COMPREHENSIVE SURVEYS ON VALIDATING WEARABLE ECG DEVICES AND RELATED FDA'S SAMD REGULATING DEVELOPMENT

Xilin Wang*

Abstract: Wearable technology is becoming more and more popular because of its convenience and accessibility, especially in connected health. Machine Learning algorithms are proven to be a practical approach in improving the accuracy of wearable technology. Recently, the FDA has established several goals of regulating AI/ML-based Software as a Medical Device (SaMD) to validate products before applying in a clinical setting thoroughly. This review comprehensively analyzes studies validating different wearable devices and algorithms conforming to the FDA standards and discusses the potential consequences of wearable devices' usage. We find an acceptable accuracy for the devices and the need for further investigation into this technology.

Keywords: SaMD, Medical AI, administrative law, wearable technology, ECG

* Student, Brown University.

Table of Contents

Introduction 96

I. Methods 98

II. Results 99

- **A. Participants and Cohorts** 101
- **B. Comparison among Accuracy Test Results** 102
- **C. Apple Watch evaluation** 104
- **D. Other devices and evaluation** 107
- **E. Consequences Brought by Wearable Technology** 108

III. Discussion 110

Conclusion 112

Introduction

Connected Health is defined as the model for healthcare management that offer a remote and more personalized service with the utilization of technology.[1] The goal of connected health is to provide opportunities for patients to participate more in their healthcare, and the emergence of low-cost costumer technologies enables this aim.[2] One of such technologies is wearable consumer electronic devices: with technological advancements in battery and computing, these devices have been able to keep track of medical data in a non-clinical setting.[3] In April 2015, Apple Inc. releases the Apple Watch Series which could measure a person's health data including fitness tracking, heart rate (HR) detection, and energy expenditure. In particular, Apple Watch Series 4, released in September 2018, has hit the market with its built-in software and hardware to perform a single-lead electrocardiogram (ECG) and detect atrial fibrillation (AF).

AF is a common type of cardiac arrhythmia from which more than five million people in the US suffer.[4] It is proven to increase the risk of stroke, heart failure, and mortality.[5] Currently, the most common technique for carrying out heart analysis is a standard 12-leads ECG, which records heart activity through putting electrodes on the body surface.[6] However, such technique requires patients and physicians to be present in the same place along with the 12-leads ECG device, which is ineffective and bothersome for the detection of AF in consideration of AF's asymptomatic nature.[7] While simpler ECG devices seem to be a solution, negative results could occur if the detection device is unable to perform long-term recording of ECG.[8] It is suggested that around 700,000 people in the US may have potential AF that is left undiagnosed, bringing up the need for the development of devices with portability, accuracy, and auto-triggered testing functionality at the same time.[9]

Several low-cost arrhythmia detection devices, including the previously mentioned Apple Watch Series, have been developed with ECG functionality.[10] Many of the devices are capable of instantaneous diagnosis and transmittion of physiological

[1] Paul Walsh, Support Vector Machine Learning for ECG Classification 10.

[2] Id.

[3] Nabeel Saghir et al., A comparison of manual electrocardiographic interval and waveform analysis in lead 1 of 12-lead ECG and Apple Watch ECG: A validation study, 1 CARDIOVASCULAR DIGITAL HEALTH JOURNAL 30–36 (2020).

[4] Dhruv R. Seshadri et al., Accuracy of the Apple Watch 4 to Measure Heart Rate in Patients With Atrial Fibrillation, 8 IEEE J. TRANSL. ENG. HEALTH MED. 1–4 (2020).

[5] Zachi I Attia et al., An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction, 394 THE LANCET 861–867 (2019).

[6] Vincenzo Randazzo, Jacopo Ferretti & Eros Pasero, ECG WATCH: a real time wireless wearable ECG, in 2019 IEEE INTERNATIONAL SYMPOSIUM ON MEDICAL MEASUREMENTS AND APPLICATIONS (MEMEA) 1–6 (2019), https://ieeexplore.ieee.org/document/8802210/ (last visited Feb 19, 2022).

[7] Id.

[8] Kenichi Hashimoto, Naomi Harada & Yuji Kasamaki, Can a Patch Electrocardiographic Device Be a Leading Actor for Detecting Atrial Fibrillation? — Diversifying Electrocardiographic Monitoring Devices —, 86 CIRC J 189–191 (2022).

[9] Seshadri et al., supra note 4.

[10] Randazzo, Ferretti, and Pasero, supra note 6.

data for any further clinical review.[11] The driving force of consumer interest in these devices culminated in the US Food and Drug Administration (FDA) clearing several portable healthcare technologies.[12] They could potentially help carry out population-level screening of AF combining with subsequent reviews from cardiologists to prevent severe cardiovascular diseases.[13] Nonetheless, the accuracy of diagnosis brought by these arrhythmia detection devices must be validated rigorously before their usage in a population-level setting.[14]

Most of the devices depend on the algorithms behind them to make diagnosis; therefore, the mechanism for these algorithms largely influence the performance of detection. Machine Learning (ML) has been more and more dominant in the design of algorithm: it is becoming as an effective approach to integrate multiple factors together to promote diagnostic accuracy.[15] For example, on Apple Watch Series 4 and later, Apple uses convolutional neural networks (CNNs) as the ML algorithm to classify data obtained from the watch's sensors.[16] There are other ML techniques that have different designs and, thus, are expected to show different levels of accuracy in the analysis of symptoms and detection of cardiovascular diseases. Some individual studies have evaluated the accuracy of arrhythmia diagnosis for a certain wearable monitoring device with/without assessments of its algorithm, and carried out experiments to compare it to a standard 12-lead ECG device currently used for clinical observation, yet few of them perform comparative analysis among the devices/algorithms to analyze the characteristics for each of them, which could provide consumers and clinicians direction for choosing the most effective and appropriate one.

