National Physical Laboratory

NPL REPORT MS 61

SCREENING OF ATRIAL FIBRILLATION USING WEARABLE PPG DEVICES – A TRUSTWORTHY AND SAFE AI LIFE CYCLE CASE STUDY

M. ALSULEMAN, P. DUNCAN, A. THOMPSON

JUNE 2025

Screening of Atrial Fibrillation using Wearable PPG Devices – a Trustworthy and Safe AI Life Cycle Case Study

M. Alsuleman, P. Duncan, A. Thompson
Data Science Department

ABSTRACT

NPL recently developed a Trustworthy and Safe AI Life Cycle (TSALC) which links the risks associated with AI systems to the software development process and AI trustworthiness metrics. In this report we apply the principles of the TSALC to a case study: the detection of atrial fibrillation (AF) using a wearable AI-based medical device equipped with a photoplethysmography (PPG) signal sensor. A hypothetical scenario is proposed in which the device is used for clinician-initiated screening of patients suspected of having AF, in order to inform prioritisation of diagnosis resources (for example ECG monitors). We describe a methodology for carrying out risk analysis and mitigation, inspired in part by NIST's AI Risk Management Framework (NIST AI RMF). Assimilating input from application experts drawn from academia, industry and the NHS, we identify specific risks and mitigations. Based upon these risks and mitigations, we outline the most relevant aspects of AI trustworthiness and the associated software engineering challenges. We also describe metrics which can be used to quantify some of these aspects of AI trustworthiness. As well as representing a useful case study in its own right, we expect that the methodology described here will be applicable to numerous other AI systems.

ISSN 1754-2960

https://doi.org/10.47120/npl.MS61

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of NPLML by
Louise Wright, Head of Science (Data Science).

CONTENTS

1 INTRODUCTION

AI system development has exploded in recent years with AI being used across multiple industries and international domains, yet it is not currently clear how regulation and standardisation can aid in addressing the growing concerns about how safe or trustworthy AI systems are for public use. Through the UK's National AI Strategy, the AI Standards Hub was established by three parts of the UK National Quality infrastructure:

- The Alan Turing Institute - National Institute for Data Science and AI
- The British Standards Institute - National Standards Body
- The National Physical Laboratory - National Metrology Institute

to bring together industry, government, regulators, consumers and civil society, and academia to promote sound, coherent and effective standards to inform and strengthen AI governance practices domestically and internationally. Through the research element of this work, the National Physical Laboratory (NPL) has led international research and analysis on key strategic questions related to AI standardisation: how to ensure that the AI systems we adopt and develop are trustworthy (they do what we expect them to do) and safe (they will not cause unexpected harm). NPL developed the Trustworthy and Safe AI Life Cycle (TSALC) framework which links the risks associated with AI systems to the software development process and trustworthiness metrics [1].

This report describes a case study in which the principles of the TSALC framework are taken and applied in the context of a wearable AI-based medical device equipped with a photoplethysmography (PPG) signal sensor. The purpose of the (hypothetical) device is to support the prioritisation of resources for atrial fibrillation (AF) for patients suspected of having the condition. We believe that, as well as being useful in its own right, this case study demonstrates the wider applicability of the approach and that the methodology followed in this case study is relevant to many other AI systems and products.

We give background on the challenge of atrial fibrillation detection from PPG signals in Section 2 before describing our hypothetical product addressing this challenge in Section 3. We then present an outline of NPL's TSALC in Section 4, focussing particularly upon risk analysis which underpins the approach. In Section 5, we then apply the TSALC in the context of the proposed product, describing both our methodology and our results, before concluding in Section 6.

2 ATRIAL FIBRILLATION DETECTION FROM PPG SIGNALS - BACKGROUND

Atrial Fibrillation (AF) is a heart condition characterised by episodes of rapid and irregular heartbeat. AF is extremely prevalent: it is estimated that 1 in 45 people in the UK are living with the condition [2], and it has been described as a 21st century cardiovascular disease epidemic [3]. Patients with AF are at increased risk of heart failure and dementia, and AF contributes to 1 in 5 strokes in the UK [4].

The risk of stroke and other ailments can be reduced with appropriate anticoagulation treatment (blood thinners), but episodes are underdiagnosed using current clinical procedures, and it is estimated that there could be around half a million undiagnosed cases of AF currently in the UK [4]. On the other hand, blood thinning medication can also lead to serious side effects such as intercranial bleeding [5], and so it is also important to reduce the risk of incorrect positive diagnosis.

Digital health solutions are often used for the timely detection of AF, and the gold standard for diagnosis of AF is the Electrocardiogram (ECG). The guidance issued by the National Institute for Health and Care Excellence (NICE) [6] is that AF should be diagnosed using a 12-lead ECG upon manual detection of an irregular pulse. If patients have undiagnosed but

suspected AF, the recommendation is to fit the patient with a wearable ECG monitor for a period of 24 hours, or longer if symptomatic episodes are more than 24 hours apart.

Photoplethysmography (PPG) technology has more recently emerged as an alternative mobile technology which can detect AF episodes. PPG uses a light source and a photodetector at the surface of skin to measure volumetric variations in blood circulation [7]. PPG wearables are appealing due to their cost effectiveness, their simplicity of operation, and their wearing comfort for users [7]. The use of PPG alone for diagnosis of AF is thought by many to be unrealistic, and as previously stated current guidelines require that an ECG is used to establish a diagnosis. However, mobile PPG technology might still be used to support AF diagnosis, and two different ways that this could happen can be distinguished: *screening* and *monitoring*. Screening refers to the identification of people who may have an increased risk of AF, even if they have not presented with symptoms that might suggest AF. Screening is typically initiated by healthcare professionals, but the possibility of self-screening has also been recently considered. Monitoring, on the other hand, refers to the use of a mobile PPG technology to monitor patients who have a known disease or condition such as AF in order to make decision about current or future treatment. In this case study, we will focus on the use of mobile PPG technology for *clinician-initiated screening* of patients suspected of having AF, and in Section 3 we describe the use case in detail and explain our reasons for selecting it.

A mobile AF detector must have the capability to automatically detect AF, which means that a vital component of any such technology is signal processing and data analysis which infers the presence or absence of AF from the raw signal. Typically, short segments of heart rhythm data (for example 30-second windows) will be analysed to determine if AF is present. It is known that certain interpretable features such as inter-beat intervals are clinically relevant [8]. However, there is no comprehensive scientific understanding of how AF can be detected from a heart rhythm segment. For this reason, Artificial Intelligence (AI) techniques which employ machine learning (ML) and data-driven modelling have been developed, and this case study focusses upon the potential use of an AI algorithm for AF detection from PPG measurements.

There are now several PPG-enabled devices or applications in circulation. Several smartwatch developers include PPG sensors and functionality for detecting AF in their devices, including Apple, Samsung, Huawei and Honor [9]. Studies such as the Apple Heart Study [10] and the Huawei Heart Study [11] have shown promising indications that PPG technologies in smartwatches are indeed able to detect AF. Meanwhile, Fibricheck have released a PPG-based AF detection smartphone app which has demonstrated promising results in studies, for example [12]. The use of AI algorithms within some of these devices, including the Apple Watch and the Fibricheck app, has also been demonstrated [13]. By taking a data-driven approach, AI algorithms have the potential to spot patterns that elude rule-based approaches, thereby giving improved classification performance. On the other hand, as introduced in Section 1, the use of AI brings with it significant additional risks and challenges for ensuring its trustworthiness.

3 AI-BASED SCREENING OF ATRIAL FIBRILLATION - DESCRIPTION OF A HYPOTHETICAL PRODUCT

In this section, we introduce a theoretical product that includes an AI component. The definition of the product is important, especially for medical applications, because it provides the product claim or benefit statement. This claim sets expectations for the users and unifies the stakeholders towards a shared goal through risk assessments, development, evaluation and deployment stages. Moreover, it helps to identify the regulation which is applicable to the product. For example, a device that claims to count heartbeats to indicate heart activity level has different expectations from a device that is used for medical decisions, such as

diagnosing AF. Furthermore, such expectations can serve as the cornerstone of a trustworthy AI system, offering transparency to stakeholders, promoting beneficence, and imposing commitment and responsibility upon the developer and deployer. In summary:

The product is a wrist-worn wearable device that helps the NHS to prioritise a waiting list for efficient resource allocation of ECG records for AF diagnosis.

As indicated earlier, there is thought to be a high number of underdiagnosed cases of AF in the UK [4], and recording ECG data for sufficient time to catch occurrences of AF for each patient might overwhelm the NHS and demand for resources needed to record and review all the generated data. Alternatively, a wrist-worn wearable device equipped with a PPG signal can perform initial AF detection, providing information about the frequency and severity of AF episodes, allowing more severe cases to be further investigated using ECG devices. It is not intended that this PPG-based device provides a diagnosis, as current guidelines require the use of ECG for diagnosis [6]. Rather, the tool assists a doctor in making informed decisions backed by objective assessments. It is worth noting that AF might present with persistent irregular heartbeats, such as AF levels 3B, 3C, and 4 [14], which are easy to catch on a short ECG record. On the other hand, levels 1, 2, and 3A irregularities are less persistent and may require longer ECG recordings (24, 72 hours or more) to identify, making these patients harder to diagnose. We think that the second group would benefit from this device.

Figures 1 and 2 show a schematic of the device and the expected steps for it to be successfully used as a screening device. The scenario assumes that patients come to the clinic with health complaints and that the doctor sees a benefit in checking for the presence of AF or ruling out this cause of the patient's symptoms. Therefore, the patient is added to a waiting list for ECG recording, which is the diagnostic tool for AF. All the patients on the waiting list get the device (initiated by the clinician), which starts to collect data and classify them based on their health and possible AF burden. The users will be provided with instructions on using the device. Signals affected by external causes such as excessive motion can be identified using other hardware and software components such as accelerometer sensors. The available data will be analysed, and actionable decisions will be presented in a readable, easy-to-understand way to the doctor and patients. The result can be summarised as the likelihood of having AF, the confidence level, and the length of ECG needed to make a diagnostic decision.

The device consists of multiple components, including additional sensors to ensure good signal quality. For example, the accelerometers help exclude motion artefact signals. However, the focus of this paper is the trustworthiness of an AI component of the system. Therefore, the quality of these other components will not be discussed in detail, and regular policies and procedures applied in NPL for other products are applicable here. This means that project management, software quality, cybersecurity, and risk management procedures at NPL are applied, and the team adheres to NPL values. As an example of software quality compliance, NPL software projects within the Data Science department adhere to TickITplus (combined ISO9001, ISO20000-1, and ISO27001) [15]. Moreover, any project at NPL requires a technical lead and project manager; thus, project management, reporting, reviewing, training, and regular technical meetings are assumed to be carried out to manage and produce this AI model, and all steps are well-documented. If a non-AI component is mentioned, it is with a view to setting additional requirements that might affect the AI component.

This product is meant to be hypothetical, reflecting our current best understanding of how PPG technology might most effectively assist clinicians with AF diagnosis. Although the product/scenario is hypothetical, it does not require the development of hardware or PPG

data acquisition beyond the current state-of-the-art. Instead, it focuses on how to design, test, and operate the AI system to achieve safety and trustworthiness.

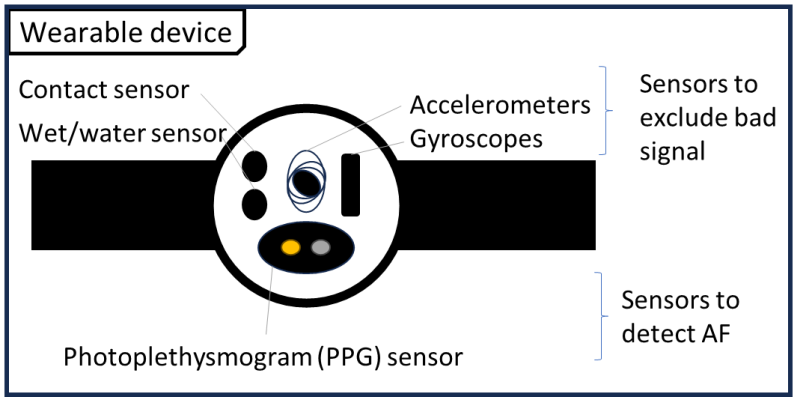

Figure 1. A schematic of the device.

