22HLT01 QUMPHY

D8 - Second report from the ethics advisor to 22HLT01

Organisation name of the lead participant for the deliverable: PTB

Due date of the deliverable: 30.06.2025

Actual submission date of the deliverable: 30.06.2025

Confidentiality Status: Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444

Deliverable Cover Sheet

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EURAMET. Neither the European Union nor the granting authority can be held responsible for them.

The project has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States.

Table of contents

Document history

Date	Version	Author versions	Change description
26.06.25	0.1	Jenny Venton	First draft, incorporating feedback from consortium
30.06.25	0.2	Nando Hegemann	Minor changes
30.06.25	1.0	Jenny Venton	Minor edits, update confidentiality status cover page
30.06.25	1.1	Jenny Venton	Added section 2.3, ‘Response to first ethics report’

1 Introduction

1.1 Outline and scope of the report

QUMPHY is a timely project and is laying the groundwork for safe and equitable health measurements for all through benchmark datasets and a good practice guide for photoplethysmography (PPG) machine learning (ML) algorithms. While not a clinical project, QUMPHY is in a strong position to be able to incorporate and implement ethical considerations at the early research stage of uncertainty quantification. This could act as a flagship example of how to incorporate ethical considerations at the earliest stages of health ML research. This ethics report will address the ethical considerations and implications of using uncertainty quantification (UQ) methods for machine ML applications in the analysis of PPG signals. To this end it will focus on the following key areas:

1. **Informed Consent**: Addressing the ethical considerations related to obtaining informed consent from individuals whose PPG data is used for model training and validation.
2. **Regulatory Compliance**: Ensuring that the development and deployment of ML models for medical applications comply with relevant ethical guidelines and regulatory standards, such as those outlined by the European Commission and international bodies like the WHO and ITU.
3. **Bias and Fairness**: Evaluating the potential for bias in ML models, particularly concerning demographic subgroups such as sex, age, and skin tone, to ensure equitable performance and avoid discrimination.

4. **Data Privacy and Protection**: Ensuring compliance with data protection regulations, such as the General Data Protection Regulation (GDPR), and addressing the ethical handling of personal health data collected from PPG signals.
5. **Transparency and Explainability**: Discussing the importance of transparency in ML models, including the interpretability of model predictions and the communication of uncertainty estimates to end-users, such as clinicians and patients.

The ethics report will aim to provide a comprehensive overview of the ethical considerations taken in the project and suggest measures to further ensure the responsible and ethical use of ML models in the analysis of PPG signals for medical applications.

1.2 Involvement of the ethics advisor

The approach taken in this ethics report is to: help the consortium highlight potential areas of risk; to propose ways that QUMPHY output can be delivered safely, equitably and adhering to existing guidelines and regulations where applicable; and with as wide a reach as possible. The mandate of the ethics advisor as described in the grant agreement is to "…help the consortium with ensuring that a risk assessment plan is delivered for the development and use of AI […] as well for ethically correct access to the widely available data." The ethics advisor has played an important role in the project, actively participating in the months 9 and 18 project meetings. During these meetings, the ethics advisor led dedicated sessions to discuss and address ethical issues relevant to the project. The consortium maintains a continuous exchange with the ethics advisor, ensuring that their suggestions and recommendations are integrated into the project's framework. Additionally, the consortium proactively identifies potential ethical concerns and engages in discussions with the ethics advisor to develop effective mitigation methods. This ongoing collaboration ensures that the project adheres to the highest ethical standards and addresses any emerging ethical challenges promptly.

1.3 Project summary

PPG signals are rich in information and easy to measure passively without any physical or mental limitations of the subject. ML algorithms are commonly used to extract physiological parameters from PPG signals for diagnosis, avoiding the need for complex and costly clinical review. As of today, no regulations specifying how these ML algorithms have to be applied, how their performance has to be measured or how their associated uncertainties have to be specified exist. At the core of this project stands the development of measures to quantify the uncertainties associated with ML algorithms applied to medical problems, in particular the analysis and processing of PPG signals. To achieve this the following tasks are being addressed: (i) benchmark datasets are being generated using publicly available in vivo, and synthetic data (ii) different ML models and uncertainty quantification (UQ) methods are being used to analyse the processing of the PPG signals and specify the associated uncertainty and (iii) a good practice guide with accompanying software repository showcasing the used models, methods and benchmarks is being developed and will be made publicly available.

1.4 Project ethics self-assessment

1. **Involvement of humans, human cells or tissue**

   In this project it is important to clarify that no humans, human cells, or tissues are directly involved in any experimental or data collection processes. The project focuses on the development and validation of ML models using existing datasets of PPG signals. These datasets are either publicly available or synthetically generated, ensuring that no new human data is collected specifically for this project. The analysis and methodologies employed are designed to work with pre-existing data, thereby eliminating the need for direct human participation or the use of human biological materials. This approach ensures that all research activities are conducted in compliance with ethical standards and regulations, prioritizing the use of non-invasive and pre-collected data to achieve the project's objectives.

2. **Personal data**
   In this project the usage of personal data is carefully managed to ensure compliance with

ethical standards and data protection regulations. All data utilized in the project are either generated through computer simulations or derived from publicly available datasets that already conform to the EU data protection guidelines. This approach ensures that no new personal data is collected directly from individuals, thereby minimizing privacy concerns. For datasets used within the project that have access restrictions, these are handled with the utmost care and will only be made available to the public in strict accordance with the data access guidelines of the respective dataset. This meticulous handling of data underscores the project's commitment to maintaining high ethical standards and protecting individual privacy.