FDA has been paying close attention to the regulation of the medical devices using AI/ML-based software, and published an action plan in 2021 discussing actions needed for an effective validation of SaMD.[17] These actions include: 1. A new framework for the AI/ML-based SaMD 2. Standards for Good Machine Learning Practice (GMLP) 3. Transparency of SaMD to the public 4. Effective validation and improvement for SaMD 5. Putting the application of SaMD in the real-world scenario.[18]

Given the rapidly grown wearable technology and environment of connected health, many devices are left without evaluation conforming to the FDA standards before their actual use, and consequences of this revolutionary healthcare system are not discussed thoroughly. This study aims to analyze the performance of currently available arrhythmia detection devices and algorithms, with a discussion of the consequences of using these devices, for a better understanding of the capability and potential usage of wearable technology in aiding clinical decisions.

[11] Kevin Rajakariar et al., Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation, 106 HEART 665–670 (2020).
[12] Id.
[13] Id.
[14] Id.
[15] Martin P. Than et al., Machine learning to predict the likelihood of acute myocardial infarction., 140 CIRCULATION 899–909 (2019).
[16] Walsh, supra note 1.
[17] FDA, Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (2021).
[18] Id.

I. METHODS

All data we use are studies that are relevant to either wearable ECG devices or algorithms that are tested in terms of their performance on diagnosis of cardiovascular diseases. This study compares the results drawn from the sources, and then provides a comprehensive analysis on these devices and algorithms, including their ability in detection as presented in the relevant studies, factors that influence the process of presenting the results, such as participants, criteria, and focuses of these studies, and the limitations brought up in this comparison process that might lead to future developments of the subject.

The following inclusion and exclusion criteria are applied to a found study to ensure its level of relevancy.

Inclusion criteria:

1. The study must either evaluate a certain electronic arrhythmia detection device or an algorithm that could be applied to arrhythmia detection.

2. Overall analysis on the current situation of the usage of electronic devices in healthcare will also be included.

3. The study must have some quantitative evaluation of information regarding the detection of a cardiovascular disease. For example, it might assess an ECG device's sensitivity and specificity of AF detection.

4. The study is presented in English.

Exclusion criteria:

1. The study only discusses devices that can perform ECG or arrythmia detection but are not easily appliable under the current environment of connected health, such as a 12-lead ECG device.

2. The study only presents an algorithm but does not evaluate its effectiveness on arrythmia detection.

This study is going to start by discussing the main topics of the sources and classifying them into different types of studies. This process helps articulate the nature of all gathered information and provides an open-up for the subsequent analysis. Then, a brief summarization of the targeted diseases and the devices and/or algorithms in the studies will be presented.

There are several indicators that could be used in the comparison of the studies. Many studies present some or all of the following descriptive medical indexes: sensitivity, specificity, positive predictive value, and negative predictive value. Therefore, even with different participants and cohorts, these studies are comparable due to the shared mathematics, which provide a direct insight into the effectiveness of the presented devices. In particular, F1 score is used in some of the studies to obtain a more balanced statistical measurement of a device's performance. This is calculated by the harmonic mean of the precision (positive predictive value) and recall (sensitivity).

Another commonly used indicator is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, which illustrates the diagnostic ability of a binary classification system, in this case the presence of the target cardiovascular disease.

Because of the relatability of the studies with common metrics such as sensitivity, specificity, and predictive value, quantitative data analysis will focus mostly on these measurements to reasonably compare results as much as possible. However, due to the different experimental conditions, influential factors in the studies will be considered to elucidate unexplainable differences. If certain metrics cannot be analyzed without its context, this study will refer to them without emphasizing them as decisive variables for the evaluation of the devices or algorithms.

In addition to the quantitative analysis, qualitative assessment, including comparison of the conclusion sections for the sources, will also be performed, and relationships such as agreement or contradiction will be noted. Any differences will be accounted for with potential causes such as different demographic features of participants or standards in measurement. Due to the outstanding numbers of studies regarding the accuracy of Apple Watch Series, they will be analyzed and discussed among themselves in detail. These studies will make Apple Watch Series a current model for the arrhythmia detection device, and any results drown from its analysis could potentially lead to advancements in such technology. Finally, as analyzed in some studies, the use of these devices might be influential to people's decision and mental health, such as their engagement with healthcare systems, or the existence of health anxiety, so the consequences of healthcare devices will be discussed at the end.

II. Results

With the inclusion and exclusion criteria, a total number of 20 studies are eventually included in this review. For convenience, these studies will be labeled [1] to [20], each of them having the same number as that in the reference list. Several factors lead to this relatively small number of literatures. First, in the rapidly expanding market of healthcare technology, there are more than 100,000 mobile health-related apps and >= 400 wearable activity monitors.[19] However, research or clinical validation is not performed on many of them before their practices directly to the consumers, leading to the limited number of available research.[20] On the other hand, there is an unbalanced number of research toward the currently most popular Apple Watch devices, while other devices receive much less attention. Lastly, most of the studies provide insights into further developments of the healthcare technology, but they also point out the need for more rigorous investigations and evaluations, so the archive is yet to be complete.

Table 1 and 2 show the classification of the studies in terms of their objectives and target disease, respectively. Among all devices, Apple Watch receives most evaluation, with 9 studies validating its accuracy and another 3 studies mentioning its usage. Other kinds of wearable or portable devices include the AliveCor, the ECG-WATCH and other patch-type devices. 4 studies assess 4 different algorithms that could

[19] Giuseppe Boriani et al., Consumer-led screening for atrial fibrillation using consumer-facing wearables, devices and apps: A survey of healthcare professionals by AF-SCREEN international collaboration, 82 EUROPEAN JOURNAL OF INTERNAL MEDICINE 97–104 (2020).
[20] Id.

help improve the detection functionality for healthcare devices. 17 studies discuss the diagnostic capability of the device. 11 studies choose the detection of atrial fibrillation for the performance assessment, with 2 focusing on myocardial infarction, 1 on heart rate variability, and 3 on ECG generation and classification.