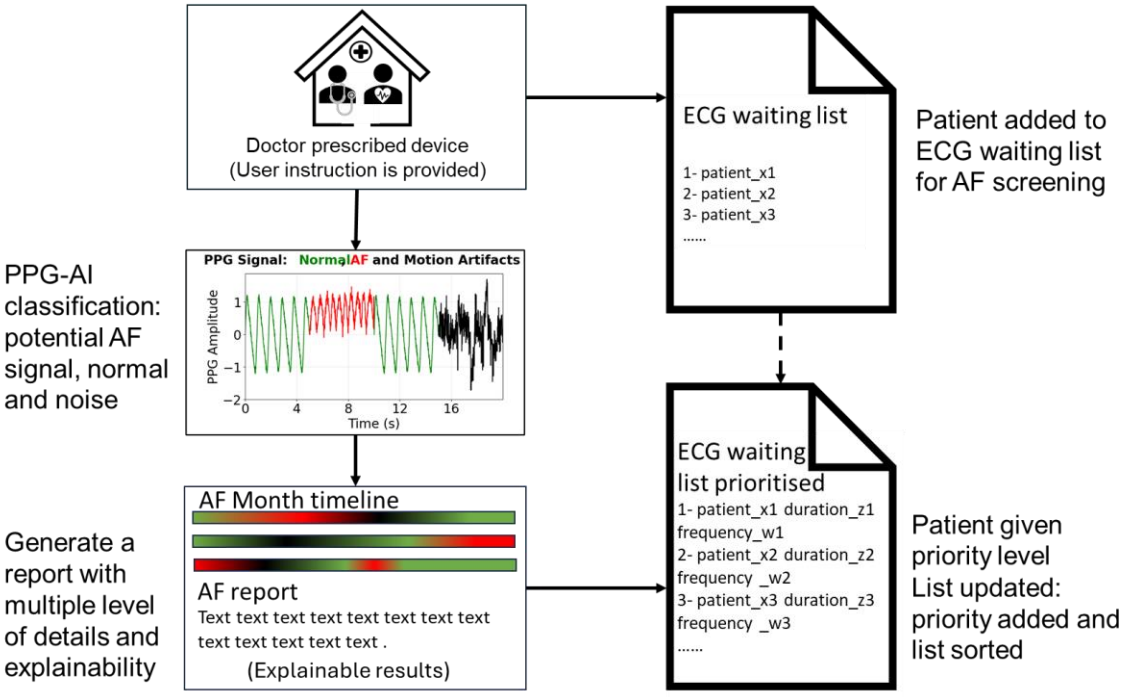

Figure 2. The steps involved in using the device and its outputs.

4 THE TRUSTWORTHY AND SAFE AI LIFE CYCLE

Understanding that AI system development requires a different approach to traditional software development, through the AI Standards Hub, NPL has been developing a life cycle approach to identify, understand, manage and, crucially, integrate the risks associated with AI systems into the system development: the NPL Trustworthy & Safe AI Life cycle (TSALC) as depicted in Figure 3. The TSALC comprises three interconnected layers: *Risk Analysis*, *Software Engineering for AI* and *Measurement for Trustworthy AI*, and each layer contains a specific methodology, method and artefact.

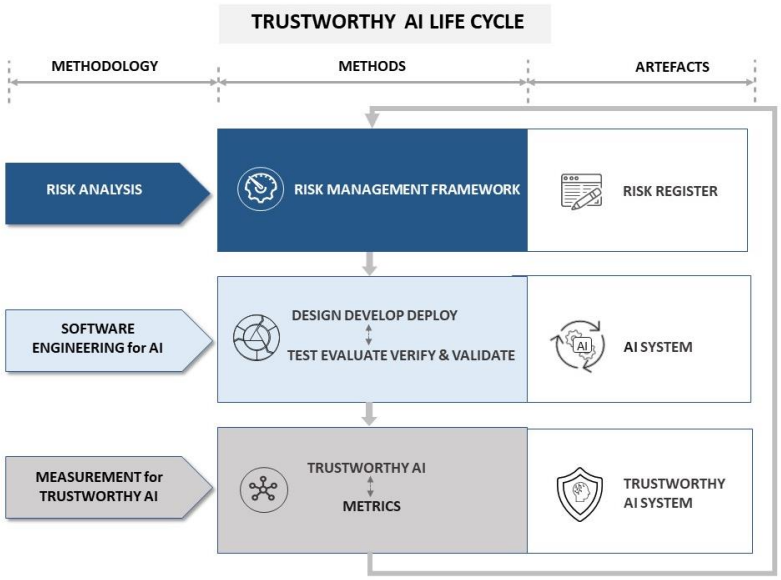

Figure 3. The Trustworthy AI Life cycle [1].

The first layer in the TSALC is concerned with the activity of *Risk Analysis* for identifying risks and assessing their potential likelihood and consequences. The NPL TSALC builds upon the principles of the NIST AI Risk Management Framework (NIST AI RMF) [16], which provides guidance on how to "govern, map, measure and manage" the risks associated with AI system development. Using tools like the NIST AI RMF will allow developers of AI systems to recognise and document the risks and benefits arising from the AI system being developed. Then, the development of a risk register will facilitate the necessary steps to track these risks and understand how they affect the *Software Engineering for AI* and *Measurement for AI* layers. There are several other international AI frameworks beyond the NIST AI RMF which aim to assess and mitigate risk within the development and use of AI systems in differing ways. These include CDEI's AI assurance roadmap [17], the European AI Act [18], ISO 23894:2023 (Information technology – Artificial intelligence – Guidance on risk management) [19], ISO 42001:2023 (AI management systems) [20], as well as OECD's Tools for trustworthy AI [21].

Once risks have been assessed and tracked in a risk register, the next layer of the TSALC is concerned with using the knowledge of risks to inform the development of the system itself: *Software Engineering for AI*. This speaks to the overarching goal of the TSALC which is to have each stage of AI system development be aware of and be influenced by the risks present in the system, ensuring that the mitigation of the risks is embedded in the system itself. The *Software Engineering for AI* layer is concerned with how the traditional design, develop and deploy (DDD) and test, evaluation, verification and validation (TEVV) cycles should be adapted for developing AI systems while being cognisant of the risks. It is important to note that the TSALC is iterative and continuous at all stages, to ensure that an AI system is not released until it is sufficiently trustworthy and safe, resolving the catalogued risks.

The final stage of the TSALC is *Measurement for AI*, where metrics that are associated with the risks identified in the system are quantified to assess the overall trustworthiness of the system itself. For example, if a risk identified about the AI system is that the system may not be used safely due to its decision-making processes lacking *explainability*, then metrics which capture the extent to which the AI system is explainable would be evaluated. Quantification of each metric of trustworthiness then influences the iterative risk analysis and software engineering.

4.1 RISK ANALYSIS

As the outline above makes clear, the trustworthy AI life cycle is risk-based, so that development, evaluation and deployment are carried out with the end goal in mind. In this way, quality, safety and trustworthiness are built throughout the life cycle. We next describe the schema and methodology involved in a risk analysis for an AI system.

ISO 31000:2018 defines risk as "the effect of uncertainty on objectives". It further elaborates that the risk is expressed in terms of risk sources, potential events, their consequences, and the likelihood of occurrence [22]. ISO 14971:2019 is a dedicated standard for risk management for medical devices, which has a generally lower level of risk tolerance and involves various stakeholders [23]. It has been acknowledged that risk is complex to identify, as each stakeholder has different views on the potential risks and level of risk tolerance. To reduce the ambiguity of a risk, it can be broken down into its components, such as describing the risk event, its causes and consequences, and giving it a likelihood and severity on a normalised scale [24]. Furthermore, a single risk source can result in multiple consequences, and one or more sources may contribute to a particular consequence [23].

Monitoring and mitigation are actions to deal with the risk and are part of the risk management [23], [24]. Monitoring informs the organisation of changes in the status of risks and whether they are emerging or under control. It refers to a measurable value in the system environment associated with the risk event. Similar to how measuring temperature or carbon dioxide might help to detect the presence of a fire event, a statistically significant difference between the outputs generated by two distinct demographic subgroups might indicate a bias in the AI system. Mitigation includes actions to mitigate the risks.

The NIST AI RMF [16] consists of four interconnected functions: *Govern*, *Map, Measure* and *Manage*. The *Govern* function identifies applicable regulation and establishes organisational frameworks. The *Map* function is the consideration of the risk context throughout the process. The *Measure* function implements quantifiable metrics to assess and monitor the risk levels and track mitigation effectiveness. The *Manage* function identifies relevant procedures and develops plans to mitigate the risks.

In Section 5.1 we apply this risk analysis methodology to our case study.

4.2 DESIGN AND DEVELOPMENT

Possible mitigation plans are identified as an output of the first layer (risk register). These mitigation plans work as design requirements to be satisfied and development criteria to be met. Due to time constraints, we discuss the approach in this layer (Section 5.2) but we did not implement it.

4.3 METRICS FOR AI TRUSTWORTHINESS

Throughout the process the context of each risk is narrowed down and defined, which allows relevant aspects of AI trustworthiness to be identified, and along with them appropriate metrics for evaluating trustworthiness. Due to time constraints, this layer is discussed (Section 5.3) but not implemented.

5 APPLYING THE TRUSTWORTHY AI LIFE CYCLE TO AF DETECTION

In this section, we will discuss the application of the TSALC to the hypothetical product for screening of AF described in Section 3. The AI life cycle outlined in Section 4 consists of three layers: risk analysis, software engineering for AI, and measurement for trustworthy AI.

Each layer guides the next one and interacts with the other. We address each of these layers in the context of this concrete case study in Sections 5.1, 5.2 and 5.3 respectively.

5.1 RISK ANALYSIS

To identify the potential risks associated with the PPG-for-AF wearable device, one hour interviews were conducted with domain experts and AI practitioners from academia, industry and the NHS who specialise in PPG technology for atrial fibrillation detection or diagnosis. The interviews aimed to introduce the participants to the device (Section 3) and a set of questions to guide them in identifying the risk with a minimum level of ambiguity. For this purpose, a risk register template was developed (Table 1). The table was presented to the participants during the interviews, which were recorded, and subsequently the template was completed based on all the participants' responses. The participants had diverse backgrounds, including AI developers, hardware inventors, NHS practitioners, clinical studies researchers, and others.

Table 1. Risk register template used for risk identification.

Column header	Column Description
Risk:	- What is the risk? - A brief description of the risk if necessary.
Cause:	- Who / what is the cause? (individuals, organisation, technology, geography etc) - A brief description of the cause of the risk - Can you describe the cause: • Give an example how the cause will lead to the risk. • What is the nature of the cause technical (e.g. hardware, software, design, quality), social (e.g. expectation, culture) other factors (e.g. unknown health condition).
The affected or impacted:	- Who / what is affected? (person, reputation, social etc) - What is the nature of the cause human (e.g. patient) social (e.g. reputation), financial (e.g. money, material), technical (e.g. faulty device) … etc - Explain the impact further by giving an example. i.e. what would happen? - What is the scale of the impact (if applicable)? - Is there an acceptable threshold? If yes, what is it?
Severity and consequences (from 1 – 5):	1 = negligible 2 = minor 3 = moderate 4 = major 5 = catastrophic
Likelihood of the risk (from 1-5):	1 = rare 2 = unlikely 3 = possible 4 = likely 5 = almost certain
Monitoring:	How is the risk monitored? (give a brief outline of the plan) Supporting questions: - Quantifiable, any metrics, definition? - Frequency? (how often) - Does it change? (changing subject being monitored)
Mitigation:	- What type of mitigation (Tolerate, Terminate, Treat or Transfer)? - Plan (include the requirement and SMART goal) - What is the related **Trustworthy AI principle**? - What is the related **policy or procedure**?

5.1.1 Risk informed by the field/application

Domain-specific risks emerge from the contexts in which the systems are deployed, consisting of consequences beyond technical performance issues. In medical applications, these risks are closely linked with patient safety, clinical workflows, regulatory compliance and established standards of care. For example, the International Medical Device Regulators Forum (IMDRF) has released ten principles for Good Machine Learning Practice (GMLP) [25]. The GMLP principles emphasise involving multi-disciplinary expertise, applying robust data practices, testing, and continuous monitoring across the entire product life cycle and representing diverse patient populations. The GMLP-compliance process includes being user-centric, clinically relevant and transparent and results in augmenting healthcare capabilities rather than replacing them.

The IMDRF is an international voluntary group of medical device regulators that includes many national regulatory agencies, such as the Medicines and Healthcare Products Regulatory Agency (MHRA) from the UK, the Food and Drug Administration (FDA) from the USA, Health Canada, and others. On a national scale, these agencies release further guidance and compliance rules. The FDA, MHRA, and Health Canada released five further principles for predetermined change control plans (PCCPs), which build on the GMLP principles [26]. PCCPs help in safely managing ML-enabled medical device modifications through risk-based, evidence-driven, transparent, and holistic approaches throughout the device's life cycle and the involvement of all stakeholders. The same national agencies also elaborated on GMLP principles seven and nine which concern transparency of the AI system [27]. This guidance stresses a holistic understanding of users, environments and workflows, using human-centred design principles to make information accessible, personalised and responsive to user needs.

The increase in the use of AI/ML in various processes in heavily regulated sectors such as healthcare might raise concerns about the complexity of developing a safe and trustworthy AI system and how this might hinder innovation. However, this risk-based approach directs the effort to where it is most necessary. For example, not all software used in medical applications is considered to be a medical device [28]. Hence, the relevant regulation might vary based on each system. The FDA made its regulatory efforts accessible online [29] and provided a feedback meeting service to facilitate innovation. Furthermore, they published a document [30] that outlined the roles regarding the adoption of AI in medical products.

The stakeholders invited to the interviews had various types of expertise relevant to PPG-based AF detection, ranging from AI developers, hardware and PPG sensor experts, practitioners and researchers. Before the interviews, the interviewers reviewed the relevant regulatory documentation cited in this section. Thus, the first half of the one-hour interview was dedicated to prompting the interviewees' opinion on potential risks for the theoretical product (Section 3), while the interviewer only guided them through the risk aspects (Table 1) and invited the interviewee to expand upon statements congruent with regulatory guidelines. The following question was used as prompt for the first part of the interview:

Based on your experience and the proposed scenario, what are the possible risks in developing and implementing this AI system?

