Final Public Report for 22HLT01 QUMPHY
Uncertainty quantification for machine learning models applied to photoplethysmography signals

Overview

Photoplethysmography (PPG) is widely used in clinical environments that use wearable devices, yet clinicians lack trust in AI-driven evaluation of these signals due to a lack of quantified reliability. The QUMPHY project met this need by developing a robust environment to quantify and validate both data-related (aleatoric) and model-related (epistemic) uncertainties of machine learning algorithms applied to PPG signals. Key outcomes included establishing six standardised clinical benchmark problems and their associated datasets, publishing a Good Practice Guide, and releasing a modular open-source software package, ensuring AI-driven medical sensing is precise, reliable, and trustworthy. These key outputs provide digital healthcare and medical device companies with options to compare and benchmark their products against, which facilitates later regulatory approval and helps them to integrate reliable machine learning solutions in enduser devices.

Need

Cardiovascular diseases, hypertension, and sleep disorders rank among the leading causes of global morbidity and mortality. Early and continuous monitoring via digital healthcare solutions, such as wearable smartwatches, has shown immense potential to mitigate these conditions through timely clinical intervention. PPG is an optical physiological sensing technique that measures blood volume fluctuations non-invasively, making it ideal for continuous, long-term health monitoring. Despite the ubiquity of PPG sensors in consumer wearables and clinical pulse oximeters, algorithmic evaluation of PPG signals to infer physiological parameters has rarely been used in high-stakes clinical environments.

One of the primary barriers to clinical translation is a fundamental lack of trust in machine learning (ML) predictions of physiological quantities. In medicine, diagnostic errors can be fatal (false negatives) or cause severe anxiety and overtreatment (false positives). Prior to this project, there were no standardising methods to specify the reliability or uncertainty associated with ML predictions.

Furthermore, PPG signals are highly susceptible to data-related quality issues, including ambient temperature, movement artifacts, and variations in sensor positioning or skin tone. Industry, clinicians, and regulatory bodies lack the tools to evaluate how these variations affect algorithmic performance. Addressing this measurement challenge required: (1) a formal framework to distinguish between data-related (aleatoric) and model-related (epistemic) uncertainties; (2) standardised clinical benchmark problems and high-quality, open-access datasets; and (3) a unified software framework to independently assess and review proposed models. Developing these tools was essential to satisfy the strict reliability requirements of European health regulations and to lay the foundations towards a path to regulatory approval for AI-driven diagnostic solutions.

Objectives

The overall objective of the project was to provide trustworthy machine learning models for analysing PPG signals in a medical context by developing methods for the quantification of uncertainty in supervised machine learning and deep learning models and generating reference datasets to benchmark those models, supported by publicly available software.

The specific objectives of the project were:

Report Status:
PU – Public, fully open

Final Public Report

Issued: August 2026

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.

European Partnership Co-funded by the European Union

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.

METROLOGY PARTNERSHIP EURAMET

1. To develop methods for quantifying the uncertainty for at least 3 existing classification and 3 existing regression supervised machine learning and/or deep learning models using photoplethysmography (PPG) data, considering the effects of both aleatoric (data) uncertainty and epistemic (model) uncertainty on model predictions.
2. To generate at least 5 measurement problems and their corresponding 5 datasets, using real and/or synthetic photoplethysmography data, that can be used to benchmark accuracy and uncertainty of supervised machine learning and deep learning models, and to make those reference problems and datasets available to the medical device and digital health communities via an online repository.
3. To validate the uncertainties obtained for existing machine learning and deep learning models of Objective 1 and to compare the accuracy and uncertainty of at least 3 classification and 3 regression machine learning and/or deep learning models in order to identify models and methods which have high accuracy and low uncertainty for a wide range of tasks.
4. To engage with the medical device, digital and health communities to promote the use of the good practice guide and accompanying software repository, support the adoption of benchmarking problems and datasets, and develop a framework for independently reviewing machine learning models proposed by industry to assist in getting regulatory approval.
5. To facilitate the take up of the technology and measurement infrastructure developed in the project by the measurement supply chain (NMIs, DIs, medical device calibration services), standards developing organisations (IEC, ISO), and end users (clinical practitioners, digital experts within the health communities, manufacturers of medical and healthcare products)

Progress beyond the state of the art and results

Methods for Quantifying Uncertainty of PPG ML Models

Prior to the project, machine learning had been applied to PPG signals for various diagnostic classification and regression problems (such as blood pressure and heart rate monitoring), but these applications were never integrated with formal uncertainty quantification (UQ). Furthermore, existing modelling approaches did not systematically distinguish between data-related quality variations and model-related approximation errors.

