DATA MANAGEMENT PLAN

Grant Agreement number	22HLT01
Project short name	QUMPHY
Project full title	Uncertainty quantification for machine learning models applied to photoplethysmography signals
Data management plan	1st ☐ 2nd ☒
Confidentiality status	PU - Public, fully open

TABLE OF CONTENTS

1 Data management plan 2
1.1 Data summary 2
1.2 Findable, Accessible, Interoperable and Re-usable (FAIR) Data 2
1.2.1 Making data findable, including provisions for metadata 2
1.2.2 Making data accessible 2
1.2.3 Making data interoperable 4
1.2.4 Increase data re-use 4
1.3 Other research outputs 5
1.4 Allocation of resources 6
1.5 Data security 6
1.6 Ethics 6
1.7 Other issues 7

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.

European Partnership Co-funded by the European Union

METROLOGY PARTNERSHIP EURAMET

1 Data management plan

1.1 Data summary

Questions	Answers
1 Will you re-use any existing data and what will you re-use them for? State the reasons if re-use of any existing data has been considered but discarded.	*The project re-used 8 datasets which originated from the following external sources (from outside of this project):* - *R. J. Mieloszyk et al. "A Comparison of Wearable Tonometry, Photoplethysmography, and* Electrocardiography for Cuffless Measurement of Blood Pressure in an Ambulatory Setting". In: IEEE Journal of Biomedical and Health Informatics 26 (2022), pp. 2864–2875 - W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh. "PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods". In: Frontiers in Digital Health 4 (2023). - J. Bacevi˘cius, V. Plu˘s˘ciauskait˙e, ˘Z. Abramikas, I. Badaras, M. Butkuvien˙e, S. Daukantas, E. Dvinelis, M. Gudauskas, E. Jukna, M. Kiseli¯ut˙e, R. Kundelis, J. Marinskien˙e, B. Paliakait˙e, A. Petr˙enas, M. Petrylait˙e, A. Pilkien˙e, G. Pudinskait˙e, V. Radavi˘cius, A. Rapalis, . . ., and V. Marozas. "Long-term electrocardiogram and wrist-based photoplethysmogram recordings with annotated atrial fibrillation episodes [Data set]". In: Zenodo (2024). doi: doi.org/10.5281/zenodo.11242869. url: https://zenodo.org/records/11242869. - Z. Liu. PPGArrhythmiaDetection. https://github.com/zdzdliu/PPGArrhythmiaDetection. Multiclass Arrhythmia Detection and Classification from Photoplethysmography Signals Using a Deep Convolutional Neural Network. 2022. - A. Bernardini, A. Brunello, G. L. Gigli, A. Montanari, and N. Saccomanno. "OSASUD: A dataset of stroke unit recordings for the detection of Obstructive Sleep Apnea Syndrome". In: Scientific Data 9.1 (2022), p. 177. - X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alc´antara, C. L. Jackson, M. A. Williams, and S. Redline. "Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA)". In: Sleep 38.6 (2015), pp. 877-888 - Moody, B., Moody, G., Villarroel, M., Clifford, G. D., & Silva, I. (2020). MIMIC-III Waveform Database (version 1.0). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/c2607m - J. Torres-Soto and E. A. Ashley, "Multi-task deep learning for cardiac rhythm detection in wearable devices," NPJ Digit Med, vol. 3, no. 1, p. 116, 2020. *These data have been used for the following purpose:* • *Training machine learning models* • *Validation of the project's results* • *Provision of benchmarks*
2 What types and formats of data will the project generate or re-use?	*The project generated the following data:* • *Code (e.g. python, matlab, shell)* • *Text (e.g. tex, md, txt, html, rst, yml, pdf)* *The project reused the following data:* • *Measurement data (e.g. WFDB, h5, npy, CSV)* • *Numerical data (e.g. WFDB, h5, npy, CSV)* • *Text descriptions (e.g. md, txt, html, pdf)*
3 What is the purpose of the data generation or re-use and its	*Project generated 1 dataset:* - Moulaeifard, M., Charlton, P. H., & Strodthoff, N. (2026). MIMIC-III-