It is important to note that the consideration of risks in this work was not intended to be exhaustive due to time limits. The AF diagnosis and management guidance [14] has not been discussed in detail here, but it was considered in the product proposal, for example the

constraint that PPG cannot be used for diagnosis. Moreover, there are other regulation guidelines that have not been considered here, such as security [31] and health data privacy [32]. However, this document demonstrates the methodology and stresses the need for collaboration between AI practitioners and domain experts to identify and manage domain-specific risks.

5.1.2 Risk informed by the technology

In this section, we focus on the risks related to the technology used, specifically AI/ML. Although in the previous section we highlighted regulatory documents that consider the technology, their focus was the nature of the application, i.e., healthcare. To focus on the AI aspect during the second half of the interview, the interviewees were asked to read a set of questions and then share any thoughts concerning additional risks. These questions were inspired by the NIST AI RMF and NIST AI RMF playbook [16], [33]. Both documents provided a categorisation of risks for AI systems. The most relevant categories for the proposed product were those that dealt with technical robustness, bias and fairness, *human in the loop* (HITL), security, risk in a broader (legal, social, and moral) context, continuous evaluation, and privacy. To these categories we added uncertainty quantification, which is an important but often-overlooked risk category for devices supporting clinical diagnosis.

Table 2 contains the questions that were introduced in the second half of the interview. The purpose of these questions is to put more emphasis on potential vulnerabilities in the proposed AI system. The questions concerning technical robustness, for instance, explore concerns about data quality, classification performance and poor generalisability. Other questions highlight bias and fairness issues that could emerge from limited training data diversity or under-represented patient subgroups. The HITL section looks at the complicated relationship between a clinician's decision-making and AI output; it addresses problems around the ability to explain, understand and be open about AI output that could hurt clinical trust and effectiveness. The questions also cover technical security measures, patient data protection, and the broader societal, legal, and ethical implications of AI deployment in healthcare. The continuous evaluation component addresses the dynamic nature of AI systems and risks associated with data drift and model maintenance.

Table 2. Set of questions shown to the participants in the second half of the interview.

These guiding questions are inspired by those from the NIST AI risk management framework.	
Technical robustness	• How might poor data quality result in poor classification performance? • In what scenarios might the AI system fail to generalise? • What risks are associated with poor classification performance?
Uncertainty quantification	• What risks are associated with classifications whose uncertainties are not quantified effectively?
Bias and fairness	• Is there a risk related to limited diversity in data training? • Are there specific patient groups or sub-groups that could be harmed by a biased AI system?
HITL	• What are the challenges for a human user (clinician) in making decisions based on the output of the AI system? • What risks arise in the case of lack of explainability, interpretability or transparency? • What risks are associated with these challenges?
Security	• How can the security of this AI system be ensured?

Risk in wider context	• What are the societal, legal, and ethical risks in the use of such technology (for either a human, an organisation or an ecosystem)? • How accountability helps in addressing AI risks?
Continuous evaluation	• Is there a potential for data drifting, how will this affect the model and how often this should be checked? • What risks are associated with errors not being detected and addressed in a timely fashion?
Privacy	• Are there patient data privacy concerns associated with this AI system?

5.1.3 Risk identification and management

Based on the interviews conducted and considering the points in 5.1.1 and 5.1.2, potential risks associated with the proposed AI system were collected. Figure 4 is a schematic of the high-level risk categories, which are divided into subcategories; only beneficiary diversity subcategories are shown in the figure as an example. The arrows indicate the ways that the product can influence different aspects of its environment, posing potential risk.

Beneficiary diversity risks are related to the features of groups and individuals that cause differences in detecting AF signals or mislabelling non-AF signals.

- Different skin tones interact differently with the greenlight wavelength used in PPG sensors, with darker skin causing less reflection. Thus, a lower-quality signal reduces the accuracy of the ML model and causes bias against patients with darker skin [34].
- Medical history, such as having other heart problems or using drugs that alter heartbeats, might mask part of the signal that is important for the classifier.
- Age is another factor in which bias can result in two ways. Firstly, most of the available datasets are from intensive care units (most positive cases are elderly patients who may experience complications resulting from medication interactions and comorbidities) or smartwatches (mostly young healthy adults), and so an imbalanced representation of the population could potentially lead to a biased model. Secondly, clinical practitioners consider the detection of AF in a younger patient to be less serious and so tend to be more forgiving in diagnosing them – a nuance that AI systems may not take into account.
- Some patient behaviours might be detected using additional sensors, for example, strong movement, such as the one associated with an active lifestyle or using an electric toothbrush. An accelerometer can exclude this type of data, improving data quality that goes into the model.
- Finally, other factors, such as biological sex, whose effects may not be known, might also impact the PPG signal, and it is therefore important to investigate bias across different subgroups throughout the life cycle.

Environmental factors include conditions that affect the signal directly or indirectly, such as changes in the signal due to seasonal change, humidity level or sensor malfunctioning. Detecting this type of error is challenging, as it often occurs coincidentally or after a thorough root cause investigation. A regular system performance check might help in detection. Other errors may be caused by the user unintentionally, such as by hardware limitation and device misuse, or intentionally, such as by an adversarial attack to increase their prioritisation score to be tested on ECG sooner.

Practitioners might resist adopting a new technology if they are not fully aware of its mechanisms, capabilities, and limitations. The black-box nature of deep learning models might cause ambiguity and reduce practitioners' confidence in using the AI system.

Understanding of the system by both practitioners and patients reduces the risk of panicking and improves patient compliance, and ultimately risk to the organisation's reputation.

Some risks are due to the nature of the data or the model. Data for training are usually collected from intensive care units or from healthy people wearing smart watches. Unrepresentative data may impact the model's prediction, necessitating evaluation and management. On the other hand, the requirement of higher explainability, as indicated earlier, may require the use of less accurate models, which reveals an important trade-off.

The sensitivity of health data is relevant for any medical device. Connected devices may expose patients to security threats and unintended use of their data. On the other hand, offline devices restrict the health care provider's capacity to take necessary action prior to the next clinical visit.

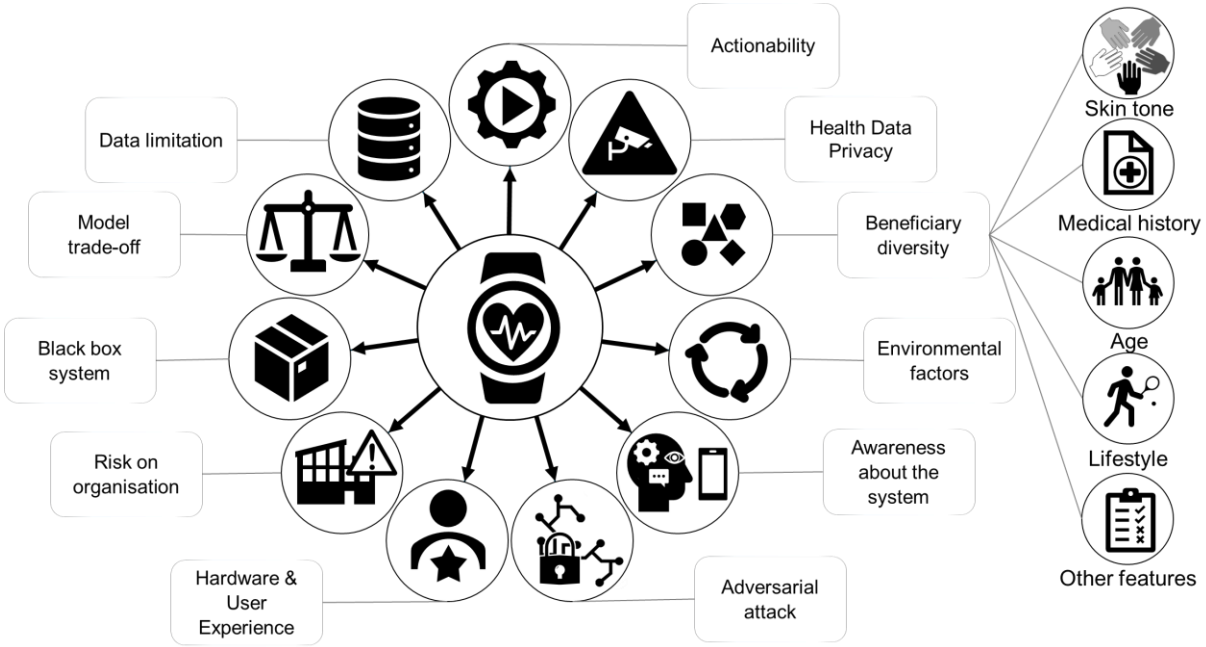

Figure 4. A schematic of the high-level risk categories identified by the risk analysis, with beneficiary diversity subcategories included as an example.

All the risks are ranked based on their likelihood and severity (Appendix table). For each risk the severity and likelihood were evaluated, and the methodology advocated in the NIST AI RMF [16] was used for mitigation. Figure 5 shows an example how the specific risk of patient social anxiety was addressed in each of NIST AI RMF functions.

The following is an example of how the NIST AI RMF approach was used to address the risk of social anxiety for patients due to wrong prediction and misalignment with the diagnosis.

For the *Govern* function, regulatory guidance was identified, namely, the Good Machine Learning Practice (GMLP) Guiding Principles [25] and the Transparency for ML-Enabled Medical Devices guidelines [27]. Additionally, NPL values [36] emphasise the need to prioritise clarity. NPL policies such as the NPL code of conduct and quality policy also address the transparency aspect. These regulatory frameworks and NPL policies and values provide the foundational requirements for the system.

The *Map* function cuts across the TSALC activities, starting with identifying the purpose and objectives of the AI system and mapping the various product requirements to the clinical context. Due to its dynamic nature, this function cannot operate in isolation from other functions. Moreover, it narrows down the context with progress through the life cycle, which

defines the outputs of other functions in more detail. For example, initially we identified that explainability is required in the measure function for patient anxiety risk. In the development stage and revisiting the risk assessment, the ML model will be more defined, which can narrow down the possible explainable AI (XAI) that can be used. As the map function defines the context, it works as a basis for both the manage and measure functions and facilitates communication to prevent the risk.

In the *Measure* function, explainability approaches, such as Shapelets XAI and counterfactual explanations, can quantify and track the system's decision-making patterns. These techniques provide better transparency, which reduces social anxiety.

In the *Manage* function, we identified the relevant internal procedures, such as the NPL Software Quality Management System and project and risk management-related procedures. These procedures encompass the frequency of risk assessment revisits, the regular monitoring of development progress, and the development of other non-AI components. Most of these procedures are general and applied to non-AI projects. We identify additional company- and product-specific procedures and plans, such as regular surveys to assess the clarity of the generated report.

In Figure 5, two contours were added around the output of the functions; the intent is to help the reader to mimic the same process for different AI systems and risks. The four functions should guide the stakeholders from high-level goals, such as patient safety and health, to more specific objectives within the company's context, product details, and ultimately to risk events. Intentionally, the boundaries of these focus areas align with the four functions. For instance, the governance function can contain company policies, and the company can use general procedures to manage product development. Other procedures might be developed specifically for this product or product type. The approach emphasises using existing processes where appropriate and adding new ones based on need.

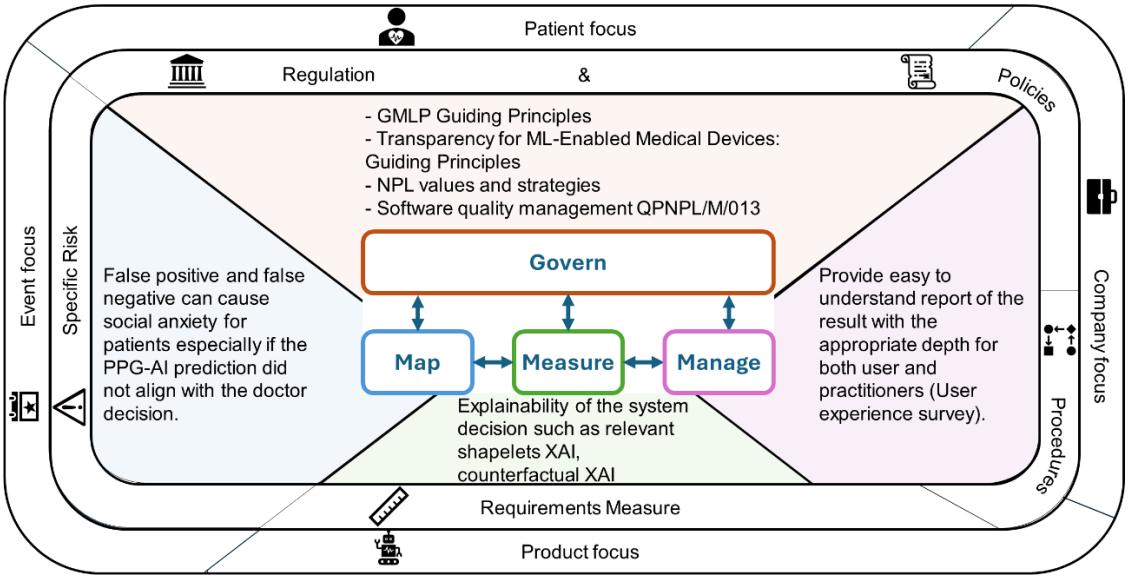

Figure 5. An example of NIST AI RMF functions mapped to a specific risk: social anxiety.

