	238/239 (99.6%)		
Specificity	238/239 (99.6%)	97,7%	< 0.0001

*Lower exact binomial one-sided confidence bound
**Test of hypothesis for sensitivity > 0.9 and specificity > 0.92
Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm

ECG 2.0 Clinical Validation Study

A second study was performed to support and validate investigational ECG app 2.0 algorithms (test device), which expand the classifiable HR range (50–150 bpm), and introduce new classification results (SR, SR with high HR, AFib, AFib with high HR, inconclusive, and poor recording). The objective of this study was to evaluate the test device's performance. Specificity and sensitivity were the primary end points assessed. Secondary end points included correct classification of the following categories of subjects on specific readable and classifiable ECG test strips: NSR (HR 50–150), SR on simultaneous 12-lead ECG as SR; AFib (HR 50–99), AFib on simultaneous 12-lead ECG as AFib, sinus tachycardia (HR 100–150), SR on simultaneous 12-lead ECG as high heart rate, and AFib with high heart rate (HR 100–150), AFib on simultaneous 12-lead ECG as AFib. Additionally, equivalence of the ECG app waveform to Lead I from 12-lead ECG—as measured by acceptable morphology of PQRST complexes and R wave amplitude agreement—were also assessed, similar to the waveform assessment performed in the ECG 1.0 clinical validation study.

The conduct of this study was approved by the appropriate institutional review boards of the respective investigational sites of the prospective, U.S.-based multicenter study. The site investigator obtained IRB approval, the ICF, and any subject-facing materials at each investigational site before participation in the study.

Figure 3: Flowchart of Subject Disposition
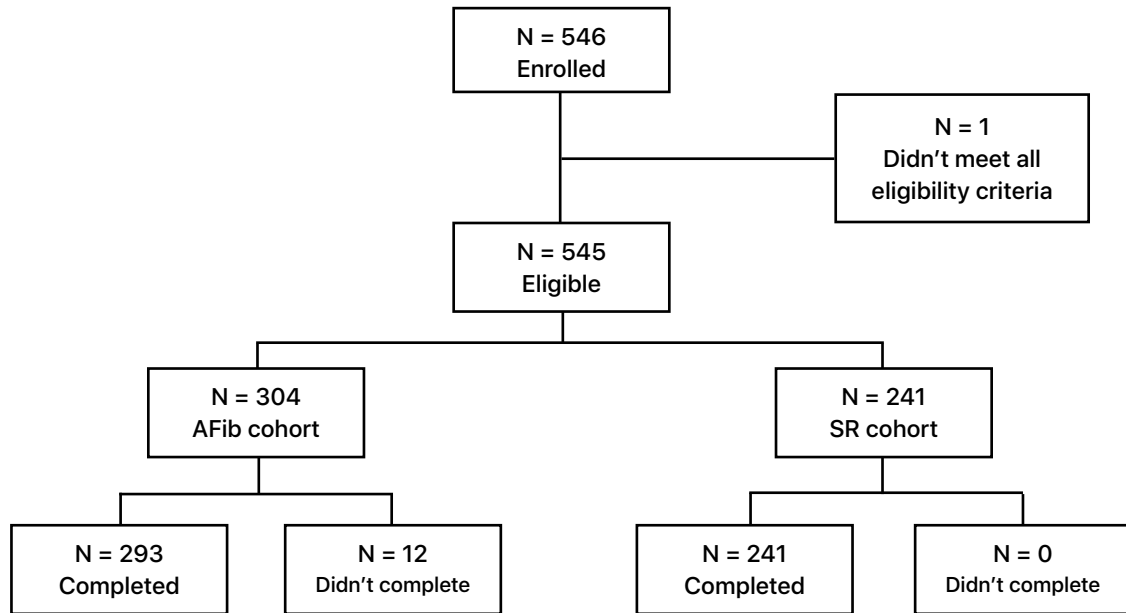
Study subjects included those with normal SR at the time of screening—with no known history of AFib—and those with known persistent, permanent, or chronic AFib who were in AFib during screening. All subjects were instructed to take a 12-lead ECG and a separate complete single lead ECG with Apple Watch simultaneously; three trials were conducted with each subject. The same process was followed for the exercise sessions; subjects who were deemed fit were asked to exercise for five minutes using a stationary bike to reach the target HR. For both rest and exercise sessions, the first trials were considered for adjudication and analysis.