The project went beyond the state of the art by training and evaluating supervised prediction models operating on multiple input representations, including raw time-series signals, extracted physiological features, and image-based representations (continuous wavelet transforms and Symmetric Projector Attractor Reconstruction). It successfully developed and integrated both model-dependent UQ methods (maximum a posteriori estimation, Monte Carlo dropout, and deep ensembles) and model-independent post-hoc calibration methods (temperature scaling, isotonic regression, Venn-Abers predictors, and conformal prediction). These implementations successfully provided rigorous estimates representing both aleatoric (data) and epistemic (model) uncertainties for both classification and regression tasks.

To address the inherent subjectivity and uncertainty associated with ordinal clinical labels, such as Fitzpatrick skin tone grades, the project introduced a novel tolerance-aware “fuzzy accuracy” metric. This metric defines a prediction of a class as correct if the ground truth falls within a neighbouring class, thereby reflecting plausible adjacent-class ambiguity. In tandem, the project formulated a custom fuzzy cross-entropy loss function utilizing neighbour label smoothing, which allocates a small probability mass to adjacent classes during training. These technical developments successfully established a standardized mathematical framework to account for both algorithmic and label-induced uncertainty in medical signal processing.

Standardised Benchmark Problems and Datasets

Standard datasets for evaluating machine learning algorithms exist, but they have lacked standardized access interfaces, uniform preprocessing pipelines, and detailed subject metadata, which has historically hindered fair performance comparison.

The project progressed the state of the art by establishing six clinical reference benchmark problems and curating corresponding high-quality datasets. These benchmark problems and datasets were integrated into a unified open-source environment. For the first problem, the estimation of systolic and diastolic blood pressure,

the project prepared finger-based and wrist-based PPG segments synchronized with continuous invasive blood pressure measurements from the Aurora BP and VitalDB (PulseDB) datasets. For the second problem, the detection of atrial fibrillation, the project curated and structured large-scale datasets from DeepBeat, TriggersAF, and a newly developed clinical dataset within the project, MIMIC-III-Ext-PPG, which captures over six million 30-second fingertip PPG waveform segments from more than 6,000 intensive care patients. For the third problem, the classification of hypertension, the project structured a three-class classification task using stratified patient cohorts from Aurora BP, aligned with the 2024 ESC guidelines. For the fourth problem, the determination of vascular age, a multiclass classification task was established using simulated waveforms from the Pulse Wave Database spanning decadal age groups and clinical records from Aurora BP. For the fifth problem, the detection of obstructive sleep apnoea, the project curated overnight polysomnography from the Multi-Ethnic Study of Atherosclerosis (MESA) sleep study and the Obstructive Sleep Apnoea Stroke Unit Dataset (OSASUD). For the sixth problem, the determination of respiratory rate, the project utilized PPG segments from the OSASUD cohort and the newly developed MIMIC PERform Large dataset.

The MESA dataset served as a primary in-distribution training and evaluation resource, consisting of 276 subjects with overall sleep quality grade 7 and fully annotated apnoea-hypopnea events. In contrast, the OSASUD cohort, which includes 30 after-stroke patients admitted to Udine University Hospital with finger PPG and SpO2 signals sampled at 80 Hz and 1 Hz, was established as a high-quality clinical testbed for evaluating out-of-distribution (OOD) generalization. Each of these datasets was converted into standardized formats, structured into balanced stratified splits (for training, validation, calibration, and testing), and documented with comprehensive subject-specific metadata, including, if available, age, weight, biological sex, and Fitzpatrick skin tone, to ensure complete traceability.

Validation and Comparison of Accuracy and Uncertainty

While uncertainty calibration is highly sought after in clinical machine learning, rigorous empirical validation under varied data quality constraints and demographic imbalances had never been systematically performed for PPG processing.

The project advanced the state of the art by implementing a comprehensive validation framework to evaluate model reliability across multiple architectures, preprocessing pipelines, and clinical scenarios. Drawing from the baseline models established in Objective 1, which benchmarked raw-signal, feature-based, and image-based models on prototypical blood pressure estimation and atrial fibrillation detection tasks, the consortium systematically compared and validated combinations of models and uncertainty quantification methods, resulting in a massive report focussing on the empirical evaluation of in-distribution and out-of-distribution performance as well as sensitivity to demographic subgroups.