5.2 DESIGN AND DEVELOPMENT

The output of Section 5.1 is a risk register which defines the risks with their severity and likelihood, and in addition the appropriate mitigation and monitoring plan according to the NIST AI RMF. These mitigation and monitoring plans represent initial guidance for the system's design and development. Figure 6 shows the most relevant AI trustworthiness principles as well as a breakdown of the safety and robustness principles for the initial

requirements of the AI system. Note that a given mitigation may address a single risk or it may address multiple risks.

Risks around safety and robustness can be mitigated through technical aspects of model development which improve the performance of the model. A low-performance model might cause incorrect prioritisation for individuals and bias toward certain demographic groups (beneficiary diversity risk), and it is therefore crucial to maintain and verify performance and robustness across all subgroups identified in the risk analysis.

Some examples of such mitigations are:

- Fine-tuning of model training to achieve the most appropriate balance between specificity and sensitivity. This requires an assessment of how to appropriately balance false negative (FN) or false positive (FP) errors.
- The use of optimisation strategies which directly enforce notions of fairness (see Section 5.3.2).
- Employing additional sensors in order to eliminate misleading signals.
- The use of uncertainty quantification to identify low decision confidence and subsequently improve the model.
- The possible extension of the classification model architecture to include detection of additional heart conditions, in order to mitigate the risk of mistaking other heart conditions for AF.
- Ensuring that the training set is as representative as possible of different subgroups.
- The use of techniques for dealing with imbalanced datasets.
- Investigation of an alternative approach in which a regression model is used to directly predict a patient's priority score.

Some mitigations influence other non-AI components of the system. To improve system security from some types of adversarial attack, an accelerometer might be fitted to identify induced strong movement by the user. To reduce access to patient information, hybrid connectivity and encryption might be used.

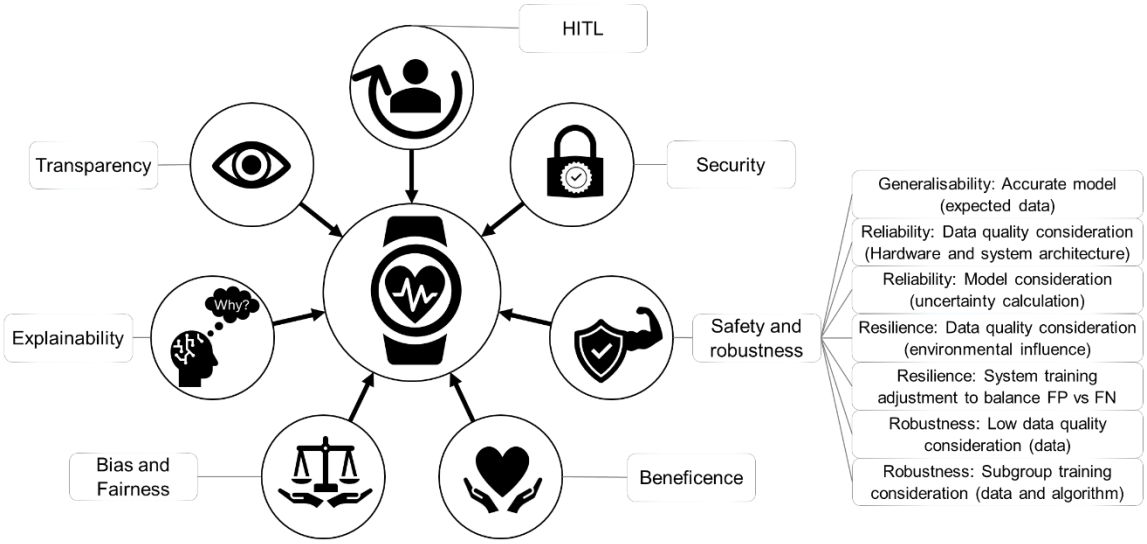

Figure 6. A schematic of the most relevant trustworthiness principles for the proposed AI system. Each principle can be divided into initial system requirements; safety and robustness are here broken down into initial requirements as an example.

The system decision must be explainable at different levels and at different levels of detail, allowing both practitioners and patients to access enough information to use the system correctly and as intended. This information should be provided at the correct time and

communicated using an appropriate method. For example, if the system detected a non-AF high-risk signal or significant high prioritisation level, the system should be able to provide an appropriate alert to the patient advising them to visit accident and emergency. Allowing humans to take control and act on information, as well as designing transparency and explainability from a human perspective, aligns with the HITL principle. One important aspect of transparency is the quantification of uncertainty, so that practitioners and patients have an appropriate understanding of the confidence they should have in the output of the system.

The system should benefit the patient and the organisation without adding complexity to the current procedures. Hence, evaluation of the system efficiency and comparison with current procedures should be done regularly. These design requirements are prioritised based on the associated risk likelihood and severity, focussing first on the most important considerations and managing trade-offs based on the higher priorities. While these represent initial requirements, the process of risk assessment should be iterative and conducted frequently across the life cycle; hence, it provides new evaluations of the risks and requirements and helps provide more detailed context to define the appropriate metric.

The appropriate identification of risk and its mitigation (Section 5.1.3; Figures 4 and 5) provides the context for choosing appropriate metrics for evaluation, and we turn in the next section to address the question of metrics.

5.3 METRICS FOR AI TRUSTWORTHINESS

In this section, we present metrics for assessing the trustworthiness of our hypothetical AI-based product. We choose to focus on selected aspects of trustworthiness out of those highlighted in the risk assessment in Section 5.1, namely *uncertainty quantification* and *fairness*. It is important to point out that the perspective of the NPL Trustworthy and Safe AI Life Cycle is that these metrics should be used not just for final evaluation but as an intrinsic part of the system development process (see Section 5.2).

5.3.1 Uncertainty quantification

The importance of uncertainty quantification was highlighted as part of the risk assessment in Section 5.1. It is vital that clinicians are provided with transparent information about the prediction of the AI system, in this case a report on the extent to which episodes of AF have been detected in a given patient. Transparency here includes not only a prediction (AF absent or present), but also a statement concerning the confidence in the prediction. Automatic PPG-based AF detection typically involves breaking down a patient's signal into segments (for example of 30 s each) and determining whether or not AF is present in each segment. A natural way to express confidence in the classification is by returning a probability that each segment exhibits AF. It follows that a trustworthy classifier for this task should be probabilistic. An AI-based approach is reliant upon an ML classifier, and many ML classifiers are probabilistic, or can be adapted to give probabilistic output. Further data analysis can then be performed to return probabilistic summary statistics, for example the probability that AF is observed *at some time* during the measurement period, or confidence intervals on the frequency of AF episodes.

A metric in this context should measure the trustworthiness of the uncertainty quantification, which at the level of the ML classifier means the reliability of the probabilities of AF assigned to each of the segments. Various tools have been developed in the ML community for validating the output of probabilistic ML classifiers (see for example [37]) and we next give a brief overview.

Proper scoring metrics. A proper scoring metric is one which is minimised when the outputted probability distribution corresponds to the true underlying distribution. Examples of

proper scoring rules are the *categorical cross-entropy* and the *Brier score* [38]. Both metrics are widely used in the ML community. In particular, categorical cross-entropy is often used as a *loss function* when training a deep neural network, which means that the criterion for optimising the weights for the network is minimising the categorical cross-entropy.

Calibration metrics. These metrics compare the probabilities returned by an ML classifier with observed proportions, which requires additional labelled data that has not been used to train the ML classifier in the first place. Any comparison with observed proportions can only be carried out by grouping the predictions in some way, and this is typically done by grouping together predictions with similar prediction probabilities. Some popular calibration metrics include Expected Calibration Error (ECE), Smooth Expected Calibration Error (smECE), Adaptive Calibration Error (ACE), Uncertainty Calibration Error (UCE) and Expected Cumulative Calibration Errors (ECCE); see [39] for a review of state-of-the-art.

Sharpness. It has been shown that any proper scoring metric can be decomposed into two components, one which captures calibration and one which captures sharpness [40]. By the sharpness of classifier, we mean the extent to which the classifier is confident about its predictions, so for a binary classification task such as AF detection a sharp classifier is one where the probability of AF is typically close to either 1 or 0. Intuitively, a classifier should be both well-calibrated and sharp. A sharp classifier whose probability estimates are over-confident and a well-calibrated classifier that has little classification ability are both to be avoid. It therefore represents good practice to calculate sharpness metrics alongside calibration metrics. Sharpness can be quantified by any of the popular metrics for assessing the predictive performance of an ML classifier, including accuracy, precision, recall, F_1 score, Matthew's correlation coefficient and receiver operator characteristic (ROC) area under the curve (AUC).

The metrics mentioned above assess uncertainty quantification at the signal segment level and further data analysis is required to validate any probabilistic summary statistics derived from the segment-level classifications. We are not aware of specific approaches that have been proposed in this regard, and so this represents an area where further work is needed.

5.3.2 Fairness

A core aspect of trustworthiness for an AI-based AF detection system is that it is robust to the range of scenarios it is expected to encounter. This leads to a crucial requirement on the ML classification models on which the AF detection system is based, namely that they generalise to all PPG signals that might need to be analysed. This aspect of trustworthiness was highlighted in the risk assessment in Section 5.1.3.

Fairness is another aspect of trustworthiness which is related to robustness. The NIST AI RMF [16] points to two aspects of fairness: harmful bias and discrimination. It is the risk of harmful bias which was specifically highlighted in Section 5.1.3. In this case, this takes the form of *beneficiary diversity risks*, or in other words bias against certain characteristics or demographic groups. Some of the characteristics that were identified in Section 5.1.3 were skin tone, medical history, age, lifestyle and biological sex.

A major cause of biased models is biased training sets. For example, if a model were only to be trained on male patients, it would be no surprise if the model performed poorly on female patients. Taking the example of skin tone, an independent review on Equity in Medical Devices recommended that the MHRA and approved bodies for medical devices should strengthen standards for approval of new pulse oximeter devices (used to obtain PPG signals) to include sufficient clinical data to demonstrate accuracy in groups with darker skin tones [34].

Bias can also be mitigated by appropriate technical ML model development methods such as techniques for dealing with imbalanced datasets and optimisation strategies which directly enforce certain technical definitions of fairness. See the review articles [41], [42] for further details.

What metrics can be used to assess fairness/bias with respect to a particular characteristic? One natural approach is to split the test set according to the characteristic in question and examine whether there are differences in performance across the different subsets.

Such an investigation was carried out in [43] in the context of high blood pressure classification using PPG signals from the Aurora BP dataset [44]. In this dataset, skin tone was categorised according to the six-point Fitzpatrick skin tone scale [45], and cross-validation classification accuracy was compared across each of the categories. The results are shown in Figure 7, reproduced from [43] with permission from the author. The bars in the figure show cross validation accuracy when the model is trained using the full training set. This training set included patients of all skin tones, though it was not balanced across the skin tones due to limited availability of data for darker skin tones. The dashed lines show cross validation accuracy when the model is trained only on data from Fitzpatrick skin tone class 1. We observe variations in performance across skin tones, and as expected we observe that performance improves for skin tone classes 2 to 6 when trained using all skin tones classes.

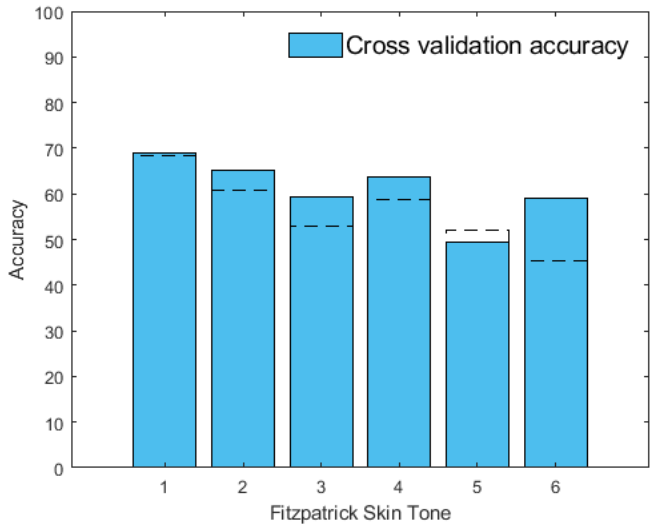

Figure 7. The accuracy for each skin tone class using cross validation for a model trained on a dataset containing all skin tone classes. The accuracies for a model trained only on data from skin tone 1 are shown as dashed lines.

6 CONCLUSIONS

We have presented a case study for NPL's Trustworthy and Safe AI Life Cycle (TSALC). In one direction, we believe that applying the TSALC to AF detection provides insight into some of the challenges and the most promising approaches for the use of AI-based PPG signal analysis in AF detection. In the other direction, we expect the methodology followed in this case study to be applicable to other AI application areas. We close by highlighting some of the lessons learned about how to apply the TSALC in context.