3. **Animals**
   In this project no animals are involved in any experimental or data collection processes. The project's focus is solely on the development and validation of ML models using existing datasets of PPG signals collected from consenting humans, which are either publicly available or synthetically generated. The only exception to this involves the potential presence of emotional support animals, which may accompany individual consortium members in certain settings to provide comfort and relieve stress. However, these animals are not subjects of the research and are not involved in any data collection or experimental procedures. This approach ensures that the project adheres to ethical guidelines and regulations concerning animal welfare.

4. **Non-EU countries**
   This project does not utilize any facilities, materials, or experimental designs from non-EU countries. While some datasets employed in the project originate from Non-EU countries, the project has not been involved in the creation of those datasets. All datasets in use have been published prior to the project, are frequently used in European research initiatives and, as far as the consortium is aware, have been collected with similar ethical and humane standards as those stated by the EU.

5. **Environment, health and safety**
   In this project, there is no risk of environmental damage, and consequently, no specific precautions are necessary to address such concerns. The experimental design and structure of the project have been carefully planned to ensure that they do not meet any criteria that could potentially harm the health or safety of the individuals involved. Additionally, the technologies employed in the project have been evaluated to confirm that they do not pose any undesirable side-effects that could endanger any of the persons participating in the project.

6. **Artificial intelligence**
   a. **Human agency and oversight:** All ML methods developed and investigated in the project are critically evaluated for their performance. The project's focus on uncertainty quantification for ML specifically emphasizes human agency and oversight, ensuring that humans can interpret the reliability of the outputs in conjunction with the associated uncertainties. The project is centred on establishing general procedures for uncertainty quantification and benchmarks ML models using specific, openly available, and anonymised datasets. Consequently, the results are not intended to inform decisions for direct clinical practice. The limitations of the employed models and the uncertainty quantification methods are thoroughly discussed and highlighted within the software and any publications produced by the project. It is crucial to note that no software or result of the project will attempt to exaggerate confidence or coerce, deceive, or manipulate individuals, as such actions are fundamentally opposed to the project's goals.

   b. **Privacy and data governance:** Privacy and data governance are handled with the utmost care and adherence to international, EU, and national laws regarding anonymity, ethical treatment of subjects, and data quality. All datasets utilised in the project are openly available and comply with EU standards. For datasets with

restricted access, these are clearly noted, and such datasets will not be published without the explicit consent of the original data holders. To ensure transparency in data processing and augmentation for machine learning purposes, scripts used to process restricted access datasets will be made publicly available. This allows for the reproduction of the project's results while maintaining data privacy. The selection of data subsets for training purposes is conducted internally, with consultation from the ethics advisor, to prevent or reduce inherent bias against minority groups. Network architectures, training procedures, and datasets are made publicly available to ensure reproducibility of the results. Furthermore, results are run multiple times to provide robust statistics for training procedures, ensuring the reliability and robustness of model outputs.

c. **Transparency:** Transparency is a cornerstone of the research methodology and dissemination strategy. All datasets utilized in the project are openly available, ensuring that the broader research community can access and verify the data used. For datasets with restricted access, these are clearly noted, and such datasets will not be published without the explicit consent of the original data holders. To further ensure transparency, scripts used to process restricted access datasets will be made publicly available, allowing for the reproduction of the project's results and providing insight into data processing and augmentation for machine learning purposes. Network architectures, training procedures, and datasets are made publicly available to ensure reproducibility of the results. Results are run multiple times to provide robust statistics for training procedures, ensuring the reliability and robustness of model outputs. All software developed in the project is tracked in an online repository with version control, allowing for traceability of the development process. An automatically generated online documentation is provided to simplify access to the software library, and scripts to reproduce any published results, including network architectures, training routines, and used hardware, will be provided.

d. **Fairness, diversity and non-discrimination:** All datasets used in the project adhere to the FAIR principles, ensuring that they are Findable, Accessible, Interoperable, and Reusable. A significant focus of the project is on investigating potential biases against demographic subgroups within these datasets, with specific activities aimed at examining biases related to sex, age, and skin tone. Considerable thought is given to the limitations of the produced results with respect to their application to demographic subgroups, and these concerns are discussed with the ethics advisor to ensure a comprehensive understanding of potential biases and limitations. The outputs of the project include discussions on the limitations of the data when applied to certain minority groups, highlighting areas where biases may exist or where data may not be fully representative. All machine learning applications developed in the project are trained on a variety of data to reduce bias within the constraints of the given datasets. Additionally, the uncertainty quantification methods employed are investigated for their potential to reflect biases against minority groups, ensuring that the project's findings are both robust and ethically sound.

e. **Societal and environmental well-being:** From a societal perspective, the project focuses on improving the reliability and trustworthiness of machine learning models used in medical diagnostics through the analysis of PPG signals. By enhancing the accuracy and uncertainty quantification of these models, the project aims to improve healthcare outcomes, particularly in the early detection and management of conditions such as hypertension. This can lead to better health management and

potentially reduce the burden on healthcare systems, thereby positively impacting societal well-being. Environmentally, the project does not involve any activities that could harm the environment. The research is conducted using existing datasets and computational models, which do not require physical experimentation or the use of hazardous materials. The project's reliance on data analysis and machine learning models ensures that there is no direct environmental impact, such as pollution or resource depletion. Thus, the project aligns with principles of sustainability and ethical research, ensuring that both societal and environmental well-being are preserved and potentially enhanced.