relation to the objectives of the project?	Ext-PPG: A PPG Benchmark Dataset for Cardiorespiratory Analysis (version 1.0.0). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/nmwb-6h34 *Purpose of the data generation:* *The data generated and re-used originate from measurements. They were used in meeting the project's objectives and in conference presentations and peer-reviewed publications.* *Data generated in relation to the objectives of the project* *The dataset has been generated by the consortium in order to meet objectives 1-4. A software repository including methods for training machine learning models and quantifying model uncertainties and scripts providing validation data has been created in the context of objectives 1 and 3. Metadata describing the developed benchmark problems have been created in accordance with objective 2. A framework for independently reviewing machine learning algorithms was build alongside the tasks of objective 4. All of these tasks rely on the generated dataset.* *Data re-used in relation to the objectives of the project* *All of the datasets listed in Question 1 have been used by the consortium in order to meet objectives 1-4. A software repository including methods for training machine learning models and quantifying model uncertainties and scripts providing validation data has been created in the context of objectives 1 and 3. Metadata describing the developed benchmark problems have been created in accordance with objective 2. A framework for independently reviewing machine learning algorithms was build alongside the tasks of objective 4. All of these tasks rely on the described datasets.*
4 What is the expected size of the data that you intend to generate or re-use?	*The overall size of the data exceeds 1TB.*
5 What is the origin/provenance of the data, either generated or re-used?	*Data generated in the project* *The provenance and context of the data listed in Question 3 were thoroughly documented. The data are accompanied by information on how they were captured, processed, analysed, and validated. Other relevant information was also provided. The data generated stem from postprocessing of already published datasets. Data processing and publication has been done in accordance with the original data holders requirements.* *Re-used data* *The existing data originates from several sources, which include: participant's pre-existing data, data from the scientific literature, and real-world measurement data. The provenance and context of the data listed in Question 1 were thoroughly documented. The data are accompanied by information on how they were captured, processed, analysed, and validated. Other relevant information was also provided. Data processing and publication has been done in accordance with the original data holders requirements.*

6 To whom might your data be useful ('data utility'), outside your project?	*The data might be useful to:* • *Stakeholders from industry: medical device and digital health companies, clinicians* • *Other scientists working in the field: uncertainty quantification, machine learning, medical imaging, data science, biosignals* • *Standardisation committees defining benchmark tests and reference accuracies for machine learning applications for PPG signals*

1.2 Findable, Accessible, Interoperable and Re-usable (FAIR) Data

1.2.1 Making data findable, including provisions for metadata

Questions	Answers
7 Will data be identified by a persistent identifier?	*Yes, the project's deposited dataset is identified by DOI.*
8 Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	*The metadata created for all of the project's deposited datasets are open under a Creative Commons Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following: datasets (description, date of deposit, author(s), venue and embargo); the European Partnership on Metrology and UK Research and Innovation funding; grant project name, acronym and number; licensing terms; persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant. Where applicable, the metadata include persistent identifiers for related publications and other research outputs.*
9 Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?	Yes, if reasonable, the following search keywords were provided in the metadata to optimise the discovery and potential re-use of the deposited datasets: uncertainty quantification, machine learning, ppg, and keywords related to the type of application.
10 Will metadata be offered in such a way that it can be harvested and indexed?	*Zenodo complies with FAIR principles (https://about.zenodo.org/principles/). The metadata are indexed in a searchable resource. Metadata are licensed under CC0, except for email addresses. All metadata are exported via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and can be harvested.* *Metadata included in the project's git repositories are openly accessible and can be crawled automatically.*

1.2.2 Making data accessible

Questions	Answers
Repository:	
11 Will the data be deposited in a trusted repository?	*The data and associated metadata, documentation and code are deposited in the trusted open access repository Zenodo (https://zenodo.org) or in one of the open access Gitlab repositories of the QUMPHY group (https://gitlab.com/qumphy).*
12 Have you explored appropriate arrangements with	*No, the data will be uploaded via a standard procedure and require no special arrangements.*

Questions	Answers
the identified repository where your data will be deposited?	
13 Does the repository ensure that the data are assigned an identifier? Will the repository resolve the identifier to a digital object?	*Yes, Zenodo assigns an identifier (DOI) to each of the project's deposited datasets. The repository resolves the identifier to a digital object. Git assigns a unique commit hash to each version of the code, which allows to distinguish between different versions of the software.*
Data:	
14 Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	*All of the data that are needed to validate the results presented in scientific publications are openly available as the default unless there is a specific reason not to publish the data.* *<u>Datasets which cannot be shared – voluntary restrictions</u>* *Other data may be made available on a case-by-case basis if it is relevant for third parties. The following data will not be made publicly available:* • *Data obtained with the permission of third parties, but the third parties have not agreed to make the data publicly available.* • *Data that discloses the identity of a manufacturer.* • *Data that compromises the protection of a participant(s) intellectual property.* *The level of data made available is also considered, for example, pre-processed data will not be provided unless there is a clear reason for doing so.*
15 If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	*The data used in scientific publications, posters and oral communications has been made available for re-use as soon as possible.*
16 Will the data be accessible through a free and standardised access protocol?	*Yes, Zenodo provides well described conditions for free and standardised access (see http://about.zenodo.org/policies/). Gitlab runs standard git protocols to ensure easy access to the data, while also providing free access to repositories listed as public.*
17 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	*There are no restrictions on the use of the published data, but users will be required to acknowledge the project and the source of the data in any resulting publications, according to the latest version of the CC-BY license.* *When third-party access restricted data are used, instruction on how to apply for access for the data are provided and scripts for processing the data into the correct format after access has been granted by the original data holders are included in the projects code.*
18 How will the identity of the person accessing the data be ascertained?	*There is no need to ascertain the identity of persons accessing the data.*