Two blinded, independent U.S. board-certified cardiologist adjudicators reviewed the 12-lead ECGs for HR and rhythm diagnosis. In the event of a discrepancy, a third adjudicator performed a review. HR was calculated for each 12-lead ECG. HR was recorded, and the HR diagnostic code that corresponds to the HR that was observed on the reference ECG was selected. The following heart rhythm diagnoses were adjudicated to the 12-lead ECG data: SR, AFib, supraventricular tachycardia (SVT), another abnormal rhythm (frequent premature atrial contractions, frequent premature ventricular contractions, atrial flutter, ventricular tachycardia, ventricular fibrillation, second-degree AV block type I, second-degree AV block type II, third-degree AV block, and other), and uninterpretable. Three blinded cardiac technicians or cardiologists reviewed the waveform during the assessment and adjudication of ECG data from the paired ECG strips. The first six consecutive distinct readable PQRST complexes without artifact that matched between the subject device strip and the reference device strip for evaluation was identified by one reviewer and used by the two other reviewers. The strips were excluded if six consecutive beats couldn't be found.

For primary end point analysis, a bootstrap approach was implemented to obtain two-sided 95 percent confidence intervals for sensitivity and specificity since data were collected from the same study subjects at rest and after exercise. Subjects with at least one adjudicated result of AFib (for sensitivity) or SR (for specificity), and with a classifiable algorithm result (SR or AFib), were selected at random with replacement. The 2.5th and 97.5th percentiles of the distribution of bootstrap estimates represented the two-sided 95 percent confidence bounds. If the lower confidence bounds for both sensitivity and specificity exceeded the preestablished performance goal associated with these performance metrics (90 percent for sensitivity and 92 percent for specificity), the null hypothesis was rejected in favor of the alternative hypothesis. For waveform comparison, data from the first trial run of the randomly selected study subjects—while subjects were at rest and after exercise—were used to test the waveform assessment hypotheses using a bootstrap sampling approach. If the 2.5th percentile of the bootstrap distribution for the morphology pass rating proportion exceeded 80 percent, the null hypothesis (the morphology pass rating proportion was less than 80 percent) was rejected. The same criterion was applied to the evaluation of the R wave amplitude agreement proportion.

Results

A total of 546 subjects were enrolled in the study; 305 were enrolled in the AFib cohort and 241 were enrolled in the normal sinus rhythm (NSR) cohort. One subject from the AFib cohort was excluded for eligibility reasons. Of the 546 subjects, 534 (293 in the AFib cohort and 241 in the NSR cohort) completed the study; 12 subjects in the AFib cohort didn't complete the study.

Three separate cardiac technicians reviewed the waveform assessment analysis set for accountability purposes. Paired watch and reference strips were used for the assessment. A total of 91 subjects (100 percent) had a readable "at-rest" strip pairing, 87 subjects (100 percent) had a readable "after-exercise" strip pairing, and 93 subjects (100 percent) had at least one strip pairing from "at-rest" or "after-exercise."

The classifiable analysis set consisted of all subjects who had readable paired ECG app 2.0 and 12-lead ECG adjudicated results. A total of 512 subjects (n = 279 in the AFib cohort and n = 233 in the NSR cohort) were considered classifiable. The waveform assessment analysis set—defined as randomly selected subjects with readable paired strips from the ECG app and Lead 1 of the 12-lead ECG—comprised of 93 subjects overall (n = 48 in the AFib cohort and n = 45 in the NSR cohort).