f. **Accountability:** Accountability is realized through a structured approach that ensures responsibility, transparency, and oversight throughout the development and operation of machine learning models. The project involves multiple stakeholders, including developers, researchers, and clinicians, who are collectively responsible for the functionality and outcomes of the ML systems developed. This accountability is supported by the project's commitment to transparency, as evidenced by the open availability of datasets, network architectures, training procedures, and software repositories. These resources are made publicly accessible to ensure that the processes and methodologies used can be scrutinized and validated by external parties. Furthermore, the project incorporates rigorous oversight mechanisms, including regular consultations with an ethics advisor and stakeholder committee, to ensure that the ML applications are developed and operated in an ethically sound manner. The project's focus on uncertainty quantification also enhances accountability by providing a clear understanding of the reliability and limitations of the ML models' predictions. This comprehensive framework will help developers and operators of the ML software applications explain how and why the systems exhibit particular characteristics or result in certain outcomes, thereby fulfilling the criteria for accountability.

2 Overall feedback

2.1 Feedback and Recommendations

General comments:

- Be clear about what the end goal of the project outputs are. For example, is this for research only or is the good practice guide and code framework to be used to evaluate medical devices?
- Where existing regulations (EU AI Act for example) have been considered, indicate this alongside any outputs and specify which aspects have been considered. This may help reassure end users who want to use the developed frameworks to support an application for regulatory approval, and improve acceptability of this work.
- There are many existing frameworks for reporting on or evaluating medical ML, and criteria for healthcare datasets that QUMPHY could benefit from. These resources are a good starting point:
    - Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD+AI): https://www.bmj.com/content/385/bmj-2023-078378
    - Prediction model risk of bias assessment tool (PROBAST+AI): https://www.bmj.com/content/388/bmj-2024-082505
    - Standards for data diversity, inclusivity and generalisability (STANDING Together): https://www.datadiversity.org/

- Note that factors such as older age, lower educational status and lower household income were associated with lower odds of wearable usage (Marvasti et al., 2024) meaning there will be less representation in the data used for QUMPHY for these groups.
- Patient, practitioner and end user feedback is increasingly important at all stages of medical software development. Include any views gathered from patients or end users on what kinds of uncertainty are important and what should impact uncertainty. If not relevant or not possible then comment on this.

Intended end user and impacted people views

- Standardised approaches and benchmark datasets for PPG are very timely. The Annex states that “The implemented algorithms will support device manufacturers, clinicians and end users in deciding on medical treatment plans for certain diseases such as diabetes and hypertension”. As the outputs of QUMPHY are intended for future medical use, and include ML tools, there is a responsibility to incorporate wider views such as that of clinicians, patients, physiologists, and other experts in the intended medical application domains.
- Guidelines describe how it is critical that teams that design, develop and test ML systems reflect the diversity of end users and people impacted by the ML system, not only in terms of gender culture or age but also in terms of professional backgrounds and skill sets (Ethics Guidelines for Trustworthy AI, 2019). Describe how are these views accounted for in the development of QUMPHY’s technical work.
- Clearly describe results of any findings or information gathering of patient/end user views on what is important to capture in terms of uncertainty quanitification, and what their views are on what should impact that uncertainty.

2.2 Essential recommendations

- How will scope of application, and limitations, be communicated to end users? For example, specify which devices the method(s) was validated for and the population demographic used to develop the data (i.e. healthy/patient, age, sex, ethnicity).
- Be clear about what the benchmark datasets can and can not be used for. Are they for ML developers generally, or specifically for PPG ML developers. As this will be publicly available resources from standard development laboratories there is a risk it could be used to validate PPG AI in settings the resources are not developed for. State the domain of each UQ method, in terms of device type, population demographic, laboratory/hospital/community setting, disease state, exercise etc.
- Alongside the code framework and good practice guide, be clear about what it can and can not be used for and what scenarios it is applicable for.
- Where findings are reported on benchmark datasets, make clear that factors independent of the dataset labels (for example healthy vs atrial fibrillation) may contribute to the model's ability to classify between the groups.
- Include some comment clarifying the position of this work on pulse oximetry (PO) and the known impact of skin tone on PO performance.
- Include a summary in the deliverable outlining feedback gathered from practitioners and clinicians about what kind or type of uncertainty is useful and important for them and what they want an uncertainty value to tell them.
- For all datasets that are used, check and state whether they comply with EU guidelines.

2.3 Response to First Ethics Report

The ethics advisor can confirm, after assessment, that the project is working towards compliance with the ethical requirements raised in these reports. A session on the ethics report will take place at the M27 project meeting.

3 Work package and task feedback

3.1 Work package 1

- Different kinds of uncertainty are discussed. Comment on how these developed uncertainty methods will report on accuracy and uncertainty for different groups of people (i.e. age, sex, ethnicity) or different settings (i.e. wrist worn, health tracker, hospital etc), or if they will.

Task 1.1:

- State whether the ML models chosen are models that are currently used for PPG or will feasibly be used in real world applications.
- How detailed are the common evaluation frameworks and are they based on any existing evaluation frameworks? Are the common evaluation frameworks to be used internally or will this also feed into the framework being developed to be used by others in Task 2.3?