Table 1 *Classification of Studies' Objective*

TYPE	NUMBER OF STUDIES (%)
DEVICE (55%)	
APPLE WATCH	9 (45%)
ECG-WATCH	1 (5%)
ALIVECOR	1 (5%)
ALGORITHM (20%)	
MYOCARDIAL ISCHEMIC INJURY INDEX (MI^3)	1 (5%)
VECTOR MACHINE LEARNING	1 (5%)
GLASGOW ALGORITHM	1 (5%)
CONVOLUTIONAL NEURAL NETWORK	1 (5%)
OTHERS (25%)	
DEVICE COMPARISON	1 (5%)
TECHNOLOGY DISCUSSION	3 (15%)
GENERAL ASSESSMENT FOR ARRHYTHMIA DEVICE	1 (5%)

Table 2 *Classification of Studies' Subject of Evaluation*

Evaluation Subject	Number of Studies (%)
Atrial Fibrillation	11 (55%)
Myocardial Infarction	2 (10%)
ECG Generation & Classification	3 (15%)
Heart Rate Variability	1 (5%)

Others	3 (15%)

A. Participants and Cohorts

Each of the studies feature their own study design, and thus incorporate different demographics as participants. As summarized in Figure 1, among all 20 studies, 5 do not include participants, either because the study is an overview of devices/algorithms or the current situation (Study [2], [8], [10], and [20]), or because the study uses ECG from existing databases (Study [7]). Other 15 studies have different recruit standards regarding the purpose of the study, yet normally only participants older than 18 could be recruited. In most cases, the summarization of participants for the studies will incorporate a section indicating the age and sex distribution, with some studies listing participants' history of diseases.

The number of participants largely varies among the 15 studies. Figure 1 maps each range of participant-number to the corresponding studies. 80% of the studies have less than 1000 participants, while Study [11] and Study [14] have 180922 and 419297 participants, respectively.[21] Most of the studies that have age statistics report a mean age of around 60, whereas Study [5] and Study [13] have participants with a mean age of 31 and 26.4, respectively.[22] There are no significant patterns for sex distribution. Note that due to the discrepancy in studies' objectives, the cohorts of different studies consist of people with different backgrounds, either healthy subjects or patients of cardiovascular diseases. In particular, Study [16] explores the impact of AF screening using devices through a questionnaire, the respondents of which are healthcare professionals.[23]

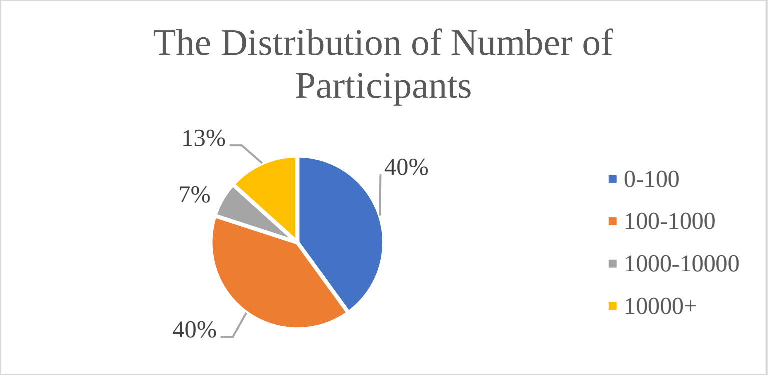


Figure 1 *The Distribution of Number of Participants*

[21] Attia et al., supra note 5; Marco V. Perez et al., Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation, 381 N ENGL J MED 1909–1917 (2019).

[22] Saghir et al., supra note 3; Ahmad Turki et al., Estimation of Heart Rate Variability Measures Using Apple Watch and Evaluating Their Accuracy: Estimation of Heart Rate Variability Measures Using Apple Watch, in THE 14TH PERVASIVE TECHNOLOGIES RELATED TO ASSISTIVE ENVIRONMENTS CONFERENCE 565–574 (2021), https://dl.acm.org/doi/10.1145/3453892.3462647 (last visited Feb 19, 2022).

[23] Boriani et al., supra note 19.

B. Comparison among Accuracy Test Results

In terms of the medical indexes used to evaluate accuracy for a test – sensitivity, specificity, positive predictive value, and negative predictive value – 7 studies have directly mentioned them as indicators for the measurement (Figure 2). Among these studies, three are for the validation of Apple Watch and its related health applications, one is for AliveCor KardiaBand (KB), and the three left are assessments for algorithm's performance.

The three Apple Watch validation studies focus on two functionalities for Apple Watch: the Irregular Rhythm Notification Feature and the single-lead ECG generation (Apple Watch Series 4 or later). For the irregular rhythm notification functionality, Study [9] and Study [14] validate its accuracy and obtain a positive predictive value of 78.9% and 84%, respectively.[24] Note that because of the new generations of Apple Watch Series 4 and later, the older versions that rely on photoplethysmography (PPG) senser for notification have received less attention. Study [9] and Study [15] present the performance of the single-lead ECG generated by Apple Watch Series 4 or later. In particular, Study [9] evaluates the effectiveness of both ECG app 1.0 and ECG app 2.0, the latter providing additional classifications such as AFib with high heart rate and differentiating between poor recordings and inconclusive recordings.[25] For ECG 1.0, the sensitivity for AFib detection reaches 98.3% (236/240), while the specificity for sinus rhythm (SR) confirmation is 99.6% (238/239), if only the two results of AFib and SR are considered.[26] With the inclusion of all inconclusive recordings – either unreadable or unclassifiable – the Apple ECG app 1.0 correctly classifies 85.2% (236/277) as AFib and 90.5% (238/263) as SR.[27] As for ECG 2.0, the first set of sensitivity and specificity is similar to that for ECG 1.0, reaching a number of 98.5% (474/481) and 99.3% (436/439), respectively.[28] The second set, in consideration of inconclusive recordings, increases with a sensitivity of 96.0% (474/494) and a specificity of 97.1% (436/449).[29] Study [15], which reports on AW 4's accuracy for AF detection, obtains a much smaller sensitivity of 41% and a 100% specificity.[30] Note that it reports a 96% sensitivity and a 100% specificity when a rhythm assessment of the AW4 generated ECG is carried out, instead of the notification provided by AW algorithm.[31]

Among the other four studies with a direct measurement of the medical indexes, Study [3] evaluates the performance of AliveCor KB, yielding an overall sensitivity of 94.4%, which is improved to 95.4% when viewing those that are appropriately diagnosed as unclassified due to sinus tachycardia as correct diagnosis.[32] Its specificity

[24] Apple, Using Apple Watch for Arrhythmia Detection (2020), https://www.apple.com/healthcare/docs/site/Apple_Watch_Arrhythmia_Detection.pdf; Perez et al., supra note 21.