- **The importance of clearly defining the product and its purpose.** The risks associated with PPG-based AF detection depend crucially upon the precise manner in which the technology is to be used. For example, the purpose of the product

proposed and analysed in this report is to support the prioritisation of diagnosis resources, and not to actually carry out diagnosis. Some of the most severe risks, such as the risk of misdiagnosis and its consequences, are taken out of the equation by defining the remit of the product more conservatively.

- **The importance of a flexible approach for trustworthy AI lifecycles.** TSALC's iterative nature allows its adoptability for research-based low-risk AI systems and applied-based high-risk AI systems, as well as the transition between them. This feature reflects the real-world scenario where numerous AI systems began as research projects before gaining practical application. Research-based systems are often associated with lower risks, a smaller scope, and fewer stakeholders. The iterative nature means revisiting the risk analysis and updating the requirements as the system is scaled up or moved closer to real users. Moreover, if the AI product is sold to another company during this transition process, the whole output of TSALC can be sold with it, maintaining traceable product development.
- **The importance of collaboration between AI practitioners and domain experts to identify and manage domain-specific risks.** Both groups are needed, and they need to work together. Without domain experts, AI systems which are trustworthy in a generic sense will be developed, and the domain-specific requirements will not be taken into account. Without AI practitioners, risk analysis will not filter down into concrete strategies for model development and evaluation and will remain at the level of high-level concepts.
- **The importance of combining a consideration of AI-related risks with more generic software quality and risk management policies.** AI systems are associated with particular risks which is why frameworks such as the TSALC and the NIST AI RMF are needed. On the other hand, many risks associated with AI systems are of a more generic nature, and existing standards, regulations and company procedures for project management and software quality should be used and extended. For example, a company with knowledge of medical applications is familiar with the regulation and has established patient-focused policies which are important for developing an AI system in medical applications. The company should leverage its expertise in medical applications to enhance its capabilities in software and AI development. Conversely, a software company would face the opposite challenge. This lowers the barriers for application- and technology-focused companies to enter the market and contribute to the innovation cycles. In Figure 5, we also demonstrated that applications with similar risk profiles often have comparable regulatory, policy, and procedural requirements, with variations occurring at a product- and event-specific level. Therefore, the experience of applying TSALC can be an accumulating process, making future development of AI systems more accessible.
- **The importance of considering risks from the perspective of both the application and the technology.** In Section 5.1, both perspectives were sought when gathering risks. Domain-specific risks emerge from the contexts in which the systems are deployed, and they are closely linked with patient safety, clinical workflows, regulatory compliance and established standards of care. It is also important to identify the relevant technical risks around the use of an AI system, and these risks typically feed into the domain-specific risks. Understanding the interplay between these two types of risk is important.
- **The importance of identifying relevant AI trustworthiness aspects and accompanying metrics before beginning the development of the AI system.** This is a key principle of the TSALC, and we have seen the importance of this principle illustrated in our case study. For example, beneficiary diversity risks must be mitigated both at the point of training data preparation (ensuring for example that different demographic groups are represented) and model training (ensuring for example that appropriate techniques are used for dealing with imbalanced datasets).
- **The importance of determining specific in-context metrics for assessing AI trustworthiness, both in development and deployment.** This is also a key

principle of the TSALC whose importance was illustrated in this case study. For example, the concept of *fairness* of an AI system can mean different things in different contexts. In some contexts, fairness refers to the imperative to ensure that models deal impartially with 'protected characteristics'. In the context of this case study, fairness more refers to the imperative to ensure that models generalise across different demographic groups. Breaking down umbrella concepts into sub-categories is important so that the most relevant specific metrics can be identified.

There is no such thing as a one-size-fits-all TSALC, but on the other hand it would be inefficient if trustworthy and safe AI life cycles are all developed in isolation without reference to a generic methodology. We believe that there is great value in the in-context application of generic trustworthy AI methodologies, and NPL hope to investigate other applications of the TSALC in the future.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the UK Department for Science, Innovation and Technology (DSIT), which funded this work as part of the AI Standards Hub programme.

The authors would also like to thank the following people (listed alphabetically by surname along with affiliations at the time the discussion took place) who provided input into both the shaping of the case study and the risk analysis.

- Tudor Besleaga (Sector Heath)
- Peter Charlton (University of Cambridge)
- Thomas de Cooman (Fibricheck)
- Glenn de Witte (Fibricheck)
- Peter Doggart (PulseAI and Ulster University)
- Albert Ferro (Guys and St Thomas' NHS Foundation Trust and Kings College London)
- Christian Heiss (Surrey and Sussex Healthcare NHS Trust and University of Surrey)
- Vaidotas Marozas (Kaunas University of Technology)
- Manasi Nandi (Kings College London)

This report, and especially the section on identification of trustworthiness metrics in the context of PPG signal analysis (Section 5.3), builds in part upon outputs of a project funded by the European Partnership on Metrology (22HLT01 QUMPHY) [46], in which NPL is a partner.

Finally, the authors would like to thank Indhu George, Louise Wright, Padmini Krishnadas and Philip Aston (all NPL) for helpful comments.

BIBLIOGRAPHY

[1] M. Levene *et al.*, 'A Life Cycle for Trustworthy and Safe Artificial Intelligence Systems'. Accessed: Mar. 05, 2025. [Online]. Available: https://doi.org/10.47120/npl.MS57
[2] 'Number of UK people with heart rhythm condition rises by 50% in a decade'. Accessed: Mar. 03, 2025. [Online]. Available: https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2023/may/new-figures-show-the-number-of-uk-people-with-heart-rhythm-condition
[3] D. Linz *et al.*, 'Atrial fibrillation: epidemiology, screening and digital health', *Lancet Reg. Health – Eur.*, vol. 37, Feb. 2024, doi: 10.1016/j.lanepe.2023.100786.
[4] Stroke Association, 'AF: How can we do better?' Accessed: Mar. 03, 2025. [Online]. Available: https://www.stroke.org.uk/sites/default/files/af-data_2018_england_eng_1.pdf

[5] Marc Holland, ‘Blood Thinners: Brain Bleeds And Strokes’. Accessed: Mar. 27, 2025. [Online]. Available: https://medshun.com/article/can-blood-thinners-cause-strokes-and-bleeding-in-the-brain
[6] NICE guideline, *Atrial fibrillation: diagnosis and management*, NG 196, Jun. 30, 2021. Accessed: Mar. 06, 2025. [Online]. Available: https://www.nice.org.uk/guidance/ng196/chapter/Recommendations#detection-and-diagnosis
[7] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, ‘A review on wearable photoplethysmography sensors and their potential future applications in health care’, *Int. J. Biosens. Bioelectron.*, vol. 4, no. 4, pp. 195–202, 2018, doi: 10.15406/ijbsbe.2018.04.00125.
[8] T. Pereira *et al.*, ‘Photoplethysmography based atrial fibrillation detection: a review’, *NPJ Digit. Med.*, vol. 3, p. 3, Jan. 2020, doi: 10.1038/s41746-019-0207-9.
[9] E. Y. Ding, G. M. Marcus, and D. D. McManus, ‘Emerging Technologies for Identifying Atrial Fibrillation’, *Circ. Res.*, vol. 127, no. 1, pp. 128–142, Jun. 2020, doi: 10.1161/CIRCRESAHA.119.316342.
[10] M. V. Perez *et al.*, ‘Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation’, *N. Engl. J. Med.*, vol. 381, no. 20, pp. 1909–1917, Nov. 2019, doi: 10.1056/NEJMoa1901183.
[11] Y. Guo *et al.*, ‘Photoplethysmography-Based Machine Learning Approaches for Atrial Fibrillation Prediction’, *JACC Asia*, vol. 1, no. 3, pp. 399–408, Dec. 2021, doi: 10.1016/j.jacasi.2021.09.004.
[12] F. Wouters *et al.*, ‘Comparative Evaluation of Consumer Wearable Devices for Atrial Fibrillation Detection: Validation Study’, *JMIR Form. Res.*, vol. 9, Jan. 2025, doi: 10.2196/65139.
[13] R. Avram *et al.*, ‘Real-world heart rate norms in the Health eHeart study’, *Npj Digit. Med.*, vol. 2, no. 1, pp. 1–10, Jun. 2019, doi: 10.1038/s41746-019-0134-9.
[14] J. A. Joglar *et al.*, ‘2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines’, *Circulation*, vol. 149, no. 1, Nov. 2023, doi: https:/doi/10.1161/CIR.0000000000001193.
[15] ‘TickITplus’. Accessed: Mar. 18, 2025. [Online]. Available: https://www.tickitplus.org/en/home.html
[16] E. Tabassi, ‘Artificial Intelligence Risk Management Framework (AI RMF 1.0)’, *NIST*, Jan. 2023, Accessed: Mar. 05, 2025. [Online]. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10
[17] Centre for Data Ethics and Innovation (CDEI), ‘The roadmap to an effective AI assurance ecosystem’. Accessed: Mar. 05, 2025. [Online]. Available: https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem
[18] D. Elgesem, ‘The AI Act and the risks posed by generative AI models’, in *Proceedings of the 5th Symposium if the Norwegian AI Society (NAIS 2023)*, Accessed: Mar. 05, 2025. [Online]. Available: https://ceur-ws.org/Vol-3431/paper3.pdf
[19] ‘ISO/IEC 23894:2023’. Accessed: Mar. 05, 2025. [Online]. Available: https://www.iso.org/standard/77304.html
[20] ‘ISO/IEC 42001:2023’. Accessed: Mar. 05, 2025. [Online]. Available: https://www.iso.org/standard/81230.html
[21] OECD, ‘Tools for trustworthy AI’. OECD Digital Economy Papers, Jun. 28, 2021. Accessed: Mar. 05, 2025. [Online]. Available: https://www.oecd.org/en/publications/tools-for-trustworthy-ai_008232ec-en.html
[22] ‘ISO 31000:2018’. Accessed: Mar. 06, 2025. [Online]. Available: https://www.iso.org/standard/65694.html
[23] ‘ISO 14971:2019’. Accessed: Mar. 06, 2025. [Online]. Available: https://www.iso.org/standard/72704.html

[24] H. Strassburg, H. Whelan, M. Alsuleman, M. Chrubasik, P. Duncan, and J. Gregório, 'An integrated framework for risk assessment in digital verification and validation', *Meas. Sens.*, p. 101496, Dec. 2024, doi: 10.1016/j.measen.2024.101496.
[25] International Medical Device Regulators Forum (IMDRF), 'Good machine learning practice for medical device development: Guiding principles'. Jan. 29, 2025. Accessed: Mar. 06, 2025. [Online]. Available: https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles
[26] International Medical Device Regulators Forum (IMDRF), *Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles*, Mar. 12, 2024. Accessed: Mar. 17, 2025. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles
[27] International Medical Device Regulators Forum (IMDRF), *Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles*, Jun. 13, 2024. Accessed: Mar. 17, 2025. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles
[28] International Medical Device Regulators Forum (IMDRF), *Clinical Decision Support Software - Frequently Asked Questions (FAQs)*, Aug. 01, 2025. Accessed: Mar. 17, 2025. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs
[29] FDA, 'Software as a Medical Device (SaMD)'. Accessed: Mar. 06, 2025. [Online]. Available: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd
[30] FDA, 'Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP are Working Together'. Accessed: Mar. 17, 2025. [Online]. Available: https://www.fda.gov/media/177030/download?attachment
[31] FDA, 'Postmarket Management of Cybersecurity in Medical Devices'. Accessed: Mar. 17, 2025. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postmarket-management-cybersecurity-medical-devices
[32] CDC, 'Health Insurance Portability and Accountability Act of 1996 (HIPAA)', Public Health Law. Accessed: Mar. 17, 2025. [Online]. Available: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html
[33] 'NIST AI RMF Playbook', *NIST*, Mar. 2023, Accessed: Mar. 17, 2025. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook
[34] 'Equity in medical devices: independent review'. Accessed: Mar. 19, 2025. [Online]. Available: https://www.gov.uk/government/groups/equity-in-medical-devices-independent-review
[35] J. Heijman, J.-B. Guichard, D. Dobrev, and S. Nattel, 'Translational Challenges in Atrial Fibrillation', *Circ. Res.*, vol. 122, no. 5, pp. 752–773, Mar. 2018, doi: 10.1161/CIRCRESAHA.117.311081.
[36] 'NPL Values', NPL website. Accessed: Mar. 18, 2025. [Online]. Available: https://www.npl.co.uk/careers/values
[37] 'Uncertainty Toolbox'. Accessed: Mar. 27, 2025. [Online]. Available: https://uncertainty-toolbox.github.io/
[38] B. Lakshminarayanan, A. Pritzel, and C. Blundell, 'Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles', in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 2017, pp. 6405–6416.
[39] C. Wang, 'Calibration in Deep Learning: A Survey of the State-of-the-Art', May 10, 2024, *arXiv*: arXiv:2308.01222. doi: 10.48550/arXiv.2308.01222.
[40] T. Gneiting, F. Balabdaoui, and A. E. Raftery, 'Probabilistic Forecasts, Calibration and Sharpness', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 69, no. 2, pp. 243–268, Apr. 2007, doi: 10.1111/j.1467-9868.2007.00587.x.