Task 1.2

- Describe how the validation of uncertainties on the prototypical datasets will translate to benchmark datasets. State any methods or assumptions made to do this.

Task 1.3

- Describe how we can verify that the uncertainty comparison on the protoypical datasets holds when translated to different datsets. For example, if the findings have used signals from healthy populations, how will the findings translate to a disease population.
- Make clear why skin tone classification is being used. For example, 'to understand whether, from an algorithm's perspective, there is a difference between signals from different skin tones'.
- The common evaluation framework being developed to distinguish in-distribution and out-of distribution data sounds very promising and perhaps a key of this kind of work. More detail on in-distribution and out-of distribution data and the links to uncertainty would be helpful.
- Activity 1.3.4. Understanding the dependence of model performance and uncertainty on factors such as sampling rate, pre-processing, signal or label noise, length of signals etc. are important findings. Will the findings cover how these factors interact? For example, if signal length gives uncertainty x to a measurement, and pre-processing gives uncertainty y, how do these combine. Is there a risk of amplifying uncertainties if an end user combined these? State what combinations (if any) the outputs are valid for.
- Activity 1.3.5. Detecting whether a sample is out of distribution is an important problem for uncertainty in medical applications. Describe how the project will determine whether epistemic uncertainty is the most appropriate for detecting out of distribution samples.
- Activity 1.3.6. Describe the limitations of the Fitzpatrick skin type scale as a skin tone classification system, and describe how this might impact the findings on skin tone in this activity. Namely, that Fitzpatrick skin typing was created to classify persons with white skin in relation to selecting UV dose for treating skin diseases, not for categorising skin tone or structure (Fitzpatrick, 1988).
-

3.2 Work package 2

Task 2.1

- State how the scope of the benchmark datasets will be communicated. For example, in a publication, on a website with the datasets, alongside the framework. Scope here means demographics, characteristics and disease status of the population in the benchmark datasets, as well as device types and setting where the data was collected.
- Describe how social and health system priorities have been accounted for when creating the benchmark datasets (Panch et al, 2020)
- Comment on how the generalisabilty of the benchmark datasets to other data will be quantified.

- WFDB format is suggested for the benchmark datasets. Explain the rationale behind this in the context of dataset formats used by device manufacturers or in a clinical setting.
- Clearly state the classification or regression problem the benchmark datasets are applicable for. For example, is the atrial fibrillation (AF) benchmark dataset benchmarking uncertainty for identifying AF in a signal, or the potential for a sinus rhythm signal to become AF.

Task 2.2

- End users may have different application problems, models or datasets. In the good practice guide, include how end users can extrapolate these findings to their own work and how different application problems, models or datasets might impact the uncertainty quantification output.

Task 2.3

- As the code framework is to support industry in attaining certification and regulatory approval, describe how the code framework has taken the EU AI Act into account.
- State whether the framework is intended to be an end to end framework for evaluating medical machine learning for PPG, or for part of the process (i.e. for model development but not training data collection). If the framework is for part of the process, describe how considerations for the rest of the development pathway will be communciated to end users. When evaluating medical AI it is important to consider the wider influences and implications.

4 Other ethics issues

At this point there are no additional ethics issues that need to be addressed by the project.

5 Potential misuse of results

Potential misuse of the results and ML models in general is carefully considered and mitigated through several integrated measures. The project focuses on developing and validating ML models for the analysis of PPG signals, which are crucial for medical diagnostics and monitoring. To prevent misuse, the project emphasizes transparency and accountability by making all datasets, network architectures, training procedures, and software repositories openly available. Furthermore, this project should ensure transparency in the scope, application domain and limitations of the benchmark datasets, good practice guide and code frameworks made publicly available. This transparency ensures that the processes and methodologies can be scrutinized and validated by external parties, reducing the risk of unethical use. It also ensures the outputs will not be used for purposes they are not validated for which could cause unintended harms. Additionally, the project incorporates rigorous oversight mechanisms, including regular consultations with an ethics advisor and stakeholder committee, to ensure that the ML models are developed and operated ethically. The focus on uncertainty quantification further enhances the reliability and trustworthiness of the ML models' predictions, providing a clear understanding of their limitations and potential biases. By discussing the limitations of the data and models, particularly concerning minority groups, the project aims to prevent any discriminatory or harmful applications of the results. Moreover, the project's commitment to open science practices, such as publishing results in open-access journals and presenting findings at international conferences, ensures that the broader research community can access and build upon the project's outcomes responsibly. These comprehensive measures underscore the project's dedication to ethical research practices and the responsible use of ML in medical applications.

6 Approval of ethics advisor for first ethics report

This report is approved by the project’s external ethics advisor

Dr. Jenny Venton
The Royal Surrey NHS Foundation Trust
Egerton Road, Guildford, Surrey, GU2 7XX, UK

Date & signature: Jenny Venton, 30 June 2025

7 References

European Commission (2019). *Ethics guidelines for trustworthy AI.* Brussels: High-Level Expert roup on Artifical Intelligence. https://digital-st

European Commission (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*. Brussels: Directorate-General Research & Innovation, Research Ethics and Integrity Sector. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-

EU Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (OJ L 119, 4.5.2016, p. 1), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1399

EU Directive 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA (OJ L 119, 4.5.2016, p. 89). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1399642418437&uri=CELEX:32006L0024

European Commission (2021). *Ethics and data protection*. Brussels: Directorate-General Research & Innovation, Research Ethics and Integrity Sector. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf

Council of Europe, European Court of Human Rights, European Data Protection Supervisor, European Union Agency for Fundamental Rights, (2018). *Handbook on European data protection law : 2018 edition.* Publications Office of the European Union. https://data.europa.eu/doi/10.2811/58814

European Commission (2021). *EU Grants: How to complete your ethics self-assessment vers. 2.0*. (2021-07-13), https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

Fitzpatrick TB. (1988). The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Arch Dermatol.*, 124(6):869–871. https://doi.org/10.1001/archderm.1988.01670060015008

Marvasti, T. B., Gao, Y., Murray, K. R., Hershman, S., McIntosh, C., & Moayedi, Y. (2024). Unlocking Tomorrow's Health Care: Expanding the Clinical Scope of Wearables by Applying Artificial Intelligence. *Canadian Journal of Cardiology*, *40*(10), 1934–1945. https://doi.org/10.1016/J.CJCA.2024.07.009/ATTACHMENT/F1A3F39A-E190-44FA-8490-7216F98C1582/MMC1.PDF

Panch, T., Pollard, T.J., Mattie, H. *et al.* "Yes, but will it work for *my* patients?" Driving clinically relevant research with benchmark datasets. *npj Digit. Med.* **3**, 87 (2020). https://doi.org/10.1038/s41746-020-0295-6

Proposal for a Regulation laying down harmonised rules on artificial intelligence (2021-04-21), https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Commission (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust (COM(2020) 65 final).* Brussels. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Ethics Report Progress Description

Interpreting 'In Compliance' column

Column value	Interpretation
Yes	Recommendation satisfied
No	Recommendation not satisfied, unclear plan for completion
Pending	Recommendation not satisfied, clear plan for completion
Unsure	Need more information to determine if the recommendation is satisfied

The overarching goal of the project is "*The overall objective is to provide trustworthy machine learning models for analysing photoplethysmography signals in a medical context, by developing methods for the quantification of uncertainty in supervised machine learning and deep learning models applied to photoplethysmography signals and generating reference datasets to benchmark those models, supported by software being developed that will be publicly available for independent review of the models.*" The recommendations have been reviewed in that context.

2.1 Feedback and Recommendations

Recommendation	Progress	Timeline	In Compliance (Yes / No / Pending)
Be clear about what the end goal of the project outputs are. For example, is this for research only or is the good practice guide and code framework to be used to evaluate medical devices?	The end goals of the project are specified in the 5 objectives described in section B1 of the JRP. In addition, publications (articles, reports, software, data) contain detailed information about the scope of the research and its applicability to different scenarios (e.g., scientific use, practical applicability, data quality). If applicable, links to published guidelines are highlighted.	Implemented in publications.	Yes The overarching end goals of the project is "*The overall objective is to provide trustworthy machine learning models for analysing photoplethysmography signals in a medical context, by developing methods for the quantification of uncertainty in supervised machine learning and deep learning models applied to photoplethysmography signals and generating reference datasets to benchmark those models, supported by software being developed that will be*

			publicly available for independent review of the models."
Where existing regulations (EU AI Act for example) have been considered, indicate this alongside any outputs and specify which aspects have been considered. This may help reassure end users who want to use the developed frameworks to support an application for regulatory approval, and improve acceptability of this work.	This will be taken into account in any future publications.	To be implemented in future publications.	Pending
There are many existing frameworks for reporting on or evaluating medical ML, and criteria for healthcare datasets that QUMPHY could benefit from. These resources are a good starting point: • Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD+AI): https://www.bmj.com/content/385/bmj-2023-078378	Many thanks for the helpful references to other frameworks. We will consider these when developing our framework.		Yes

• Prediction model risk of bias assessment tool (PROBAST+AI): https://www.bmj.com/content/388/bmj-2024-082505 • Standards for data diversity, inclusivity and generalisability (STANDING Together): https://www.datadiversity.org/			
Note that factors such as older age, lower educational status and lower household income were associated with lower odds of wearable usage (Marvasti et al., 2024) meaning there will be less representation in the data used for QUMPHY for these groups.	Thanks for raising this issue which relates to the wider issue of bias in ML models arising from the many potential biases in the training datasets. We have not done any data collection for this project and so cannot take account of these factors. However, the details relating to the collection of the data for each benchmark dataset are described in detail in the D4 deliverable report which has already been submitted. Work in progress in Task 1.3 is considering out of distribution data, which will provide insights into the impact of models on groups that are not well represented in the training data.	Complete.	Pending To be discussed at M27 meeting
Patient, practitioner and end user feedback is increasingly	Feedback from practitioners and end-users is constantly sought through engagement	Ongoing until the end of the project.	Pending

important at all stages of medical software development. Include any views gathered from patients or end users on what kinds of uncertainty are important and what should impact uncertainty. If not relevant or not possible then comment on this.	with the Stakeholder Committee. Feedback from patients is not planned for this project. The various uncertainty measures are highly technical and so best practice for these has been taken from the literature. In particular, the Good Practice Guide will be presented to the Stakeholder Committee (A2.2.2) and their feedback will be incorporated into the Guide.		
Standardised approaches and benchmark datasets for PPG are very timely. The Annex states that "The implemented algorithms will support device manufacturers, clinicians and end users in deciding on medical treatment plans for certain diseases such as diabetes and hypertension". As the outputs of QUMPHY are intended for future medical use, and include ML tools, there is a responsibility to incorporate wider views such as that of clinicians, patients, physiologists, and other experts in the	We have clinicians (Surrey) and physiologists (KCL) as part of the QUMPHY consortium who are both making valuable contributions to the project. We have not included patients as described above. The Stakeholder Committee also includes experts with a wide range of experience in this area. The ML models developed in QUMPHY could be taken up by device manufacturers and further developed and implemented in medical devices. However, this is beyond the scope of the QUMPHY project. We will not be using our models for live implementation.	Complete.	Yes Gathering patient feedback is beyond the scope of the project, QUMPHY is a research project