[25] Apple, supra note 24.

[26] Id.

[27] Id.

[28] Id.

[29] Id.

[30] Dhruv R. Seshadri et al., Accuracy of Apple Watch for Detection of Atrial Fibrillation, 141 CIRCULATION 702–703 (2020).

[31] Id.

[32] Rajakariar et al., supra note 11.

is 81.9% with unclassified readings seen as false and is increased to 90.7% when all unclassified readings are excluded.[33] The positive and negative predictive values for the KB are 54.8% and 98.4%, respectively, with the former increasing to 72.3% once unclassified diagnoses are excluded again.[34]

The other three studies present an assessment for the proposed algorithm. Study [1] measures statistics regarding the performance of a machine learning algorithm called myocardial ischemic injury index, MI^3, in the detection of type 1 myocardial infarction.[35] The algorithm provides an index with threshold values of low-risk, intermediate-risk, and high-risk for the patients to make clinical decisions. Within the low-risk range, the algorithm has a negative predictive value of 99.7% and a sensitivity of 97.8%; above the high-risk range, it reaches a positive predictive value of 71.8% and a specificity of 96.7%.[36] The MI^3 algorithm reaches an overall AUC of 0.963, indicating a well discrimination between those with and without type 1 myocardial infarction.[37] Note that this study also compares the target algorithm with other diagnostic strategies to gain an understanding for MI^3's comparative accuracy.[38] With the given threshold, MI^3 is better at identifying low- and high-risk patients, the primary reason being its flexibility of the testing condition and simplicity of stratification of risks using a single index.[39] Study [7] evaluates the use of Support Vector Machine in ECG readings and classification.[40] This study classifies heartbeat types into five labels, and the algorithm's F1 score is the major indicator for its performance.[41] The algorithm reaches a weighted F1 score of 0.97, calculated by the average of metrics for each label, in consideration of the number of samples as weights; when the data is unweighted, F1 score is 0.82 due to the unbalance of dataset towards many normal heartbeats.[42] Study [11] focuses on convolutional neural network, the algorithm currently applied by Apple Watch. Two analyses of its performance are determined: the first analysis tests the model on the first sinus rhythm ECG for each patient, while the second includes multiple ECG data for the same patients, thus indicating whether additional information could yield better results.[43] The first analysis obtains results as follows: AUC 0.87, sensitivity 79.0%, specificity 79.5%, F1 score 39.2%, and an overall accuracy of 79.4%; in the second analysis, all the statistics improve: AUC 0.90, sensitivity 82.3%, specificity 83.4%, F1 score 45.4%, and an overall accuracy of 83.3%.[44]

Table 3 outlines the results for the seven studies in this section. Since different values could be yielded because of different standards of data processing, the lower value will be considered in the graph.

[33] Id.
[34] Id.
[35] Than et al., supra note 15.
[36] Id.
[37] Id.
[38] Id.
[39] Id.
[40] Walsh, supra note 1.
[41] Id.
[42] Id.
[43] Attia et al., supra note 5.
[44] Id.

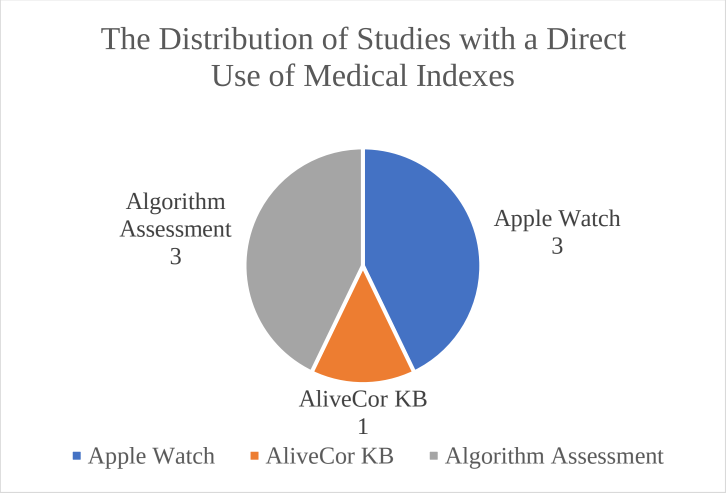


Figure 2 *The Distribution of Studies with a Direct Use of Medical Indexes*

Table 3 *Statistics for Accuracy Test Results*

	Device/ Algorithm	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**	**F1 Score (%)**	**AUC**
Study [1]	MI^3	97.8	96.7	71.8	99.7	N/A	N/A
Study [3]	AliveCor	94.4	81.9	54.8	98.4	N/A	N/A
Study [7]	Vector Machine Learning	N/A	N/A	N/A	N/A	82	N/A
Study [9]	Irregular Rhythm Notification	N/A	N/A	78.9	N/A	N/A	N/A
	ECG 1.0	98.3	99.6	N/A	N/A	N/A	N/A
	ECG 2.0	98.5	99.3	N/A	N/A	N/A	N/A
Study [11]	Convolutional Neural Network	79	79.5	N/A	N/A	39.2	0.87
Study [14]	Irregular Rhythm Notification	N/A	N/A	84	N/A	N/A	N/A
Study [15]	ECG	41	100	N/A	N/A	N/A	N/A

C. Apple Watch evaluation

Though without direct use of the common medical indexes, another six studies add to the comprehensive evaluation of Apple Watch apart from the above three studies. Table 4 lists the studies and the corresponding topics of interest. They are intended to assess different aspects of Apple Watch: Study [4] and [13] determine the Apple Watch's accuracy of measuring heart rate, and heart rate variability, respectively. Study [5] compares the ECG generated by Apple Watch to a standard 12-lead ECG, and Study [19] supplements the ECG evaluation by concentrating on the QTc interval measurement. Finally, Study [12] is designed to comprehensively discuss the influence of using Apple Watch on patients' life, while Study [18] focuses on diagnostic testing at clinical settings as the consequence of using Apple Watch's abnormal pulse notification.