[41] T. P. Pagano *et al.*, ‘Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods’, *Big Data Cogn. Comput.*, vol. 7, no. 1, Art. no. 1, Mar. 2023, doi: 10.3390/bdcc7010015.
[42] M. Shah and N. Sureja, ‘A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions’, *Arch. Comput. Methods Eng.*, vol. 32, no. 1, pp. 255–267, Jan. 2025, doi: 10.1007/s11831-024-10134-2.
[43] P. Aston, ‘Does Skin Tone Affect Machine Learning Classification Accuracy Applied to Photoplethysmography Signals?’ Accessed: Mar. 19, 2025. [Online]. Available: https://www.cinc.org/archives/2024/pdf/CinC2024-038.pdf
[44] R. Mieloszyk *et al.*, ‘A Comparison of Wearable Tonometry, Photoplethysmography, and Electrocardiography for Cuffless Measurement of Blood Pressure in an Ambulatory Setting’, *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 2864–2875, Jul. 2022, doi: 10.1109/JBHI.2022.3153259.
[45] T. B. Fitzpatrick, ‘The Validity and Practicality of Sun-Reactive Skin Types I Through VI’, *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, Jun. 1988, doi: 10.1001/archderm.1988.01670060015008.
[46] ‘QUMPHY’. Accessed: Mar. 27, 2025. [Online]. Available: https://www.qumphy.ptb.de/home

APPENDIX

This is a rough draft of the detailed risk assessment provided for guidance:

ID	Category	Risk	Cause	The affected or impacted	Severity and consequences	Likelihood of the risk	Matrix [1-6] [7-12] [13-25]	Monitoring	Mitigation	Mitigation summary
1	Actionability	(Intended Use) The recorded and analysed data may not be immediately actionable.	Device usage: - There may be a time delay between when an event is detected and when a follow-up test is reviewed (sometimes monitoring time might be too long for some patients in dire situations).	There is a delay in treatment or testing.	2	4	8	An additional framework or algorithm (which can be rules or symbolic AI) could flag patients with a high burden of detected AFib for moving to ECG testing soon.	HITL: - The system can ask the patient to visit the doctor sooner (increase human involvement when needed, HITL). - Even if the PPG monitoring period is long, regularly identify and highlight higher-risk patients for healthcare providers.	HITL: indication for a human to take control
2	Actionability	(Clinical Action) What is the action point?	How many positives/burdens are actionable?	Monitoring someone for 30 days is likely to generate at least some FPs. Therefore, the threshold cannot be "any instance of atrial fibrillation".	3	3	9	Generalisability (accuracy) can be used during the ECG recording to provide feedback on the PPG performance.	Resilience, robustness, and beneficence: Test different models—for example, a regression model that gives the probability immediately vs another model that determines the probability based on the duration and frequency of the episode. Age and other factors can be included in the calculation to give different thresholds for different groups (young is less likely).	Beneficence: Complexity-Aware ML Models
3	Actionability	The actionability of the high-risk cases is crucial. Priority should depend on more than just the frequency and duration of AF signals. We might need to consider other clinical risk factors.	The model does not include additional data, such as age or medical symptoms, in its training or final prioritisation calculations.	These conditions can lead to incorrect priorities and delays in high-risk cases.	5	3	15	The pilot study compared the two methodologies. The calculation determines the robustness of both models.	Classification problem: The goal is to define rules for calculating priorities based on the output of the ML model and other clinical data. Regression problem: The goal is to directly incorporate this data into the ML model, with the model's output serving as the final priority level.	Beneficence: Complexity-Aware ML Models

4	Actionability	Irregular heartbeats might be related to more serious illnesses; thus, dealing with these signals as AF ones might cause missing and delaying treating these cases.	Similar symptoms: - Some serious illnesses could show very similar symptoms to afib detected by the device; for example, a stroke could produce similar signals.	Patients may not receive the appropriate treatment in a timely manner, as the diagnosis may be biassed towards afib-related cases.	4	4	16	The system should undergo a regular check to ensure its robustness.	Beneficence: It might be useful to increase the number of classes in the classifier to include other causes of irregular heartbeats. Alternatively, it is possible to develop another model that detects the other cases (including serious cases and illnesses). Treat: A live connection to the medical organisation (NHS) in serious cases and emergencies might inform the NHS to intervene.	Beneficence: Complexity-Aware ML Models
5	Adversarial attack	Fake measurements	Patients are attempting to induce AF detection to gain a higher priority ranking in the list and expedite the ECG process. This goal might be achieved by external factors such as movement, tooth brushing, etc.	Organisation (NHS) This risk has the potential to result in financial and resource waste due to inaccurate information, which can prolong the process and increase the chances of mistreatment. Patients: It also might cause mistreatment.	3	2	6	Validation of the process is required. Using a system to track the number of both patients who are detected late and prioritised healthy people.	We can use other sensors, such as accelerometers, to remove low-quality data and signals caused by movement. A possible solution is to give more importance to night data, where patients do less movement and record higher data quality. The algorithm might consider episodes that are long enough to cause a stroke, minimising the use of random, irregular signals that are caused by movements or other activities. It is also helpful to provide uncertainty in the prediction answer.	Beneficence: System efficiency
6	Awareness about the system's limitations and capabilities	Patients might feel that there is ambiguity in the process of the entire system.	The system's priority rules are unclear, potentially resulting in orders that could cause trouble for certain types of patients.	The wrong priorities might lead to delays or missed checks for some patients.	3	3	9	The system should ensure a certain level of AF was detected efficiently and that prioritisation is directly linked with patients at higher risk.	HITL and Transparency: Practitioners should be involved in setting the rules, such as the weight of the frequency or episode length for the patient's final priority.	Transparency: Patient knowledge HITL: System building with the collaboration of practitioners
7	Awareness about the system's limitations and capabilities	The system could cause anxiety for the patients.	Misinterpretation and the overwhelming amount of information given to the patient	The patient's anxiety over the illness is causing panic.	3	3	9	Survey to test patient interaction with both methods.	Transparency: It is important to provide information at the correct time and with enough details. if the patient is provided with the results in real time or at the doctor's appointment. Real-time might provide an opportunity to ask the patient to stand still to increase the collected data quality. While providing the results at the clinic, reduce the patient's panic. Educate the patient about AF, the device, and the process. In a way, the patient does not panic.	Transparency: Patient knowledge
8	Awareness about the system's limitations and capabilities	Patients' compliance issue	Patients don't wear the device or perform experiments with	Patients can lose interest in wearing the device; this could result in no data or weird records being	3	3	9	Surveys to check patient compliance. Furthermore, analyses of touch	Transparency (the device is working): Show the patient any information, like time or heart rate, that can let them know the device is functioning.	Transparency: Patient knowledge

			it; they shake it to see if it is working.	registered from their experiments.				sensor data can be useful.		
9	Awareness about the system's limitations and capabilities	Patients' compliance issue	A long record might cause the patient to lose interest in wearing the device. Investigating for AF in both PPG and ECG recordings means that the patient is being tested for a long period of time, so they might get bored.	Patients might cease to wear the device, fail to return to the doctor for a check-up, and end up losing the NHS device and resources.	3	3	9	Compliance survey	Transparency: We should communicate the importance of the test to the patient and provide multiple choices of watch straps so it is comfortable to wear.	Transparency: Patient knowledge
10	Beneficiary diversity (age)	Performance issues, like accuracy and biases	There is a bias in the training data (under-represented) towards certain subgroups, such as young patients, women, active people, and patients on certain drugs.	Lower accuracy with some subgroups.	3	4 (patients on certain drugs, such as beta blockers; this is common for monitoring diagnosed patients)	12	There should be frequent measuring of bias in the subgroup during training, testing, implementation, and use. Explainability and generalisability are measured to evaluate the reason and the output.	We can include data that represent this subgroup in the training set. Labelling the data explicitly and using it in training and prediction might help improve accuracy. We should evaluate this rather than allowing the model to independently identify the different behaviours within the subgroup.	Bais: Subgroup training consideration (data and algorithm)
11	Beneficiary diversity (age)	(False positive) bias	Bias dataset quality: - If collected from smartwatches, the dataset gathered is likely from a younger demographic. - If collected from clinics, then the population is elderly dominant.	Patient, healthcare professional: - Training the device on a biassed dataset reduces its accuracy.	3	4	12	To monitor this risk, we can measure overall accuracy over time and accuracy for subgroups.	To treat the risk, we should: - improve robustness - Use a dataset that includes a greater variety of features (age, race, sex, lifestyle, etc.).	Robustness: Subgroup training consideration (data)

12	Beneficiary diversity (general)	Performance issue	Normally, the distribution of errors is not uniform, leading to low accuracy for specific groups of features. Different skin colours and hydration are relevant.	There is a risk of low priority for the subgroup, which could cause delays or missing the diagnosis of these patients from these groups. False positives can be due to features such as lifestyle and being active.	3 (severity is naturally lowered, as the result is not a wrong diagnosis but prioritisation of the patient).	4	12	The accuracy of the model should be regularly measured.	Robustness and reliability: Training the model should be done with the presence of these types of errors with good generalisability to ignore these errors. The source of error should be regularly reviewed to include them in the training data. Practitioners should be consulted to design the training dataset. The real-world scenario should be close enough to the training scenarios. For example, if the difference between age groups affects the detection of AF irregular signals, then we include different age groups in the training set. But if other features, such as eye colour, don't affect the detection of the Af signal, then there is no need to include them in the training dataset.	Robustness: Subgroup training consideration (data)
13	Beneficiary diversity (lifestyle)	(Interpretations) Respiratory Sinus Arrhythmia	Very healthy individuals have large respiratory-driven heart rate variability, which will set off Afib alarms at low heart rates.	Patient: Patient can generate a very large FP Afib burden.	3	4	12	Algorithm design must take this into account. Long window lengths allow us to detect the sinusoidal RSA pattern, which aligns with the breathing rate.	-Improving reliability might be possible by considering different scenarios. -RSA-driven HRV declines with age, so this condition could also be managed by patient selection.	Reliability: Data quality consideration
14	Beneficiary diversity (lifestyle)	(Misuse) The patient wears the device during an activity that is contraindicated. This activity results in a patient not receiving proper treatment.	Patient: - Patient wears the device whilst undertaking cardio activity such as jogging.	Patient, healthcare professional: It is much more likely to get FPs during these periods even with other sensors and inaccurate data.	3	3	9	Monitoring for activity levels is possible using other sensors to exclude this data and not use it for Afib screening.	Robustness: Robustness might be improved by including this type of data in the training and checking if the model can learn it.	Reliability: Data quality consideration Transparency: Patient knowledge

15	Beneficiary diversity (medical history)	Accuracy problems: there are two types of patients. 1. Previous patients who are monitoring their situation 2. New patients who are checking the cause of their symptoms.	A simple classification system might not be enough for detection. Patients already diagnosed with AF are more likely to exhibit AF signals, leading to a higher number of false negative cases. While a new case that has not been diagnosed previously might end up with more false positive cases.	An accuracy problem might cause a wrong priority listing.	3 (4 if the number of false predictions is high to disturb the priority list)	4	12 to 16	To monitor, we can measure false positives and false negatives for subgroups.	Including and excluding criteria: We should include in the dataset all potential cases to improve the accurate detection or implement a model for each type (new patient and old patient). We should validate the model results with ECG data for both cases. This tendency to prefer false positives or negatives might be different for new patient vs diagnosed patient. A potentially personalised model might be possible. All these models should be evaluated and compared and accept the one that shows an acceptable residual risk that we can tolerate.	Beneficence: Complexity-Aware ML Models
16	Beneficiary diversity (medical history)	(Interpretations) There might be irregular rhythms which is not AF.	It's difficult to differentiate other rhythms on PPG. For example, VT (high risk) and/or frequent premature atrial contractions (maybe more common than Afib in some groups like the younger generation).	Ultimately, patients with high-burden PACs are at higher risk of Afib, but treatment is less clear. For example, using ECG, even if the doctors detected a few Af in young patients, they won't treat them for AF. But the same appears with an elderly person; the doctors might diagnose them with AF.	3	4	12	To monitor this risk, we should be able to assess the performance of the algorithms on other common arrhythmias.	This is a well-known problem with PPG arrhythmia sensors and even ECG sensors (inconclusive on Apple Watch for PACs). One potential solution could involve incorporating subgroup characteristics, such as age, into the priority level calculation. This will take into account the presence of AF more in elderly people, for example.	Beneficence: Complexity-Aware ML Models
17	Beneficiary diversity (medical history)	(Interpretations) Presence of sleep apnoea which might trigger AF detection.	Patients with known (or unknown) sleep apnoea may cause FP events due to well-known heart rate accelerations and decelerations.	This can generate a very large FP Afib burden.	2	3	6	Accuracy: We can monitor this risk by testing model accuracy using sleep apnoea patients data.	Robustness: Include in the training and testing sets cases with sleep apnoea, and testing the accuracy might treat the problem. Another solution is to exclude patients with known sleep apnoea from screening on PPG and provide alternatives like ECG records directly.	Beneficence: Complexity-Aware ML Models (System Architecture)