intended medical application domains.			
Guidelines describe how it is critical that teams that design, develop and test ML systems reflect the diversity of end users and people impacted by the ML system, not only in terms of gender culture or age but also in terms of professional backgrounds and skill sets (Ethics Guidelines for Trustworthy AI, 2019). Describe how these views are accounted for in the development of QUMPHY's technical work.	The QUMPHY consortium is large, consisting of 16 partners and so we have a team which is diverse in terms of gender, ethnicity, nationality, age, professional background and skill sets who are all contributing to the multidisciplinary aims of the project. We have regular WP1 and WP2 meetings that many people attend and contribute to.	Complete.	Not actionable All members of the consortium are given equal opportunity to contribute. The make up of the team is fixed at this stage.
Clearly describe results of any findings or information gathering of patient/end user views on what is important to capture in terms of uncertainty quantification, and what their views are on what should impact that uncertainty.	The views of patients/end users have not been sought for this project as uncertainty quantification for machine learning is a highly technical and specialised area. However, the consortium is in constant exchange with experts from the respective fields through the Stakeholder Committee and attendance at international conferences. Information obtained by these sources is collected and integrated in the project's publications. In addition,	Ongoing.	Pending

	the deliverable D6 report will include a recommendation that patient/end users views should be sought when implementing ML and UQ for live use.		

2.2 Essential Recommendations

Recommendation	Progress	Timeline	In Compliance (Yes / No / Pending)
How will scope of application, and limitations, be communicated to end users? For example, specify which devices the method(s) was validated for and the population demographic used to develop the data (i.e. healthy/patient, age, sex, ethnicity).	The scope of application and limitations will be communicated to the end users in the Good Practice Guide, which is currently in progress. The D4 deliverable report provides information on the collection device and population demographics for each of the benchmark datasets. This information will also be included in the appropriate publications.	The GPG is in progress and is due by May 2026. The D4 report has been submitted.	Pending
Be clear about what the benchmark datasets can and cannot be used for. Are they for ML developers generally, or specifically for PPG ML developers. As this will be publicly	We can be clear about what the benchmark datasets *contain* (and these are clearly described in the D4 deliverable report, already submitted). However, we do not feel that we can specify what they can or cannot be used for, since this is up to	Complete.	Pending Details of dataset intended use to be included in the repository and/or a link to the D4 deliverable.

available resources from standard development laboratories there is a risk it could be used to validate PPG AI in settings the resources are not developed for. State the domain of each UQ method, in terms of device type, population demographic, laboratory/hospital/community setting, disease state, exercise etc.	the end user to take responsibility for. The UQ methods are generally applicable to any machine learning regression or classification problem. It is the datasets that contain the information regarding the device type, population demographic, setting, etc. and these are all described in the D4 deliverable report for each of the benchmark datasets.		Discuss at M27 meeting
Alongside the code framework and good practice guide, be clear about what it can and cannot be used for and what scenarios it is applicable for.	The applicable scenarios for each of the benchmark datasets are described in the D4 deliverable report which has been submitted. The code framework will be made openly available, and its intended use is described in the D2 deliverable report. The Good Practice Guide is in progress and will include recommendations for good practice which should be followed.	Complete.	Pending Details of intended use in good practice guide. Details of code framework intended use to be added to repository. Wording to be discussed at M27 meeting
Where findings are reported on benchmark datasets, make clear that factors independent of the dataset labels (for example healthy vs atrial fibrillation) may contribute to the	This will be included in the D3 deliverable report which is currently in progress.	Due March 2026.	Pending

model's ability to classify between the groups.			
Include some comments clarifying the position of this work on pulse oximetry (PO) and the known impact of skin tone on PO performance.	Most PPG datasets currently available do not include information on skin tone so it is then impossible to determine the impact of skin tone on machine learning performance. However, A1.3.6 aims to classify PPG signals by skin tone and this activity will highlight how well machine learning can distinguish between signals based on the skin tone which will indicate how much of a problem this issue is. A presentation and paper at the Computing in Cardiology conference in 2024 considered the effect of skin tone on classification accuracy for a simple binary classification of high/not high blood pressure. However, this study used the AuroraBP dataset which is not balanced across the skin tones so further work could be done if a more skin tone balanced dataset could be obtained. The D3 deliverable report (in progress) will include conclusions from this work regarding the use of PPG devices.	A1.3.6 is due in December 2025	Pending Add statement to A1.3.6 saying that the project isn't looking at pulse oximetry / blood oxygen measurements.