Questions	Answers
19 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	*This consortium has a Data Access Committee (DAC) consisting of the coordinator, the WP leaders and one additional institution who has experience with data management.* *Their remit will be to select the data that will be openly accessible on a case-by-case basis. Ethical aspects and data security, including intellectual property requirements, will be considered as will access requests to personal/sensitive data.*
Metadata:	
20 Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	*In Zenodo, metadata are licensed under CC0, except for email addresses. All metadata are exported via OAI-PMH and can be harvested.*
21 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data are no longer available?	*The data will remain available and findable for the lifetime of the Zenodo repository, which is expected to be a minimum of 20 years.* *If data are withdrawn from Zenodo, the DOI and the URL of the original object are retained. In case of closure of the Zenodo repository, best efforts will be made by Zenodo to integrate all content into suitable alternative institutional and/or subject based repositories.*
22 Will documentation or reference about any software be needed to access or read the data and will this be included? Will it be possible to include the relevant software (e.g. in open source code)?	*The one dataset listed in Question 3 is published in a common format and can be read using widely available software libraries (open source or commercial). The projects software repository contains code to interact with the dataset using open source python implementations. Code snippets (python) to interact with the data are also listed in the dataset description available on PhysioNet (see bibliographic reference in Question 3).*

1.2.3 Making data interoperable

Questions	Answers
23 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	*The datasets use the trusted repository's basic metadata schema for administrative data, which is compliant with the recommended standards used by DataCite (https://search.datacite.org/) and OpenAIRE (https://www.basesearch.net/).* *For individual datasets, the following discipline-specific vocabularies, standards, formats, and methodologies are used additionally to standard formats:* *WFDB (hierarchical file format; biological time-series)*
24 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to	*Mapping is not required as the terminology used will is chosen to be compatible with the existing literature.*

allow their re-use, refinement or extension?	
25 Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?	*Yes, the project's datasets that will be deposited in the chosen repository (e.g. Zenodo) include qualified references to other datasets from the same project or previous research.*

1.2.4 Increase data re-use

Questions	Answers
26 How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	*A short README file (e.g. Markdown) is provided together with the data listed in Question 3, in order to enable data analysis and to facilitate data re-use. Reports have been published in the project describing data processing and example use.*
27 Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement?	*The data are either licensed under the latest available version of the Creative Commons Attribution International Public License (CC BY) or a license with equivalent rights as set out in the Grant Agreement. Users are required to acknowledge the consortium and the source of the data in any resulting publications. Alternatively, the Creative Commons Public Domain Dedication License (CC 0) is used.* *In cases where data are third-party with limited access, information on how to apply for access are provided and scripts to process the data accordingly after access has been granted are provided through the project's software repository. (see bibliographic reference in Question 3).*
28 Will the data produced in the project be useable by third parties, in particular after the end of the project?	*Any data published in open-access journals are usable by third parties after their deposition in Zenodo. The data that do not relate to peer-reviewed publications are made available for re-use on a case-by-case basis.*
29 Will the provenance of the data be thoroughly documented using the appropriate standards?	*Yes, the provenance and context of the data is thoroughly documented to meet relevant standards. Data are accompanied by information on how they were captured, processed, analysed, and validated. Other information that enables interpretation and use are also provided.*
30 Describe all relevant data quality assurance processes.	*Data quality is assured through several quality assurance procedures:* • *Adherence to standards for data recording.* • *Use of controlled vocabularies and standard terminology.* • *Validation of the data collected.* • *Provision of test results along with the data generated or re-used in the project (see Questions 1 & 3).* • *Peer-review of publications based on those data.*
31 Further to the FAIR principles, DMPs should also address research outputs other	*Allocation of resources* *The estimated costs for making the data and other research outputs FAIR are 7500 € (personnel costs) (see question 34). The costs for*