As reported by Study [5], a moderate to strong agreement is shown in the Apple Watch generated ECG.[45] Specifically, the agreement between the AW ECG and the 12-lead ECG is analyzed in terms of heart rate detection, RR, PR, QRS, ST, QT, and QTc intervals.[46] A weakness for the AW ECG is the fact that it only carries lead 1 information, which may largely differ from other information carried by a standard 12-lead ECG.[47] Therefore, though AW can accurately measure heart rate and interval lengths on healthy subjects in this study, its effectiveness and potential on directing clinical decisions should be further tested when it comes to a wider population with a variety of medical pathologies.[48] Study [19] adds to the evaluation of AW ECG by validating its QTc measurement, and a similarly strong agreement is observed.[49] The heart rates detection for AW ECG matches that on a 12-lead ECG, and all patients identified as high risk are identified by the smartwatch.[50] The QTc measurements is adequately accurate, and when adjusting the smartwatch position on patients, the AW ECG performs better, indicating a need for identification of the best smartwatch position.[51]

Study [4] presents the heart rate measurement from AW Series 4, suggesting a correlation coefficient (r_c) of 0.7 between AW readings and the telemetry.[52] The r_c for patients who were in AF is larger than that for those who were not ($r_c = 0.86$ for patients in AF, and $r_c = 0.64$ for patients not in AF); this, nevertheless, could be caused by patients' own awareness of AF conditions, with those who are in AF being more careful and thus precise at getting the AW reading.[53] Similarly, Study [13] carries out an evaluation for AW's ability of measuring heart rate variability using a standard ECG as reference and yields a reasonable agreement between the two.[54] However, as observed

[45] Saghir et al., supra note 3.
[46] Id.
[47] Id.
[48] Id.
[49] Marc Strik et al., Validating QT-Interval Measurement Using the Apple Watch ECG to Enable Remote Monitoring During the COVID-19 Pandemic, 142 Circulation 416–418 (2020).
[50] Id.
[51] Id.
[52] Seshadri et al., supra note 4.
[53] Id.
[54] Turki et al., supra note 22.

during the experiment, the watch must be worn properly tight, otherwise an inaccurate measurement will occur.[55]

Study [18] analyzes the healthcare utilization following the irregular pulse notification function of Apple Watch.[56] The result shows that a clinical actionable cardiovascular diagnosis after only occurs in 11.4% (30/264) patients.[57] Patients who experienced symptoms are more likely to undergo clinical diagnosis than those who did not, while there is no difference of seeking clinical diagnosis between patients who received a direct alert from the pulse detection and those who did not.[58] The limited amount of clinical actionable diagnosis indicates a high false positive rate among the AW's function of abnormal pulse notification, which could potentially lead to an excessive use of healthcare resources.[59] On the other hand, being a more general assessment of the impact of Apple Watch on patients' quality of life and healthcare utilization, Study [12] leverages a more patient-focused experiment through using the Atrial Fibrillation Effect on QualiTy-of-life (AFEQT) questionnaire score as the primary.[60] Secondary outcomes include a set of patient-reported information regarding the use of personal digital devices, healthcare utilization investigation, and the data of Apple Watch using, such as number of irregular rhythm notification or heart rate records.[61] Although this study is not yet to be done, it shows the need for a patient-centered environment for research regarding personal health devices. There are also limitations of the study. First, there might be a possibly incomplete ascertainment as to patients' use of healthcare systems.[62] Second, people with AF might be more likely to purchase AW, even though those with a history of AF are not recommended for using ECG feature on AW.[63]

Table 4 *Corresponding Topic of Interest for Apple Watch Studies*

Apple Watch Studies	**Topic of Interest**
Study [4]	Heart rate accuracy
Study [5]	ECG generation
Study [12]	Influences on patients
Study [13]	Heart rate variability accuracy

[55] Id.
[56] Kirk D Wyatt et al., Clinical evaluation and diagnostic yield following evaluation of abnormal pulse detected using Apple Watch, 27 Journal of the American Medical Informatics Association 1359–1363 (2020).
[57] Id.
[58] Id.
[59] Id.
[60] Sanket S. Dhruva et al., Apple Watch and Withings Evaluation of Symptoms, Treatment, and Rhythm in those Undergoing Cardioversion (AWE STRUCk): A Pragmatic Randomized Controlled Trial (2021), http://medrxiv.org/lookup/doi/10.1101/2021.07.10.21260230 (last visited Feb 19, 2022).
[61] Id.
[62] Id.
[63] Id.