Key validation results demonstrated that while image-based models (such as continuous wavelet transform scalograms combined with the ResNet18 network architecture) generally achieved the highest predictive accuracy across regression and classification tasks, raw-signal-based networks (such as temporal convolutional networks) often provided the best calibration metrics. Post-hoc calibration methods were shown to consistently improve upon uncalibrated baselines; in classification tasks, Bayesian binning and isotonic regression reduced the expected calibration error by 60% to 94%, while Gaussian process-based methods and neural-network g-layers provided the best negative log-likelihood for regression. Furthermore, the project validated that while dataset perturbations and pre-processing choices (such as high-pass/low-pass filtering and down sampling) affect model accuracy, they have a remarkably minor impact on the stability of uncertainty calibration. The project successfully demonstrated techniques to disentangle aleatoric and epistemic uncertainties, showing that aleatoric uncertainty dominates in real-world clinical regression and correlates with prediction error, whereas epistemic uncertainty acts as an extremely powerful and reliable detector for out-of-distribution data, effectively reducing false positive diagnostic alerts.

Additionally, this validation framework was applied to evaluate algorithmic fairness, showing that joint-sex training does not induce substantial demographic bias, and that raw PPG signals carry biological sex characteristics that can be classified with moderate accuracy (up to 71%). The framework was also applied in two clinical publications to validate sleep apnoea severity characterization, where incorporating PPG morphological features with arterial oxygen saturation improved out-of-distribution generalization by up to 16%

in F1-score, and to examine Fitzpatrick skin tone classification, where utilizing a novel “fuzzy accuracy” metric overcame subjective label noise and demonstrated that PPG signals are systematically and discernibly affected by skin tone.

Community Engagement and Independent Review Framework

Medical device manufacturers and clinicians lacked standards, guidelines, or frameworks to evaluate the clinical safety and regulatory readiness of AI-based tools.

The project addressed this by creating a Good Practice Guide, which combines multiple peer-reviewed publications of the project, and releasing the open-source QUMPHY Software Package. To assist developers in getting regulatory approvals in the future (e.g., under the EU Medical Device Regulation), the project established a general, open-source review protocol that interfaces with black-box and white-box ML algorithms without requiring knowledge of their internal mechanisms. This protocol was integrated into a modular Python package (qumphy) alongside a public code documentation, example configurations, and an intuitive graphical user interface (GUI) to lower the barrier to entry for non-expert reviewers and clinical practitioners.

All objectives were successfully achieved.

Outcomes and impact

Key dissemination and communication activities

To drive community engagement and technology uptake, the project executed a comprehensive dissemination strategy, resulting in 21 scientific publications and 30 conference or workshop contributions. The publication portfolio comprises 11 peer-reviewed journal articles in leading open-access venues (such as eClinicalMedicine, IEEE Journal of Biomedical and Health Informatics, PLOS Digital Health, and Scientific Data), 6 conference papers (at IEEE/ACM CHASE, Computing in Cardiology, and FUZZ-IEEE), and 4 preprints deposited in public repositories like arXiv and TechRxiv. The consortium delivered 23 oral presentations and 7 poster presentations at key national and international venues, including invited lectures at Vienna University of Technology, the 14th Metrology for Digital Transformation Day, the VITAL Training School, and the Cardiac Physiome workshop.

To foster continuous dialogue, the consortium established a Stakeholder Committee of 13 registered collaborators and stakeholders representing 8 countries. This committee brought together 3 industrial/SME entities, 5 academic institutions, 4 metrology institutes, and 1 public healthcare agency. The project engaged this community through 3 stakeholder workshops, held in September 2023, April 2024, and April 2026, to capture requirements, select benchmark problems, and present final project outcomes. Additionally, knowledge transfer was supported by 2 major training sessions, including an NPL seminar on uncertainty-aware machine learning and a cardiovascular data science special session at the Computing in Cardiology conference.

Active external communication was maintained via the project website and a dedicated LinkedIn Discussion Group. General public outreach also included 2 popular science interviews published in Lithuania on elektronika.lt and delfi.lt explaining how smartwatches can prevent serious health issues, a poster exhibition at the Institute of Physics and Engineering in Medicine, and technical feature articles in the periodic publications of the Instituto Português da Qualidade.