[1] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

Questions	Answers
than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.	*making other research outputs FAIR are included in the project's budget and will be claimed if compliant with the Grant Agreement's conditions. The consortium's Data Access Committee had the overall responsibility for managing other research outputs (see question 36). Where feasible, long-term preservation is ensured by depositing the other research outputs in repositories. The Data Access Committee decided on a case-by-case basis on which other research outputs were deposited and for how long.* *<u>Security of other research outputs</u>* *All participants are either accredited to, or work in compliance with, the ISO 17025 standard on the "General requirements for the competence of testing and calibration laboratories". The participants store other research outputs on their organisations' networks, which are protected by firewall, backups etc. Other research outputs are also stored in the project's SharePoint environment, with password-protected login. Deposition in public repositories provides additional security as they have multiple replicas in a distributed file system which is backed up on a nightly basis. This project did not generate sensitive other research outputs. The other research outputs are safely stored in open access repositories.* *<u>Ethical aspects</u>* *There are issues that could impact on the sharing of other research outputs.* • *Information relating to other research outputs acquired from third parties, e.g. manufacturers, will not be shared without their explicit consent.* • *Information relating to other research outputs collected by the consortium at commercial sites will not be shared without the site owner's explicit consent.* *Ethical issues have been addressed as the project prepared two ethics reports.* *The project does not share other research outputs with identifiable personal information. Sensitive information relating to the other research outputs are collected, separated as soon as possible and kept secure.* *Please also see the information provided in section 1.3 below.*

1.3 Other research outputs

Questions	Answers
32 In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new	*The software developed in the project has been released under the EUPL license. The software repositories are hosted on Gitlab under the QUMPHY group and can be accessed publicly through the group's website on gitlab.com (https://gitlab.com/qumphy).*

Questions	Answers
materials, antibodies, reagents, samples, etc.).	
33 Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.	*As far as possible, the FAIR data approaches specified in questions 7-30 above have been applied to the management of this project's other research outputs. This commitment has been met by releasing the new software that was developed in the project under license.*

1.4 Allocation of resources

Questions	Answers
34 What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?	*The estimated costs for making the data and other research outputs Findable, Accessible, Interoperable and Re-usable (FAIR) are 7500 € (personnel costs). These costs have been kept to a minimum by using a free repository (Zenodo) and by making only relevant data and other outputs FAIR.*
35 How will these be covered? Note that costs related to research data/output management are eligible as part of the European partnership on metrology grant (if compliant with the Grant Agreement conditions).	*The costs for making the data FAIR are included in the project's budget and will be claimed if compliant with the Grant Agreement's conditions.* *Long term preservation will be ensured by depositing the data within repositories (Zenodo). There are no costs associated with the long-term preservation of the data in these repositories.*
36 Who will be responsible for data management in your project?	*The consortium's DAC has overall responsibility for data management. The coordinator leads this committee and is responsible for coordinating updates to the data management plan. The committee is responsible for organising data backup and storage, data archiving and for depositing the data within the repositories (Zenodo).*
37 How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?	*Long term preservation will be ensured by depositing the data within repositories (Zenodo). There are no costs associated with the long-term preservation of the data in these repositories.* *The data serve as benchmark cases for the developed methods and tools developed in the project. It enables stakeholders to test and understand the methods developed in the project. Hence, the data facilitate the uptake of the project's results.* *The Data Access Committee decided on a case by case basis on what data are kept and for how long.*

1.5 Data security

Questions	Answers
38 What provisions are or will be in place for data security (including data recovery as well	*<u>Data recovery and secure storage</u>* *The participants store data on their organisations' networks, which are protected by firewall, backups etc. Data are also stored in the project's SharePoint environment, with password protected login.*

as secure storage/archiving and transfer of sensitive data)?	*Deposition in the Zenodo public repository provides additional security as it has multiple replicas in a distributed file system which is backed up on a nightly basis.* *Transfer of sensitive data* *This project did not generate sensitive data.*
39 Will the data be safely stored in trusted repositories for long term preservation and curation?	*Yes, the data are safely stored in the Zenodo open access repository. Zenodo and the underlying Invenio Framework for digital repositories were designed according to the Open Archival Information Systems (OAIS) reference model. Zenodo is working towards ISO 16363 certification.*

1.6 Ethics

Questions	Answers
40 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics report(s) and the ethics section in the Annex 1.	*There are issues that could impact on data sharing.* • *Data acquired from third parties, e.g., manufacturers, will not be shared without their explicit consent.* *Data collected by the consortium at commercial sites have not and will not be shared without the site owner's explicit consent.*
41 Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	*There were no market or customer surveys conducted in the project.*

1.7 Other issues

Questions	Answers
42 Do you, or will you, make use of other national / funder / sectorial / departmental procedures for data management? If yes, which ones (please list and briefly describe them)?	*Data management is compliant with:* *The research data policy of the European Partnership on Metrology;* *European laws about data security and the protection of privacy (e.g. GDPR);* *Institutional guidelines;* *Scientific community guidelines*.

2 Open science: research data management

Statement	Put an X in the box to confirm	Or, list any exceptions to this
All participants have adhered to the requirements of the project's GA and CA with respect to open science: research data management (GA Article 17 and its Annex 5) for this reporting period	☒	