Study [18]	Occurrence of follow-up diagnostic testing
Study [19]	QT interval

D. Other devices and evaluation

Three studies discuss devices and algorithms other than Apple Watch using their own criteria of evaluation. Study [6] introduces the ECG-Watch as a low-cost wearable device, being able to provide heart records with a 10-second single ECG and a built-in algorithm capable of detecting AF.[64] This study indicates the portability of wearable devices as the primary advantage over the traditional 12-lead ECG, because in a traditional setting, patients and physicians should be in the same physical location to carry out an ECG recording, rendering sporadic ECG anomalies such as AF undetected in most cases.[65] The experiment section of the study shows a favorable agreement between the ECG-WATCH recording and the 12-lead ECG.[66] The other study, Study [10], provides a brief introduction to the Glasgow Algorithm, an ECG interpretive algorithm, and compares its criteria of myocardial infarction to the AHA/ACCF/HRS recommendations for such criteria.[67] It is shown that the Glasgow Algorithm's capability exceeds the standard criteria of MI detection.[68]

Study [2], on the other hand, is a comprehensive review of the commercially available devices in the detection of paroxysmal AF.[69] Since such AF could be asymptomatic, ECG devices should either be capable of recording for a long period of time or have an auto-trigger function, otherwise false negative results could occur.[70] Figure 3 shows the duration of monitoring for each kind of devices verses their burden on patients, as evaluated by Study [2]. Patch-type devices, such as eMemo and Zio Patch, are the most well-balanced in terms of the duration and burden, with a small number of electrodes and no electrode leads.[71] The currently most reliable devices for long-term detection of AF are insertable cardiac monitors (ICM); however, they are quite invasive compared to the wearables.[72] Ambulatory ECGs (AECG) have the highest AF diagnostic accuracy due to its largest number of electrodes, yet the conventional AECG can only record for 24 hours.[73] While there are commercially available long-term AECGs, it still might be burdensome to patients because the

[64] Randazzo, Ferretti, and Pasero, supra note 6.

[65] Id.

[66] Id.

[67] Stryker Emergency Care, What is the Glasgow Algorithm? - stryker emergency care (2010), https://www.physio-control.com/uploadedFiles/learning/clinical-topics/The%20University%20of%20Glasgow%2012-Lead%20ECG%20Analysis%20Algorithm%203304421.B.pdf.

[68] Id.

[69] Hashimoto, Harada, and Kasamaki, supra note 8.

[70] Id.

[71] Id.

[72] Id.

[73] Id.

electrodes cannot be changed throughout the observation.[74] Finally, with the development of algorithms, healthcare products such as the Apple Watch are available in the market, but their accuracy of detection is still under further investigation.[75]

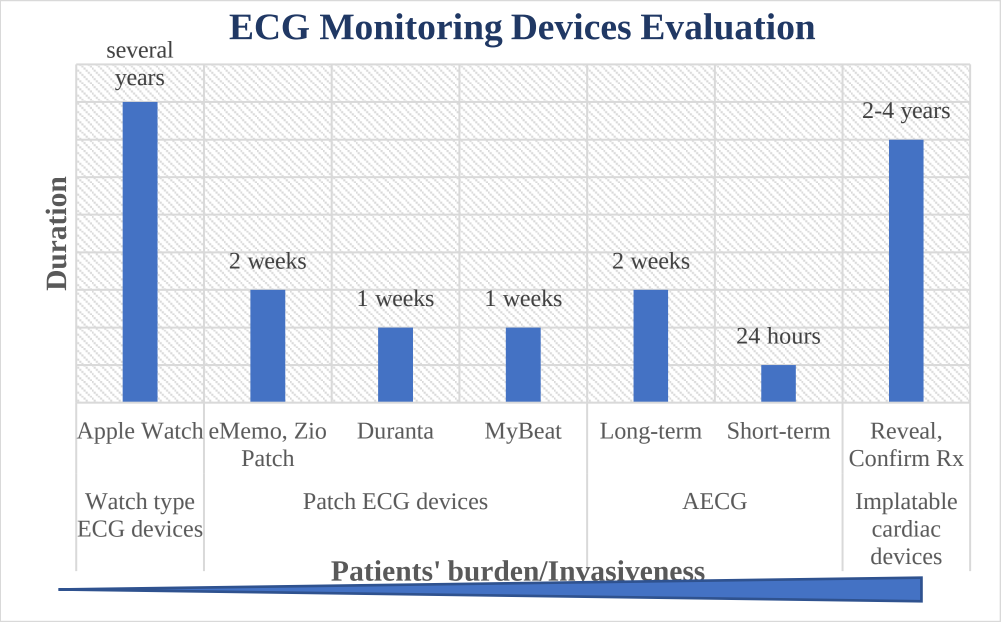


Figure 3 *ECG Monitoring Devices Evaluation*

E. Consequences Brought by Wearable Technology

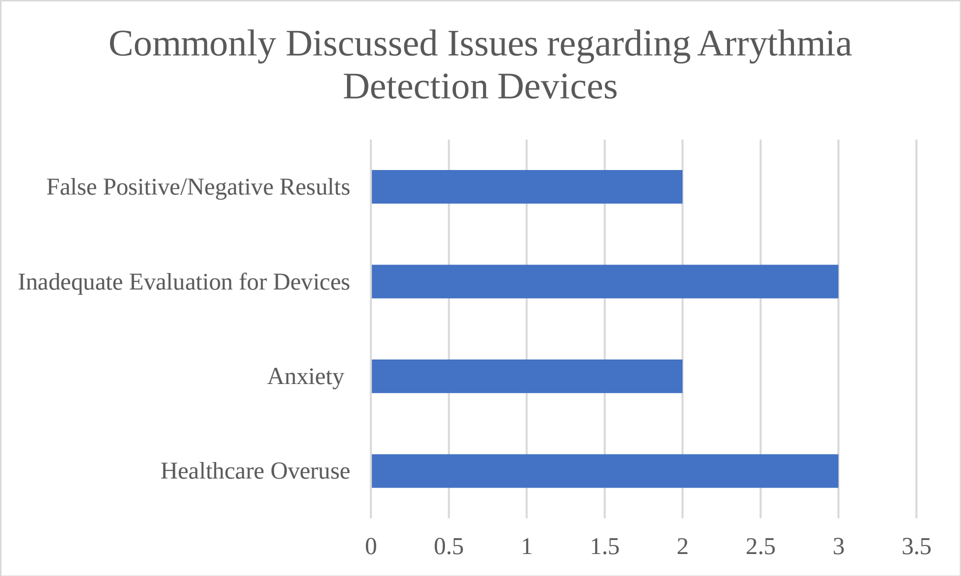
Apart from all the evaluations for the arrythmia detection devices and algorithms, four studies manifest some undesired problems and consequences brought by such technology. Figure 4 summarizes the commonly discussed issues among those studies. Study [8] focuses on the substantially increasing healthcare utilization due to personal ECG devices, which could cause a burden on the cardiology services.[76] In particular, two of the currently most reliable devices, AliveCor and Apple Watch, give the responsibility of any result other than normal sinus rhythm back to users by suggesting them to consult a physician, essentially causing the extra workload for cardiac physiologists.[77] This study points out that further developments in deep learning-based detection algorithms might result in a reduction of any false positive results and a better use of personal ECGs.[78]