Table 6 presents the cross-classified data of the ECG app 2.0 and the adjudicated 12-lead ECG results for combined at-rest and after-exercise data. A high level of agreement between the adjudicated 12-lead ECG results and the ECG app device output was achieved for the SR and AFib categories. For 12-lead ECG strips adjudicated as SR (n = 470), a total of 436 results were also classified as SR by the ECG app device output. For 12-lead ECG strips adjudicated as AFib (n = 521), a total of 474 results were classified as AFib by the ECG app device output. The ECG app device classified 9 results as SR, 37 results as AFib, 13 results as inconclusive, and 9 results as poor recording that were adjudicated as other (SVT or other) on the 12-lead ECG.

Table 6. Cross-Classification Table of ECG App and Adjudicated 12-Lead ECG Results (At-Rest and After-Exercise)—Classifiable Analysis Set

ECG app device output	Ground Truth Using Adjudicated 12-Lead ECG Results					
	SR (50≦HR≦150)	AFib (50≦HR≦150)	Other (SVT or other with HR 50–150; HR<50; HR>50)	Uninterpretable	Not available	Total
SR [SR (50–99); NSR sinus tach (100–150)]	436	7	9	0	2	454
AFib (50≦HR≦150)	3	474	37	0	6	520
Inconclusive [including HR<50; HR>150]	10	13	13	0	1	37
Poor recording	14	23	9	0	3	49
Not available	7	4	1	1	0	13
Total	470	521	69	1	12	1073

Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm, HR = heart rate, SVT = supraventricular tachycardia

The specificity of the ECG app 2.0 was estimated to be 99.3 percent, and the sensitivity was estimated to be 98.5 percent (Table 7). The lower confidence bounds of the specificity estimate (98.4 percent) and sensitivity estimate (97.3 percent) exceeded the preestablished performance goals associated with these performance metrics; therefore, the null hypotheses for specificity and sensitivity were rejected in favor of the alternative hypotheses, and the primary end points of specificity and sensitivity were met. At-rest specificity was estimated to be 99.1 percent, and at-rest sensitivity was estimated to be 98.5 percent. After-exercise specificity was estimated to be 99.5 percent, and after-exercise sensitivity was estimated to be 98.6 percent. An additional analysis incorporating the "inconclusive" algorithm classification category into the calculation was done for the combined at-rest and after-exercise data. The resulting specificity was estimated to be 97.1 percent, while the sensitivity was estimated to be 96.0 percent.

A high level of agreement was observed between the adjudicated 12-lead ECG HR rhythm results and the ECG app device output. The level of misclassification was low when compared with the 12-lead ECG reference strip results. Table 8 provides the summarized results that confirm that the secondary end points for correct classification of patients with NSR (HR 50–99, SR on simultaneous 12-lead ECG) as SR, AFib (HR 50–99, AFib on simultaneous 12-lead ECG) as AFib, sinus tachycardia (HR 100–150, SR on simultaneous 12-lead ECG) as high heart rate, and AFib with high heart rate (HR 100–150, AFib on simultaneous 12-lead ECG) as AFib (high HR) on a readable and classifiable ECG app strip. The percent of SR (HR 50–99) correctly classified was 98.3 percent; the lower confidence bound of 96.6 percent exceeded the preestablished performance goal of 90 percent for this end point. The percentage of AFib (HR 50–99) that was correctly classified was 98.9 percent; the lower confidence bound of 97.5 percent exceeded the preestablished performance goal of 81 percent. The percentage of SR (HR 100–150) that was correctly classified was 90.7 percent; the lower confidence bound of 86.7 percent exceeded the preestablished performance goal of 81 percent. The percentage of AFib (HR 100–150) that was correctly classified was 83.0 percent; the lower confidence bound of 77.8 percent exceeded the preestablished performance goal of 75.5 percent.