Include a summary in the deliverable outlining feedback gathered from practitioners and clinicians about what kind or type of uncertainty is useful and important for them and what they want an uncertainty value to tell them.	Feedback from practitioners and clinicians will be obtained as part of A2.2.4 and included in deliverable D5.	D5 is due by June 2026	Pending
For all datasets that are used, check and state whether they comply with EU guidelines.	The EU AI Act contains Article 10 on Data and Data Governance which only applies to high-risk AI systems. According to Article 6, a high-risk AI system is one that is used in a product, which is not the case for our research project. However, we have been cognisant of the recommendations in Article 10 for factors under our control. We have not been involved in any data collection and so can only report on these aspects, which we have done in the D4 deliverable report. We have also clearly documented all aspects of the benchmark datasets in the D4 deliverable report to assist anyone who may want to use these datasets in developing AI systems.	Complete.	Unsure To be discussed further

3.1 Work Package 1

Recommendation	Progress	Timeline	In Compliance (Yes / No / Pending)
Different kinds of uncertainty are discussed. Comment on how these developed uncertainty methods will report on accuracy and uncertainty for different groups of people (i.e. age, sex, ethnicity) or different settings (i.e. wrist worn, health tracker, hospital etc.), or if they will.	Analysing accuracy and uncertainty by sex and age is included in A1.3.2 which is currently in progress. Ethnicity is only available in one of the benchmark datasets (the MESA dataset for the benchmark problem of detection of sleep apnea) so results for this dataset will be considered with respect to ethnicity.	A1.3.2 is due by M27.	Pending
State whether the ML models chosen are models that are currently used for PPG or will feasibly be used in real world applications.	The selected benchmark problems are of real-world interest but the ML models developed in this project are for research purposes only and will not be implemented in actual devices as part of the QUMMPHY project.	Complete.	Pending Include some text with the code to describe the intended purpose
How detailed are the common evaluation frameworks and are they based on any existing evaluation frameworks? Are the common evaluation frameworks to be used internally or will this also feed into the framework being	The evaluation framework that we have developed is very comprehensive and comprises 6 metrics for classification problems and 4 metrics for regression problems. The metrics are a mix of local and global reliability. All the metrics are taken from the literature, with the exception of Variation Calibration Error (VCE) for	Complete.	Yes

developed to be used by others in Task 2.3?	classification problems which has been developed in the project. A small number of these metrics is often considered but this comprehensive framework is not based on any existing framework. The framework will be used internally, specifically for the D2 report and paper, but it will also be made openly available for others to use. Some aspects of the evaluation framework may be used in Task 2.3, but this has not been decided yet.		
Describe how the validation of uncertainties on the prototypical datasets will translate to benchmark datasets. State any methods or assumptions made to do this.	The same methods used for the validation of uncertainties on the prototypical datasets will be used for the benchmark datasets also in A1.3.1, which is currently in progress. There are no extra assumptions made for this.	A1.3.1 will be completed soon.	Pending
Describe how we can verify that the uncertainty comparison on the prototypical datasets holds when translated to different datasets. For example, if the findings have used signals from healthy populations, how will the findings translate to a disease population.	The uncertainty *methods* translate to different datasets, but the *results* are only applicable to the type of data contained in the dataset. So if the dataset only contains signals from a healthy population, it cannot be assumed to be applicable also for a disease population. If this is required, then a dataset including both healthy and	Complete.	Yes

	disease populations would be required.		
Make clear why skin tone classification is being used. For example, 'to understand whether, from an algorithm's perspective, there is a difference between signals from different skin tones'.	The idea of skin tone classification is to determine whether or not skin tone affects the collected PPG signals. This can be done by a statistical analysis of features derived from the signals for different skin tones, but we consider that a better way is to try a classification of skin tone using machine learning. If a high classification accuracy is obtained, then it implies that the PPG signals are affected by skin tone, but if a low accuracy is obtained, it implies that skin tone is not a significant factor in the collection of the signals. A similar rationale applies to the classification of biological sex in A1.3.2, since clearly classification of sex is not of any practical interest.	Complete.	Yes
The common evaluation framework being developed to distinguish in-distribution and out-of-distribution data sounds very promising and perhaps a key of this kind of work. More detail on in-distribution and out-of-distribution data and the links to	We intend to consider external datasets (i.e. completely separate patient populations/measurement devices) as these are certainly out of distribution. We may also consider subsets of PulseDB with models trained on one source (e.g. Vital) being tested on another source (e.g. MIMIC or Combined). For example, a model	A1.3.3 is due by September 2025.	Pending

uncertainty would be helpful.	trained on CalibFree Vital is tested on CalibFree MIMIC or AAMI MIMIC, which introduces out of distribution conditions due to differences in patient populations, sensor hardware, or blood pressure distributions (see Table IV https://arxiv.org/pdf/2502.19167).		
Activity 1.3.4. Understanding the dependence of model performance and uncertainty on factors such as sampling rate, pre-processing, signal or label noise, length of signals etc. are important findings. Will the findings cover how these factors interact? For example, if signal length gives uncertainty x to a measurement, and pre-processing gives uncertainty y, how do these combine. Is there a risk of amplifying uncertainties if an end user combined these? State what combinations (if any) the outputs are valid for.	The intention is to consider each of the factors listed in A1.3.4 individually. It would be a much more complex study to consider combinations of factors. However, we could consider this if time and resources allow.	A1.3.4 is due by December 2025.	Pending