Study [16] seeks for the opinions of healthcare professionals (HCPs) in terms of their advice for the available wearable devices/apps for AF.[79] 57% of respondents have suggested using these devices; among the respondents, electrophysiologists and general cardiologists are more likely to advise their uses compared to other specialist

[74] Id.
[75] Id.
[76] Rob Brisk et al., Personal ECG Devices: How Will Healthcare Systems Cope? A Single Centre Case Study (2019), http://www.cinc.org/archives/2019/pdf/CinC2019-335.pdf (last visited Feb 19, 2022).
[77] Id.
[78] Id.
[79] Boriani et al., supra note 19.

physicians.[80] When asked about the disadvantages of using wearable devices/apps, 65% HCPs refer to the anxiety in people who test positive for AF, and 40% point to the false reassurance because of a negative result.[81] Lastly, the study emphasizes the need for clarifications of the validated devices and Apps so that patients are assured to use those that are rigorously tested.[82] Study [17], on the other hand, compares the degree of healthcare use for those who use wearables and those who do not, concluding that while the mean pulse rates was similar between the two cohorts, individuals using wearables had a higher healthcare use score.[83] The need for more data to guide the utilization of wearable devices is again supported.[84]

Study [20] illustrates the health anxiety among people from the use of wearable devices by describing a case of a 70-year-old woman with paroxysmal AF.[85] There was an excessive use of smartwatch for cardiac monitoring one year after her initial AF diagnosis, and the patient was shown to believe that notifications from the smartwatch were a sign of worsening cardiac function, leading to additional clinical visits.[86] As noted by the study, ambiguous data including inconclusive readings could trigger similar behavioral response as compared to irregular rhythm notification.[87] The study concludes that compared to the traditional clinics, wearable devices strengthen the personal access to health data, which could bolster the belief that those already under appropriate therapy should still use such data to seek for medical care, which is in fact not necessary.[88]



[80] Id.
[81] Id.
[82] Id.
[83] Libo Wang et al., Association of Wearable Device Use With Pulse Rate and Healthcare Use in Adults With Atrial Fibrillation, 4 JAMA NETW OPEN e215821 (2021).
[84] Id.
[85] Lindsey Rosman, Anil Gehi & Rachel Lampert, When smartwatches contribute to health anxiety in patients with atrial fibrillation, 1 CARDIOVASCULAR DIGITAL HEALTH JOURNAL 9–10 (2020).
[86] Id.
[87] Id.
[88] Id.

Figure 4 *Commonly Discussed Issues regarding Arrythmia Detection Devices*

III. Discussion

Wearable technology has allowed people to participate in their own healthcare management, which is the main goal of connected health. The emergence of more and more wearable detection devices should come by no surprise. This prompts FDA to publish standards that ensure an effective validation process for these devices. Rigorous research must be done in accordance with the FDA standards for SaMD to maximize the utility of wearable devices and minimize unfavorable consequences. This study thus provides insight regarding the current position of wearable technology by reviewing and comparing its published validations and discussions.

As classified in the Result section, the 20 studies included in this review use various metrics, include different cohorts, and focus on different perspectives of the problem. While it is difficult to discuss all the studies in a shared context, this review will analyze the studies consistently by referencing studies making connection throughout the discussion. Specifically, quantitative assessments, including indications behind the data, will be compared under the same context, as to whether the device/algorithm performs well under evaluation and obtains enough validation to be in use. Studies arguing about the problems of wearable technology will be combined for a wholistic review of the barriers of this technology, and a guidance for future research directions.

To start with, Apple Watch receives the most research interests among watch type devices, following by a few other devices such as AliveCor and ECG WATCH. Study [2] summarizes the advantages and disadvantages of different kinds of ECG devices in terms of their duration of monitoring and the burden on patients, indicating a tradeoff between the two.[89] While watch type devices have the longest duration of monitoring because they can be worn all the time, they do not have the approved accuracy compared with other prescriptive ECG devices.[90] Studies evaluating them are therefore valuable for any decision to incorporate them into clinical use.

Published by Apple, Study [9] reports a positive predictive value of 78.9% for Apple Watch's irregular rhythm notification, and 98.5% for its ECG 2.0, in detecting AF.[91] Study [14] obtains a similar positive predictive value of 84% for the irregular rhythm notification function when patients with a first notification received a second notification and a clinical ECG concurrently, but among all participants, AF was only observed in 34% of those receiving a notification on a clinical ECG performed later.[92] The actual predictive value would not be as low as the second figure since AF is usually paroxysmal, yet this finding points out the need of a durable monitoring function for wearable detection devices if they are applied to the real world setting, as pointed out by the 5th FDA action plan.[93] For the detection algorithm along with Apple Watch's ECG, Study [15] reports a much smaller sensitivity of 41% in distinguishing AF and

[89] Hashimoto, Harada, and Kasamaki, supra note 8.
[90] Id.
[91] Apple, supra note 24.
[92] Perez et al., supra note 21.
[93] FDA, supra note 17.