Outcomes for industrial and other user communities

The project advanced the integration of trustworthy, uncertainty-aware algorithms into wearable medical devices and digital healthcare solutions. The Good Practice Guide (Objective 4), standardized benchmark datasets including the MIMIC-III-Ext-PPG (Objective 2), and the qumphy software package (Objective 4) empower industrial developers to build prediction systems that use uncertainty budgets for medical predictive tasks, thus enhancing clinical decision-making. Notably, KTU's sleep apnea algorithms (from Objective 1 and validated in Objective 3) is poised for integration into UAB Teltonika Telemedic's TeltoCare wristband, projecting a 5-year turnover of 50,000€ to 100,000€. WhaleTeq Co., Ltd further plans to utilize the project's IP for artifact simulation (Objective 1, Objective 3), anticipating 25,000€ to 50,000€ in turnover. Moreover, the robust uncertainty quantification framework (Objective 1, Objective 3) facilitates regulatory approval for medical

device companies and significantly aides certification bodies like Scarlet, which expect a substantial turnover of 1,000,000€ to 5,000,000€ over five years by streamlining medical software certification for PPG-based devices (Objective 4).

Outcomes for the metrology and scientific communities

The project introduced a rigorous mathematical and metrological framework for evaluating machine learning uncertainty in medical signal processing. By developing and proving methods to disentangle data-related noise (aleatoric) from model-related error (epistemic), the project established a pioneering precedent for constructing uncertainty budgets in complex, deep-learning-based inverse problems applied to PPG signals. Working closely with prominent European metrology institutes such as VSL, INRIM and MathMet and academic collaborators the consortium has provided the scientific community with 11 peer-reviewed journal articles, 6 conference proceedings, and 4 preprints, alongside public benchmarking datasets. These resources provide researchers worldwide with standardized tools to compare and validate new machine learning architectures.

Outcomes for relevant standards

The project's findings were disseminated directly to key standards and regulatory bodies, such as the BIPM Joint Committee for Guides in Metrology Working Group 1, the OIML Digitalisation Task Group, JCGM Work Group 1, the EURAMET Working Group on Metrology for Digital Transformation and the Mathmet 2026 Meeting. The consortium presented project activities and research results directly to these groups, highlighting the project results and how the evaluation of uncertainty can be beneficially used in metrology for medical and healthcare applications. Crucially, the Good Practice Guide, the standardized clinical benchmarks, and the modular software repository developed during this project lay the essential technical and conceptual foundations towards future standardization work on measurement uncertainty evaluation for AI-based medical software.

Longer-term economic, social and environmental impacts

In terms of social and health impacts, the project laid the foundations for trustworthy remote patient care and chronic disease self-management. By equipping machine learning models with reliable uncertainty quantification, patients will be able to confidently monitor cardiovascular and respiratory conditions using wearables, such as UAB Teltonika Telemedic's future TeltoCare wristband, dramatically reducing false-positive anxiety. In clinical environments, integrating uncertainty estimates into monitoring hardware will mitigate alarm fatigue for nursing staff, ensuring diagnostic alerts correspond to genuine clinical emergencies and allowing for early, life-saving interventions. Furthermore, the project's explicit focus on ethical aspects, such as algorithmic fairness across sex, age, and skin tone, directly helps prevent bias in future clinical applications, ensuring AI-driven diagnostic technologies perform reliably and equitably for all patient demographics.

The economic benefits of the project are substantial for both healthcare systems and the digital health sector. Catching physiological deterioration early prevents expensive hospital readmissions, optimizing public healthcare resources. For medical device manufacturers, the Good Practice Guide and modular qumphy review framework will streamline compliance with safety regulations like the EU Medical Device Regulation, as observed with certification bodies like Scarlet, significantly reducing development cost and time-to-market for AI-driven wearable products. The project's knowledge and IP will also drive future developments, such as WhaleTeq Co., Ltd's "Algorithms and Software Framework for Simulating Photoplethysmogram Artifacts", further contributing to innovation in the sector.

List of publications

Bench, C. et al (2025) 'Uncertainty quantification with approximate variational learning for wearable photoplethysmography prediction tasks', *Machine Learning: Health*, 1(1) p. 015013. Available at https://doi.org/10.1088/3049-477X/ae0b74

Bench, C. et al (2026) 'A systematic evaluation of uncertainty quantification techniques in deep learning: a case study in photoplethysmography signal analysis', *Machine Learning: Health*, 2(1) p. 015011. Available at https://doi.org/10.1088/3049-477X/ae4c8e

Charlton, P. et al (2025) 'Determinants of photoplethysmography signal quality at the wrist', *PLOS Digital Health*, 4(6) p. e0000585. Available at https://doi.org/10.1371/journal.pdig.0000585

Hanna, L. et al (2024) 'Safety and procedural success of daycase-based endovascular procedures in lower extremity arteries of patients with peripheral artery disease: a systematic review and meta-analysis', *eClinicalMedicine*, 75 p. 102788. Available at https://doi.org/10.1016/j.eclinm.2024.102788