Activity 1.3.5. Detecting whether a sample is out of distribution is an important problem for uncertainty in medical applications. Describe how the project will determine whether epistemic uncertainty is the most appropriate for detecting out of distribution samples.	We will have some prior knowledge of the extent to which examples are out of distribution and will then observe the correlation between this and the magnitude of epistemic uncertainty (vs. other sources).	A1.3.5 is due by December 2025.	Pending
Activity 1.3.6. Describe the limitations of the Fitzpatrick skin type scale as a skin tone classification system and describe how this might impact the findings on skin tone in this activity. Namely, that Fitzpatrick skin typing was created to classify persons with white skin in relation to selecting UV dose for treating skin diseases, not for categorising skin tone or structure (Fitzpatrick, 1988).	We are aware that the six-point Fitzpatrick skin tone scale was not originally intended as a skin tone labelling system, but now, when skin tone is considered in conjunction with PPG signals, the Fitzpatrick scale is widely used. The Aurora BP dataset is one example of a PPG dataset where the Fitzpatrick skin tone class is recorded for a subset of the participants. A recently proposed alternative is the ten-point Monk scale, but we are not aware of any PPG datasets with the Monk class recorded as well. A major limitation of the Fitzpatrick scale (and all other similar scales) is that the classification of an individual is subjective and does not involve any measurement of skin tone	A1.3.6 is due by December 2025	Pending The results of on skin tone and PPG must include an indication of the limitation that Fitzpatrick scale is not a skin tone scale, so conclusions on PPG and skin tone in A1.3.6 in deliverable D3 will have this limitation.

	at the site of the wearable device, which would be ideal. Thus, there is some inherent inaccuracy in the skin tone labels. To take account of this, in A1.3.6 we report the usual accuracy for a machine learning six-class classification, but also what we term the "fuzzy accuracy", where a skin tone prediction is counted as correct if it is either the same as the skin tone label or differs by one class. Preliminary results show the accuracy is quite low, but the fuzzy accuracy is very high.		

3.2 Work Package 2

Recommendation	Progress	Timeline	In Compliance (Yes / No / Pending)
State how the scope of the benchmark datasets will be communicated. For example, in a publication, on a website with the datasets, alongside the framework. Scope here means demographics, characteristics and disease status of the population in the benchmark datasets,	This will be communicated in different ways: The D4 report contains such information and has already been submitted. Moreover, as described in A2.2.5, a Gitlab repository is under construction for dissemination of the benchmark datasets and associated codes.	A2.2.5 is due by March 2026.	Pending

as well as device types and setting where the data was collected.			
Describe how social and health system priorities have been accounted for when creating the benchmark datasets (Panch et al, 2020).	We have selected benchmark problems that involve major clinical priorities (e.g. prediction of blood pressure and detection of AF) rather than niche problems that only apply to a small number of patients. Reflecting the diversity of patients is a crucial point for the selection of suitable benchmark datasets. In the D4 deliverable report, these aspects are discussed for every benchmark problem and dataset, highlighting possible biases.	Complete.	Yes
Comment on how the generalisabilty of the benchmark datasets to other data will be quantified.	The easiest way is to evaluate the models generated on the benchmark datasets on external out of distribution datasets with the same metrics as before and compare the performance, which forms part of A1.3.3. One exemplary study is already finished and available at https://arxiv.org/abs/2502.19167	A1.3.3 is due by September 2025.	Pending
WFDB format is suggested for the benchmark datasets. Explain the rationale	The reason to suggest WFDB format as one option for the benchmark datasets is the fact that this is the file	Complete.	Yes

behind this in the context of dataset formats used by device manufacturers or in a clinical setting.	standard on PhysioNet. However, most of our code that we will publish starts with e.g. numpy arrays and csv files such that most benchmark datasets will also be provided as npy and csv files or we will provide code to generate npy and csv files.		
Clearly state the classification or regression problem the benchmark datasets are applicable for. For example, is the atrial fibrillation (AF) benchmark dataset benchmarking uncertainty for identifying AF in a signal, or the potential for a sinus rhythm signal to become AF.	The problems that the benchmark datasets are intended to address are clearly described in the D4 report which has been submitted. The benchmark datasets/code will be clearly linked to the D4 report that describes their intended use.	Complete.	Pending Wording to include in the repositories to be discussed at the M27 meeting
End users may have different application problems, models or datasets. In the good practice guide, include how end users can extrapolate these findings to their own work and how different application problems, models or datasets might impact the uncertainty quantification output.	The Good Practice Guide is currently in progress. We will reflect this recommendation while working on the Good Practice Guide.	A2.2.7 is due by May 2026.	Pending

As the code framework is to support industry in attaining certification and regulatory approval, describe how the code framework has taken the EU AI Act into account.	The code framework is not yet finished. We will take these aspects into account.	A2.3.4 is due by May 2026.	Pending Details of code framework purpose are pending. EU AI Act will be taken into account for relevant parts when purpose decided.
State whether the framework is intended to be an end-to-end framework for evaluating medical machine learning for PPG, or for part of the process (i.e. for model development but not training data collection). If the framework is for part of the process, describe how considerations for the rest of the development pathway will be communicated to end users. When evaluating medical AI, it is important to consider the wider influences and implications.	The framework is intended to be end-to-end, but excluding data collection which it is assumed that industry will perform for their problem of interest.	A2.3.4 is due by May 2026.	Pending Specify what the start and end of the end-to-end framework is in any documentation. Include text alongside the code framework that makes clear the framework is to support approval and cannot be relied on as sole evidence of model/UQ performance. Developer takes responsibility for evaluation but may find this framework helpful.