Table 7. Sensitivity and Specificity Results (Primary End Point Results)—Classifiable Analysis Set

Parameter	Value	Two-sided 95% bootstrap confidence interval*
Combined (at rest and after exercise)		
Reference result = AFib (50–150 HR)	481	
ECG app = AFib (50–150 HR)	474/481 (98.5%)	
ECG app = SR (50–150 HR)	7/481 (1.5%)	
Sensitivity	474/481 (98.5%)	(97.3%, 99.6%)
Reference result = SR (50–150 HR)	439	
ECG app = AFib (50–150 HR)	3/439 (0.7%)	
ECG app = SR (50–150 HR)	436/439 (99.3%)	
Specificity	436/439 (99.3%)	(98.4%, 100.0%)

*Results obtained from bootstrap resampling
Abbreviations: AFib = atrial fibrillation, SR = sinus rhythm, HR = heart rate

Table 8. Percent Correct of Classifiable Rhythms of AFib and SR

Parameter	Value	Bootstrap confidence interval**

Table 8. Percent Correct of Classifiable Rhythms of AFib and SR

Parameter	Value	Bootstrap confidence interval**
Combined (at rest and after exercise)*		
Percent correct for SR (HR 50–99) reference	230/234 (98.3%)	(96.6%, 99.6%)
Percent correct for AFib (HR 50–99) reference	272/275 (98.9%)	(97.5%, 100%)
Percent correct for SR (HR 100–150) reference	186/205 (90.7%)	(86.7%, 94.6%)
Percent correct for AFib (HR 100–150) reference	171/206 (83.0%)	(77.8%, 88.0%)
*Percent correct calculation includes ECG app 2.0 classifiable rhythms of AFib and SR **Confidence interval obtained from bootstrap resampling Abbreviations: AFib = atrial fibrillation, HR = heart rate, SR = sinus rhythm		

Additionally, waveform comparison between the ECG app algorithm 2.0 to Lead I from a 12-lead ECG revealed that the total pass rating percentage for combined at-rest and after-exercise morphology results was 100 percent, with a total of five excluded results. The lower confidence interval bound was 97.9 percent, which exceeded the preestablished morphology pass rating proportion of 80 percent. The total pass rating percentage for combined at-rest and after-exercise R wave assessment results was 97.2 percent; the lower confidence interval bound was 93.6 percent, which exceeded the preestablished R wave amplitude agreement proportion of 80 percent. At-rest and after-exercise groups showed similar R wave.

Conclusions

Apple Watch customers now have access to two optional features that enable the detection of irregular heart rhythms: the Irregular Rhythm Notification Feature (available on Apple Watch Series 1 and later) and the ECG app (available on Apple Watch Series 4, Series 5, and Series 6).

The Irregular Rhythm Notification Feature is a software that classifies opportunistically collected tachograms in the background, notifying consumers who activate the feature to the presence of an irregular heart rhythm. The Apple Heart Study demonstrated that of the participants who received a notification during concurrent wear of Apple Watch and an ECG patch, 78.9 percent showed concordant AFib on the ECG patch and 98.2 percent showed AFib and other clinically relevant arrhythmias. These results indicate that in the majority of cases, the notification will accurately represent the presence of AFib, but in some instances, a notification might indicate the presence of an arrhythmia other than AFib.

In Apple Watch Series 4, Series 5, and Series 6, an electrical heart sensor is available that—along with the ECG app and algorithm—generates an ECG waveform similar to a single lead (Lead I) ECG and is able to classify the waveform to look for the presence of AFib. The proprietary algorithm designed to classify these ECGs demonstrated sensitivity of more than 98 percent and specificity of more than 99 percent in identifying AFib, compared with ECGs recorded with a reference device and interpreted by independent clinical experts.

Consumers who want to use these two features must complete a user experience that provides education about the interpretation and non-diagnostic nature of these findings and the limitations of the algorithms. Both features received clearance by the FDA by De Novo or 510K classification.