18	Beneficiary diversity (medical history)	(Other Devices) Pacemakers and ICDs and other cardiac-modifying drugs (like beta blockers) might alter the heartbeat signal.	Patients and healthcare professionals: - Patients with pacemakers or ICDs may exhibit strange PPG signals. - Patients with beta-blockers or rate control drugs may not exhibit AFib characteristics even during AFib.	Patient, healthcare professional: This may cause false positives, false negatives, or other algorithm issues.	3	3	9	Monitoring this risk can be done by measuring accuracy and stress testing with known scenarios.	Robustness: This risk might be treated by: - using other sensors to exclude bad-quality data, such as that caused by movement or with high noise. - Improving the dataset to see how the model deals with this type of data.	Reliability: Data quality consideration (system architecture)
19	Beneficiary diversity (medical history)	(Other Devices) Wearing other devices might restrict blood flow (long-term blood pressure monitoring).	Patient: - Patients with poor circulation or other devices that restrict blood flow (such as cuffs for monitoring blood pressure overnight) have weak PPG signals. Some patients wear two watches on the same wrist. There is a need to understand how the device would behave in such a situation.	Patient, healthcare professional: This will reduce the device's accuracy and functionality; specifically, it will likely cause poor signals to be captured (high uncertainty), leading to more false positives and false negatives.	2	4	8	Monitoring can be done by collecting patients' records (if they are wearing other devices).	Transparency: - Instructions should include a note and instructions to preclude patients from using other devices of this nature during the monitoring period or remove this device first.	Transparency: Patient knowledge
20	Beneficiary diversity (medical history)	(Uncertainty) Presence of high uncertainty correlates with the detection of AF.	Device, model, software: - Uncertainty measures from the AI model may become associated with disease or sub-groups of disease.	Patient: - This could lead to all periods of Afib being labelled as uncertain and therefore ignored. missing treatments	4	2	8	This risk might be monitored by frequent stress testing (reliability and accuracy)	Treating this risk might be achieved by: - improving reliability. - validating that the amount of “uncertain” results is the same in both diseased and non-diseased patients. We have observed in real-world data that inconclusive results on smartwatches are significantly more common among diseased patients.	Reliability: Model consideration (uncertainty calculation)
21	Beneficial diversity (skin tone)	Performance issues, like accuracy and biases, expose patients or subgroups to risk.	Sensor limitations, such as using green light wavelengths which have less penetration with a higher concentration of melanin, can cause performance issues.	This can cause lower accuracy with some subgroups—people with a higher melanin (skin colour). This bias could potentially result in a lower priority for this particular subgroup.	3	4	12	This risk can be monitored by frequent measuring of bias in the subgroup during training, testing, implementation and use.	This risk can be reduced by using a sensor with better penetration at different wavelengths (https://ieeexplore.ieee.org/document/8234348) and provide uncertainty with the prediction answer.	Bais: Subgroup training consideration (data and algorithm) Reliability: Data quality consideration (hardware)

								Additionally, explainability and generalisability should be measured. This is done to verify the cause and the result.		
22	Beneficial diversity (skin tone)	Bias can occur due to subgroup features like colour of skin, race, gender, etc.	Device: - Detection of signals is affected by skin tones. Dataset: - Data used to train software may be biased towards a certain group of people.	Patients can be affected because of: - inequalities in receiving treatments --darker skinned people may get overlooked more due to biased results.	3	4	12	Monitoring can be done during the test stage: - detecting bias for different subgroups	To treat this risk, we should improve generalisability by including patients with darker skin and use infrared instead of visible light (better penetration in darker skin).	Bais: Subgroup training consideration (data and algorithm) Reliability: Data quality consideration (hardware)
23	Black box nature of the system	Validation of the prioritisation is needed to ensure the system's effectiveness.	Giving more importance to certain features, such as episode length or frequency, and neglecting other important features or feature interactions can lead to wrong prioritisation.	Wrong prioritisation might lead to delay in checking patients with higher risk and reduction in efficiency of using resources.	3 (severity is naturally lowered, as the result is not a wrong diagnosis but prioritisation of the patient).	3	9	Monitoring can be done by counting delayed patients and falsely high-prioritised cases.	To manage the risk validation of the prioritisation regarding AF levels and the associated risks to patient health, it should be conducted.	Explainability: validation of prioritisation.
24	Black box nature of the system	The model might detect background noise which affects the final decision.	Datasets might include two types of data: patients who are in the hospital and healthy whose data have been collected from smartwatches.	The model is able to detect patients, as their record shows the PPG signal of a person in a horizontal position (from a hospital) vs. a vertical one (healthy).	4	3	12	Two aspects of the model should be monitored: Robustness: By testing the accuracy versus the benchmark dataset. Explainability: It should be tested after each training, as the relevant AF signal might change with learning the new dataset by	Explainability: This can help in feature extraction and understanding if the model detects relevant features. HITL in the explainability: Does the practitioner agree on the reason for a certain decision, i.e., linking the explanation with domain expert knowledge? Device: Using other sensors might help the model detect the AF signal and exclude noise signals.	Explainability: validation of prioritisation. HITL: Domain expert knowledge

								the model. We might also need to use multiple explainability methods.		
25	Data collection and environmental factors	Data quality can cause a performance issue.	This risk is associated with low-quality data from movements or scenarios that are not included in the training set.	Poor prediction could lead to incorrect priorities and delay the diagnosis of real patients.	3	4	12	We can monitor this risk by tracking cases associated with a high value of uncertainty and investigating which subgroup they belong to.	Data with low quality can be excluded using other sensors, such as a strong movement detected in the accelerometer, and some signals can be excluded if the prediction is associated with a high value of uncertainty.	Reliability: Data quality consideration (hardware) Robustness: Low data quality consideration (data)
26	Data collection and environmental factors	Low data quality can cause risk.	Low data quality can propagate errors to the model.	This can cause wrong detection and labelling of AF signals.	3	3	9	(See mitigation first.) Evaluate the data quality before and after notifying the patient to reduce its movement.	Provide the patient with notification to reduce movement if AF is detected so the collected signal quality increases. Transparency: Explaining and teaching the patient to use the device can increase the data quality.	Reliability: Data quality consideration (hardware) HITL: an indication for a human to take control.
27	Data collection and environmental factors	Season changes might affect the PPG signal.	There might be a change in the signal between seasons leading to variability in the model output.	This can cause performance issues.	2	3	6	This risk can be monitored by measuring robustness (accuracy).	Including data from different seasons explicitly (i.e., data including the date of the recording) or implicitly in the training data might improve the accuracy.	Resilience: Data quality consideration (environmental influence)
28	Data collection and environmental factors	Data drifting might cause risk.	Possible drifting in the data (sensor issue) or model decision can cause this risk.	An increase in false detections can result in the incorrect prioritisation of patients.	3	3	9	Monitoring this risk can be achieved through -Regular checks using predefined cases to assess the drift in the model's decisions. -Comparison test between the prerecorded PPG signal and the collected data.	Reducing the risk can be done by regular calibration of the sensors and the model on a regular basis. Sensor drifting is considered a robustness issue; thus, another model (the statistical traditional model, maybe) that detects the drift can be implemented.	Robustness: Low data quality consideration (data)
29	Data collection and environmental factors	Data collection issues can be due to: (Other Devices) Electrode / Contact Issues (Other Devices) Skin Irritation	Device: - Device must be worn tight enough to capture good signals. - Some patients will have skin irritation despite materials being generally well tolerated.	Patient: -Poor device comfort might lead to poor adherence to monitoring.	2	3	6	This risk can be monitored by a survey about patient compliance.	HITL: - Practitioners can regularly contact the patients to follow up and provide support. - Different materials (silicon, fabric and other straps) should be available to the patient to choose from.	Reliability: Data quality consideration (hardware) Robustness: Low data quality consideration (data)

30	Data collection and environmental factors	Environmental factors can impose risk. Examples of these factors are: - Not patient-related, such as bright lights (including sunshine), motion or vibration (cars, buses, planes, etc.), sweat and humidity and dirt on the sensors. - Patient-related, such as using dominant vs non-dominant wrist/upper arm, pressure changes on sensors, the presence of tattoos, different skin tones, and having dry skin.	Environmental factors might introduce noise in PPG signals to varying degrees.	Patient, healthcare professional: - This risk might cause a poor signal to be captured (high uncertainty).	3	4	12	Monitoring this risk can be done by using additional sensors and algorithms for detection [external cause] and testing the model with this noise (robustness accuracy).	To mitigate this risk, few aspects can be targeted: -reliability (data quality): Additional sensors and algorithms are needed to exclude this data. [External Environmental Cause] -Transparency: teaching how the device should be used, possibly via video instructions. These primarily need to be dealt with by patient onboarding and assessment of their physical characteristics.	Reliability: Data quality consideration (hardware) Robustness: Low data quality consideration (data)
31	Data collection and environmental factors	(Safe Use) lack of device Robustness (like waterproofness and resistance to impact) might cause a risk.	A patient may accidentally break the device if it is delicate, especially if it is on the wrist.	The lack of robustness might leave the device unusable (water ingress or blunt force damage may destroy it). Another impact could be chemical burns or fire (battery damage may cause).	3	2	6	Monitoring the risk can be done by testing the device's resilience with water.	To mitigate this risk, device resilience should be improved. The device should be of the same quality and waterproofing standard as expected consumer electronics like smartwatches and phones.	Transparency: Patient knowledge (activity indicator)
32	Hardware & User Experience	There is a risk around device power inefficiency.	Using many sensors might cause a drain in the battery's usage.	The device requires prolonged charging, which often leads to the loss of captured data.	2	4	8	Monitoring the risk can be done by calculating the overall use of the power.	Reducing this issue can be done by removing redundant sensors; for example, a gyroscope is not needed, as an accelerometer is more efficient and provides enough information.	Hardware: device robustness (activity indicator)
33	Hardware & User Experience	(Lifetime) A small battery can cause a risk, as it very frequently needs charging.	How is the device powered? How is it recharged, and when?	Any recharge time will be lost monitoring. If the device takes a long time to charge or requires charging often, patients will plug it in overnight, which is the best monitoring period (lack of movement).	2	4	8	Monitoring can be done by measuring the charge consumption and checking compliance from the contact sensor.	To mitigate the risk, clear instructions on how long the device battery lasts and how/when to charge it are needed.	Hardware: device robustness (activity indicator)

34	Hardware & User Experience	(Lifetime) The risk of device reuse, including the hygienic concerns.	If the device is reusable (between patients), you need to consider sterilisation and/or unpairing of devices from patient-linked accounts/records.	Patients prefer things to be packaged nicely when they get them. Even if devices are recycled/reused, you need a method to make patients feel like it is new to them. Need to ensure that devices don't cross patient records or allow patients to see other people's data. Patient: As the device is transferred from one patient to another, their private health data can be leaked to the next patient if data is not removed. Using a used device carries extra health concerns (unclean device). The patient might show less compliance or be repulsed knowing it's been used.	3	3	9	Monitoring this risk can be done by conducting a survey about patient compliance.	Mitigating this risk can be done by: - Biohazard risk will need closer assessment. - sterilisation - Unpairing of devices from patient-linked accounts/records. - Devices are packaged nicely when they get them.	Hardware: device robustness (activity indicator)
35	Hardware & User Experience	(Misuse) Underuse of the device is a potential risk.	Patients may underuse or remove the device (if there is no feedback on whether it is working over long monitoring periods).	Patient, healthcare professional: This risk leads to reducing the monitoring period of the patient.	2	5	10	Monitoring this risk can be done by conducting a survey about patient compliance.	Mitigating this risk can be done by improving transparency. Patients should have some way to verify the device is working. Even if it's just that they can access their heart rate trend or see the time.	Transparency: Patient knowledge
36	Hardware & User Experience	(Misuse) Patients might not use the device as indicated.	Patient: - Discomfort (taking it off at night as it affects sleep) might cause data loss when the device is not in use.	Patient, healthcare professional: - The low patient compliance might reduce the monitoring period and really at the times when the signals will likely be cleanest.	2	5	10	Monitoring this risk can be done by conducting a survey about patient compliance.	Mitigation can be done by improving transparency; for example, explain to patients why they should wear the device at night.	Transparency: Patient knowledge