proposes a potential solution of combining a rhythm assessment with the AW generated ECG.[94] This indicates the need for a better algorithm to classify AF and other types of cardiovascular diseases with the generated ECG. Study [5] and Study [19] provide information regarding the evaluation of ECG diagrams generated by AW, both of which show strong agreement between AW ECG and a standard 12-lead ECG.[95] However, there are some drawbacks of using AW, including its sensitivity to motion artifacts and the limitation of using only 1 lead in the generation of ECG.[96] Similarly, the heart rate and heart rate variability measurements are mostly accurate as assessed in Study [4] and Study [13], but are easily affected by wearing positions to maintain such accuracy.[97] These results support the fact that Apple Watch is indeed currently one of the most reliable wearable arrhythmia detection devices, yet not as positive as the Apple's report, studies reveal a decent number of issues through experimentation. To fulfill the FDA's actions of promoting GMLP and an effective evaluation process, improvements in algorithms are needed. AliveCor and ECG WATCH are also brought into consideration in Study [3] and Study [6], respectively. The AliveCor assessment reaches a sensitivity of 94.4%, with a significant number of unclassified readings and false positives.[98] Study [6], on the other hand, is a brief introduction for ECG WATCH, which does not include much data for analysis. However, it comments on the usefulness of other devices: for example, AliveCor, as discussed in Study [6], filters too much signal and thus loses important information regarding heart activity.[99] In general, while Study [6] and Study [19] show confidence in the further use of the wearable devices, other studies such as Study [4] and Study [13] assert needs for further validation before its use in aiding clinical decisions. In addition, as the false positive rates and unclassified readings are still the main issue of wearable technology, more advanced algorithms need to be tested and put in practice.

Most of the currently available algorithms apply Machine Learning in the detection of cardiac diseases as a more inclusive and precise approach. The four studies describing algorithms feature MI^3, Vector Machine Learning, Glasgow Algorithm, and Convolutional Neural Network, respectively. The MI^3 algorithm in Study [1] is trained to output an indicator that takes into consideration of various influential factors, proving that the proposed algorithm is more adaptive to different conditions with better performance than traditional algorithms.[100] In Study [7], the Vector Machine Learning algorithm can classify different types of ECG and thus is potentially applicable in the wearable devices as a powerful ECG reading tool.[101] The Glasgow Algorithm in Study [10] employs multiple factors in detecting myocardial infarction, which is similar to MI^3 in terms of their finer resolution for thresholds.[102] Study [11] presents a Convolutional Neural Network trained to detect the signatures of AF patients' ECG, which is essentially the most relevant to the current issues of false results produced by

[94] Seshadri et al., supra note 30.
[95] Saghir et al., supra note 3; Strik et al., supra note 49.
[96] Saghir et al., supra note 3.
[97] Turki et al., supra note 22.
[98] Rajakariar et al., supra note 11.
[99] Randazzo, Ferretti, and Pasero, supra note 6.
[100] Than et al., supra note 15.
[101] Walsh, supra note 1.
[102] Stryker Emergency Care, supra note 67.

wearable devices.[103] These algorithms are continuously proven to be effective, and further research should focus on the comparison among algorithms to determine the specialty of each of them. If systematically trained and tested, they can be essential in the popularization of wearable technology.

There are many potentially unfavorable problems with this market of wearable technology when they are used in reality. The biggest issue, as mentioned in Study [8], [12], [17], and [18], is the healthcare overuse after a more personal level of engagement realized by wearables.[104] Because of the false positive results, the frequency of people seeking for clinical assurance of their health is bound to grow. Both Study [17] and [18] indicate that such increasing amount of healthcare utilization is not proportional to the actual effect on patients' health[105].

What comes next is the patients' anxiety induced by overusing wearables. Study [16] reports the opinions of healthcare professionals, most of whom point out the problem of anxiety, especially due to false positive results.[106] Additionally, Study [20]'s description of a patient suffering from anxiety reveals the fact that a lot of people are not informed with the correct way of using this technology and interpreting its results.[107] In fact, being an unprecedently penetrative tool, wearable device is almost inevitably overused by patients regardless of its accuracy. This information gap is also present in the validation of devices: as the market spreads rapidly without restriction, too many devices are available, whereas few of them are validated to be in use. The 3rd and 5th FDA action plan both points to the importance of a patient-centered environment for the use of SaMD.[108] To fulfill these goals, actions to effectively promote transparency for consumers and control the wearable technology in reality are necessary.

There are also problems with regard to the nature of wearable technology. Study [4] reflects on whether the awareness of having AF affects people's use of Apple Watch, because readings from the watch have to be manually obtained, those without being aware of a possible AF might not have the incentive to obtain a reading, and therefore cannot find out their real conditions.[109] Furthermore, there is an asymmetry between the buyers and users of wearable technology: most young people consume products like Apple Watch, but those who need to be tracked with their heart activity are normally the elders. If wearable devices are to be implemented to a large scale, these issues must be considered and resolved to realize an effectively controlled and managed environment to the interest of the public.

Conclusion

Studies examining wearable devices and potentially applicable algorithms show relatively high accuracy for the detection of AF. Still, almost all of them indicate a need for further research on a larger cohort with a similar structure to the real world, one of

[103] Attia et al., supra note 5.
[104] Brisk et al., supra note 76 at 1; DHRUVA ET AL., supra note 60; Wang et al., supra note 83; Wyatt et al., supra note 56.
[105] Wang et al., supra note 83; Wyatt et al., supra note 56.
[106] Boriani et al., supra note 19.
[107] Rosman, Gehi, and Lampert, supra note 85.
[108] FDA, supra note 17.
[109] Seshadri et al., supra note 4.

the FDA action plans for SaMD. There are also underlying problems of the wearable technology use, most primarily due to the high level of personal engagement in healthcare it provides, such as the overuse of healthcare resources and patients' anxiety given the unlimited access to data, especially when false-positive results may occur. These issues must be formally resolved before the large-scale application of wearable technology in aiding clinical decisions. Given the rapid development of connected health, wearable devices such as the Apple Watch are bound to become the mainstream tool for this more patient-centered environment of healthcare distribution. However, many of them do not conform to the FDA regulation for SaMD. Therefore, any new wearable device should be under a formal validation process so that such technology can be appropriately applied to the public.