Moulaeifard, M. et al (2026) 'Machine-learning for photoplethysmography analysis: Benchmarking feature, image, and signal-based approaches', *Biomedical Signal Processing and Control*, 120 p. 109831. Available at https://doi.org/10.1016/j.bspc.2026.109831

Moulaeifard, M. et al (2026) 'MIMIC-III-Ext-PPG, a PPG-based Benchmark Dataset for Cardiovascular and Respiratory Signal Analysis', *Scientific Data*, 13(1). Available at https://doi.org/10.1038/s41597-026-07335-8

Moulaeifard, M., Charlton, P., Strodthoff, N. (2025) 'Generalizable deep learning for photoplethysmography-based blood pressure estimation—A benchmarking study', *Machine Learning: Health*, 1(1) p. 010501. Available at https://doi.org/10.1088/3049-477X/ae01a8

Rinkevičius, M. et al (2024) 'Obstructive Sleep Apnea Characterization: A Multimodal Cross-Recurrence-Based Approach for Investigating Atrial Fibrillation', *IEEE Journal of Biomedical and Health Informatics*, 28(10) p. 6155-6167. Available at https://doi.org/10.1109/JBHI.2024.3428845

Rodway, A. et al (2024) 'Prognostic and predictive value of ultrasound-based estimated ankle brachial pressure index at early follow-up after endovascular revascularization of chronic limb-threatening ischaemia: a prospective, single-centre, service evaluation', *eClinicalMedicine*, 68 p. 102410. Available at https://doi.org/10.1016/j.eclinm.2023.102410

Vardanega, S. et al (2025) 'Attractor Image-Based Deep Learning of Arterial Pulse Waves for Age Classification', *Computing in Cardiology Conference (CinC)*, 52. Available at https://doi.org/10.22489/CinC.2025.343

This list is also available here: https://www.euramet.org/repository/research-publications-repository-link/

List of technical deliverables and the data management plan

Deliverable number	Title	Confidentiality status	Clickable links to 'Public' documents
D1	Report on developing methods based on existing supervised machine learning and/or deep learning models using PPG signals representing at least 3 classification and 3 regression uncertainty determinations, including a comparison of model performance for one prototypical regression and one prototypical classification problem	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529
D2	Report on developing methods on uncertainty quantification for one prototypical machine learning and/or deep learning classification and one prototypical regression problem using PPG signals including the effects of both aleatoric (data) uncertainty and epistemic (model) uncertainty on model predictions	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529

D3	Report on validating the uncertainties obtained for existing machine learning and deep learning models from D1 and D2. This includes the comparison of accuracy and uncertainty of at least 3 classification and 3 regression machine learning and/or deep learning models in order to identify models and methods which have high accuracy and low uncertainty for a wide range of tasks	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529
D4	Report that at least 5 datasets using real and/or synthetic PPG data have been generated, that can be used to benchmark accuracy and uncertainty of supervised machine learning and deep learning models. This includes making the 5 reference problems and their respective 5 datasets available to the medical device and digital health communities	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529
D5	Good practice guide for quantifying uncertainties for machine learning models applied to PPG signals. This includes an email confirmation to EURAMET that a repository containing the developed software has been made publicly available	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529
D6	Report containing guidelines to support the adoption of the benchmarking problems and the development of a framework for independent review of the proposed machine learning models by e.g., medical device, digital and health communities to assist them in getting regulatory approval	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529
D7	First report from the ethics advisor to 22HLT01	Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444	
D8	Second report from the ethics advisor to 22HLT01	Classified C-UE/EU-C – EU CONFIDENTIAL under the	

		Commission Decision No2015/444	
n/a	Data management plan	PU - Public, fully open	https://doi.org/10.5281/zenodo.22919529

Project start date and duration:	2023-07-01, 36 months	
Coordinator: Nando Hegemann, PTB E-mail: nando.hegemann@ptb.de Project website address: https://www.qumphy.ptb.de/home		
Internal Beneficiaries: 1. PTB, Germany 2. CMI, Czechia 3. IMBiH, Bosnia and Herzegovnia 4. IPQ, Portugal 5. LNE, France	External Beneficiaries: 6. FC, Belgium 7. FVB, Germany 8. KTU, Lithuania 9. THM, Germany 10. UGent, Belgium 11. Uni Oldenburg, Germany	Unfunded Beneficiaries: -
Associated Partners: 12. KCL, United Kingdom 13. NPL, United Kingdom 14. Sector Health, United Kingdom 15. SURREY, United Kingdom 16. UCAM, United Kingdom		