37	Hardware & User Experience	(Misuse) More than one patient might wear the device.	Patient: - The patient might give the device to someone else to try (the “try this” effect) (puts it on their cat).	Patient: Wrong data that doesn't represent the patient can be recorded, analysed and used for decision-making; as a result, a wrong decision is made. Healthcare professional: Practitioners cannot differentiate between the actual patient and the other person (or object) in the recording.	3	3	9	Monitoring this risk can be done by conducting a survey about patient compliance.	Mitigation can be done by explaining to patients why this is not a good idea!	Transparency: Patient knowledge
38	Hardware & User Experience	(Clinical Action) Long or extended monitoring time might be needed to give the final diagnostic decision to the patient.	Patient: Patients might be less compliant with the follow-up test (ECG) if they feel over-monitored.	Patient, healthcare professional: A patient might refuse or reduce the monitoring period of the ECG, which is essential (the approved diagnostic tool) to also find Afib.	4	3	12	Monitoring this risk can be done by conducting a survey about patient compliance.	To mitigate this risk, transparency should be improved by explaining to patients why they should wear the device at night.	Transparency: Patient knowledge
39	Hardware & User Experience	(Clinical Action) Patients who are PPG positive but ECG negative might lose trust in the organisation.	Patients: (psychological cause) Patients might become worried that the abnormality was simply missed on the ECG. Patients don’t understand the limitations of the technology (PPG). Therefore, they might not understand how it can be found on one test but not another.	Patients might experience a prolonged consultation process. Thus, healthcare professionals need longer time with patients to communicate difficulties and PPG limitations. This might mean financial waste for the organisation.	2	4	8	To monitor this risk, notes of such instances should be taken, then the rate at which such an issue would happen should be calculated.	Improving transparency might help mitigate this risk by making it clear to patients that the results are not conclusive and merely help with the consultation process when showing results on the device.	Transparency: Patient knowledge Explainability: Ability to link the findings with the causes
40	Health Data Privacy	Privacy concerns: Cloud-based data analysis could potentially lead to leakage of private information.	Data leakage might happen due to insufficient security of the AI system.	This risk might affect patient’s quality of life (insurance at higher prices, limited job prospects, all because of the result leaks).	3	2	6	To monitor this risk, life quality should be compared between the patients and healthy people, for example, if	Privacy: The model utilises internal data collected to process and make predictions; data sent to medical facilities like the NHS has to undergo encryption.	Security: System Architecture

								there is a significant difference between insurance prices. This could potentially point to data leakage, which may or may not be associated with this device.		
41	Health Data Privacy	Data leakage might occur.	If the process is streamed and processed on the cloud.	This can cause a potential security risk of patient data.	3	2	6	System security is checked regularly.	Security: Mitigating this risk can be done by processing the data on the device and sending (optional or if it has an advantage) only some information (minimising the data) to the cloud. Applying good software development practice is also essential.	Security: System Architecture
42	Health Data Privacy	Streaming or sharing data might pose privacy risks.	The risk of medical information leakage might be induced by: (Device/system) Sharing or streaming data. (Patient) Uncontrolled companion devices such as a person using their phone + an app. (Device) If a device has its own comms, then you need to consider things like WiFi/mobile comms availability.	Patients might become victims of private healthcare data misuse (insurance company rejection or increased prices due to recently leaked health reasons). There is a need for a robust method to retrieve data from the devices. Some patients might have difficulties dealing with the technology (examples: old people who might be less familiar with modern devices and lose connection outside of city areas).	3	3	9	This risk might be monitored by reports submitted by patients on suspicious activities (hyper-tailored ads, service changes (like the insurance example).	Improving privacy and security might mitigate the risk by: - Store updates of new data on internal servers and process the data on the device itself. - Increase memory of device for the entire monitoring period - considering a hybrid system for communication. Another aspect to be improved is transparency by: - implementing an alert system: real-time intervention in case of emergency.	Security: System Architecture
43	Health Data Privacy	Low data security is a potential risk.	Device, streaming system: Data leakage might occur in data transferring to organisations (the NHS) due to lack of security in data protection infrastructure.	This might put the patient in an unfavourable situation when using other systems (insurance companies refusing due to potential AFib) which violates the patient's privacy (patient's confidentiality).	3	2	6	Monitoring this risk can be done during the testing stage by cybersecurity checks.	Improving the security and mitigating this risk can be done by: - Data is encrypted and anonymised if it is streamed. - Establish policies and better design for data security infrastructure during the development stage.	Security: System Architecture

44	Model generalisability limitation	The risk associated with model generalisability can be the presence of a high number of false positives and inaccuracies in prediction or detection.	False positives and false negatives can cause social anxiety for patients, especially if the PPG-AI prediction did not align with the doctor's decision. (vibration sensitivity affected by lifestyle, instead of vibration caused by possible Afib episodes).	This will lead to waste of resources (allocating medical resources to healthy people, wasting time and money, making queues longer, etc.) and a psychological effect on the patient, such as losing trust between doctor and patient, causing treatment compliance issues.	3	4	12	Monitoring of this risk can be done by surveys that show the users are fully aware of the device they are using.	To mitigate the risk associated with the psychological effect: Transparency can be improved by better communication between the device and users. The transparency should be dynamic in both its detail level and complexity level. For example, the device should provide a full answer, i.e., the prediction + its uncertainty (diff level of explainability). To improve transparency, patients should be educated using easy-to-understand context to convey the severity of predictions (providing references for patients to compare with, e.g., the odds of getting cancer are the same as getting hit by a car ten times over) and provide additional advice that may not necessarily be a doctor's appointment. HIL should be in the transparency: emphasis on the doctor role in explaining the output of the results to the patients.	Transparency: Patient knowledge. Explainability: Ability to link the findings with the causes (with uncertainty indicated) HITL: indication for a human to take control
45	Model generalisability limitation	Risk due to failure in the intended use due to false negatives Afib Alerts.	(Software / AI) Risk can arise due to multiple reasons, such as: detection is not robust, Unusual patient presentation, Episodes are too short.	This risk can have a psychological effect (over-reliance on AI, hinder consultation processes due to communication difficulty, ignore other symptoms); this might lead to delays in treatments.	3	3	9	(accuracy study) we would require follow-up testing in a percentage of patients who do not show any signs of Afib on our test to verify the FNR. (FNs are unknown unknowns, so very challenging.).	Mitigating this risk can be done by: - improving generalisability (example: at times when the heart rate doesn't match movement levels regardless of the Afib algorithm). - improving resilience (example: some way of patients recording symptoms/a higher sensitivity mode if the symptom right now flag is set.)	Generalisability: accurate model (vs expected data) Explainability: Ability to link the findings with the causes (with uncertainty indicated) HITL: indication for a human to take control
46	Model generalisability limitation	(Measurements) Heart rate is calculated incorrectly.	Software / AI detection is not robust.	Poor data quality reduces overall trust in the system. This is very obvious when things go wrong (large spikes) or HRs that aren't physically possible.	2	4	8	Monitoring HR for abrupt changes and sensible min/max limits might be useful.	Excluding unrealistically high HR can help in mitigation.	Robustness: Low data quality consideration (data)
47	Model generalisability limitation	(Uncertainty) On-device algorithm learning	Software/AI: - PPG signals tend to be non-stationary, and the waveform is dependent on device position, pressure and other factors.	Patient, healthcare professional: - On-device learning may overfit to a signal shape which abruptly changes. Another problem could be a degraded performance of the detection algorithm, high uncertainty.	3	3	9	Monitoring of the risk can be done by frequent testing of the robustness (accuracy) of the system on the benchmarking dataset.	Mitigating this risk can be done by improving robustness and generalisability, which are addressed at the algorithm design stage. (Any on-device learning needs to be on a moving scale (exponentially weighted moving averages, etc.).)	Generalisability: accurate model (uncertainty calculated)

48	Model generalisability limitation (trade-off)	There is a trade-off between specificity vs. sensitivity.	There is often a trade-off between sensitivity and specificity. Improving one can sometimes lead to a decrease in the other. The balance between them depends on the specific requirements and consequences of the classification problem.	Sensitivity: If the test has high sensitivity, it means it correctly identifies most of the people who have the disease (few false negatives). Specificity: If the test has high specificity, it means it correctly identifies most of the people who do not have the disease (few false positives).	4	4	16	Monitoring this risk can be done by calculating both values. Specificity, sensitivity.	Mitigating this risk can be done by having two modes, one for screening and another for monitoring. One of the models prefers sensitivity, and the other specificity.	Beneficence: Complexity-Aware ML Models
49	Model generalisability limitation (trade-off)	(Interpretations) if the measurement window is too short, short episodes may be missed. .	Most PPG algorithms trade off window length vs specificity. Short episodes are likely to be missed.	Patient: - Patients with high-burden short runs of afib will likely be missed. (This is problematic because more "new" Afib patients start with paroxysmal Afib.) - Long windows can also artificially inflate Afib burden time.	3	4	12	This risk can be monitored by measuring accuracy.	Mitigating this risk can be done by improving resilience: - experimenting with the minimum target length of an afib episode detected. - Need to select a sensible analysis window length.	Beneficence: Complexity-Aware ML Models
50	Risk on organisation	Patients might lose trust in the NHS system.	This might happen in case the patients get a low prioritisation ranking.	Patients feel forgotten by the NHS.	3	3	9	Monitoring can be done by counting missing cases based on small studies, like comparing the procedures with using the PPG vs without.	Mitigating this risk can be done by improving transparency by communicating clearly and effectively with the patients the need to revisit the clinic if symptoms change, yet they have not been evaluated.	Transparency: Patient knowledge
51	Risk on organisation	The system might generate more workload on the NHS staff, which defeats the purpose of the product.	The system might require intensive contributions from the human (NHS to review long PPG record), yet these signals are not fully understood.	This means NHS staff will be reviewing long PPG records manually. This causes more workload.	3	3	9	Monitoring this risk can be done by comparing the efficacy of the new procedure with the previous procedure (captured cases/hours spent). Different procedures can have different HITL roles.	HITL and Transparency: The human role in reviewing the PPG record is reduced. The NHS staff role is to review surmised results from the record. However, there is a need for the ability to review long records if needed. This is coupled with the explainability of the PPG signal and provides the correct level of details and depth at the right time for each of the stakeholders (NHS staff here most relevant). This means high HITL in the explainability of the ML model. The provided output for the practitioner should be actionable; for example, a certain level of frequency and length of episodes means x% priority and	Generalisability: accurate model (vs expected data) Explainability: Ability to link the findings with the causes (with uncertainty indicated) Transparency: practitioners understand how

									certain action to take. Thus, the action is traceable back.	the system works and what the results mean.
52	Risk on organisation	Performance issues such as too many false positive Afib alerts might pose a risk.	(Software / AI) detection is not robust.	This affects the organisation and wastes public funds and resources. Patients might have psychological effects (unnecessary panic). And overall there might be a delay in the treatments.	3	4	12	Monitoring this risk can be done by estimating FP results using validation data, which must be representative of the real world (always worse).	Mitigating this risk can be done by improving resilience, for example to adapt the sensitivity of devices in the field/where the algorithm is run to reduce FPs if too large (adjust the sensitivity of detection).	Resilience: system training adjustment to balance FP vs FN.
53	Risk on organisation	(Interpretations) There might be resistance from the NHS staff to use new technology.	Signals are not clinically useful/actionable.	Primarily clinicians will have poor acceptability of such a system being used. "Why can't we just use ECG patches?"	2	5	10	Physician feedback on the implemented system will be key (Survey).	Transparency: The system must be as easy for clinicians to use and follow. The workflow for patient registration, device fitting, data capture, review and referral must be clear.	Generalisability: accurate model (vs expected data) Explainability: Ability to link the findings with the causes (with uncertainty indicated) Transparency: practitioners understand how the system works and what the results mean.
54	Risk on organisation	(Clinical Action) Risk Scoring should be clearly defined.	Clinicians will want proof that this screening is worthwhile and identifies risk better than scoring like CHA2DS2-VASc. This also ties into the risk for Afib (patient groups).	Clinician buy-in to actually support the use of the device is important. Are young people likely to get treatment for small amounts of Afib anyway? If not, there might be no point in doing the test.	3	3	9	Monitoring this risk can be done by surveys that target the healthcare providers and run efficacy study frequently.	Mitigating this risk can be done by improving transparency and explainability that target the healthcare providers. It is important to improve robustness (validation) to prove that the new procedure/system is more efficient.	Beneficence: Complexity-Aware ML Models

55	Risk on organisation	Too many people might seek consultation from detection.	Too many patients might seek the detection for AF using the device.	(Healthcare system/organisation) Workload might increase due to an influx of people without differentiating who's in need, as a result wasting the organisation's resources. Patients might experience prolonged queues and delay treatments in serious situations.	3	3	9	Monitoring this risk can be done by tracking the number of patients detected vs. time required to detect the case, i.e., an efficacy study of the new procedure.	The device was initiated by doctors; hence, the organisation has control and the ability to reduce or increase the number based on their capability.	Explainability: validation of prioritisation
56	Training data limitation	Low robustness can pose a risk.	The model provides an accurate prediction, but the features used for the model decision are related to the dataset but not the disease.	The system might fail to detect cases that are not representative in the training data.	3	4	12	Monitoring this risk can be done by regular checks for robustness.	Mitigating this risk can be done by improving the robustness of the model by training on a wide range of data (using multiple datasets and including various conditions and subgroup considerations) and fine-tuning the model.	Robustness: Subgroup training consideration (data and algorithm)
57	Training data limitation	Performance issues might cause a risk.	Training labelled data usually are collected from specific settings like cardiology clinics. (The dataset from the hospital has a way higher proportion of Afib than the general community.).	The developer might not have control over the input data, thus using available datasets which might have a limited number of scenarios that don't reflect all cases in real-life practice.	3 (severity is naturally lowered, as the result is not a wrong diagnosis but prioritisation of the patient)	4	12	Accuracy measured on the benchmarking dataset that considers all possible cases.	Mitigation of this risk can be done by improving robustness and reliability by collecting data with control settings or combining different datasets. Alternatively, we might consider extracting a dataset from the available ones to teach the model how to deal with these scenarios.	Robustness: Subgroup training consideration (data and algorithm)