Leveraging Interpretable Counterfactuals to Inform the Design of AI-Enabled Technologies for Human-Centered Automation in Healthcare

by

Raul Isaac Ramirez Batista

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Systems Design Engineering

Waterloo, Ontario, Canada, 2026

© Raul Isaac Ramirez Batista 2026

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis.

This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Raul Ramirez was the sole author of Chapters 1, 2, and 5, which were written under the supervision of Dr. Catherine Burns and were not written for publication. The findings presented in Chapters 3 and 4 are being submitted for publication.

Exceptions to sole authorship of material presented in Chapter 3 are as follows:

Portions of Chapter 3 build on a graduate course project completed at the University of Waterloo for the course HLTH 650B: Machine Learning Techniques in Health, instructed by Dr. Helen Chen and Dr. Zahid Butt. The original course project was conducted by Raul Ramirez and Naomi Loughlin, both master's students of the Department of Systems Design Engineering. The initial dataset exploration and implementation of baseline traditional machine learning models were developed for that course. These components serve as the technical foundation upon which the work in this chapter is extended, but the results are not reported in this thesis.

For the purposes of this thesis, Raul Ramirez reconceptualized and expanded the original work. Specifically, Raul Ramirez:

- Redesigned and formalized the machine learning pipeline architecture, replacing the simpler final-project train/test workflow with reusable validation, preprocessing, model-building, evaluation, plotting, and output-saving functions;
- Reframed the analysis around bad-quality PPG detection, explicitly treating Quality = 0 as the target evaluation class and reporting metrics such as class-0 F1, recall, precision, F0.5, balanced accuracy, and AUROC;
- Changed the data-preprocessing strategy, using subject-info.csv and quality-hr-ann.csv, preserving the original Quality labels, creating a person_id grouping variable, and one-hot encoding individual motion conditions instead of reducing motion to a simple binary variable;
- Implemented group-safe model evaluation, using StratifiedGroupKFold to prevent subject/person leakage between training and test data;

- Integrated dimensionality-reduction variants, evaluating models with no reduction, PCA, and LDA rather than only using a single scaled feature representation;
- Integrated additional comparative benchmarking procedures, including GridSearchCV, RandomizedSearchCV, cross-validation result export, model-family summaries, learning-curve support, and combined model-comparison plots;
- Reran the models in two additional experiments: a motion-only feature experiment and a balanced-training experiment that undersamples the majority class only within the training split;
- Expanded the explainable-AI analysis from basic SHAP and LIME plots to systematic SHAP, LIME, and DiCE outputs for each trained model;
- Implemented DiCE, a model counterfactual explanation method, using bad-to-good quality counterfactuals that constrain changes to movement conditions and specifically test movement-to-rest swaps;
- Implemented a large language model, Llama, as an orchestrator to translate model outputs into natural language;
- Compared the outputs of Llama against GPT-5.5;
- Identified cases for the human validation study presented in Chapter 4.

Naomi Loughlin contributed to expanding the original coursework project, including implementing a Multi-Layer Perceptron (MLP), with and without early stopping. The initial findings of this work were presented as a poster at the 2026 HFES Health Care Symposium, hosted by the Human Factors and Ergonomics Society in New York.

The subsequent methodological expansion, interpretability integration, validation procedures, incorporation of additional models (including DiCE-based counterfactual methods), and the broader conceptual framing presented in this thesis were developed by Raul Ramirez. The scope and analytical depth of the work as presented in this thesis extend beyond both the original coursework submission and the conference poster. No portion of prior coursework is reproduced in its original or complete form.

Secondary Use of Data

Chapters 3 and 4 used either the Brno University of Technology Smartphone PPG Database (BUT PPG), version 2.0.0, or the outputs of algorithms developed using this dataset. The BUT PPG database was published on PhysioNet by Nemcova et al. (2024) and was used in this thesis for secondary analysis in accordance with its stated usage notes and Creative Commons Attribution 4.0 license. The database was originally collected to support the evaluation of PPG signal quality and heart rate estimation algorithms.

Research Dissemination

The research reported in this thesis was accepted and presented in the following venues:

1. The 25th Annual Human Factors Engineering Inter-University Workshop. University of Toronto-Human Factors Interest Group (Toronto, Canada). September 22, 2025. Oral Presentation.
2. The 2026 HFES International Symposium on Human Factors and Ergonomics in Health Care. Human Factors and Ergonomics Society (New York, United States). March 22-25, 2026. Poster Presentation.
3. Supporting Health by Technology Conference. University of Groningen (Groningen, The Netherlands). June 4-5, 2026. Oral Presentation.
4. A Safety & Security Colloquium. Delft University of Technology (Delft, The Netherlands). June 15, 2026. Oral Presentation.
5. BMS-CODE Section Meeting. University of Twente (Enschede, The Netherlands). June 22, 2026. Oral Presentation.
6. BMS-HIB Section Meeting. University of Twente (Enschede, The Netherlands). June 29, 2026. Oral Presentation.

Abstract

Hypertension is a prevalent cardiovascular condition characterized by blood pressure exceeding 130/80 mmHg, which often remains asymptomatic and undiagnosed. In Canada, 1 in 5 people with hypertension are unaware that they have the condition. The increasing integration of artificial intelligence (AI) into digital health technologies (DHTs), such as wearables, offers opportunities to support hypertension screening and monitoring; however, high levels of automation in DHTs may provide users with limited information about the conditions that influence the measurements or limitations of underlying algorithms. This is particularly important in systems that decide when to involve the user in the decision-making and action selection process, where limited visibility into the system's reasoning may affect how users interpret an unsuccessful measurement or the absence of a notification. This thesis investigates how explainable AI (XAI), model counterfactual explanations, large language models (LLMs), and behavior-informed interface design can be integrated to support more transparent and actionable communication between highly automated health systems and their users.

The research first examined the use of explainable AI to evaluate binary photoplethysmography (PPG) signal quality classifiers beyond predictive performance alone. Statistical analysis identified the acquisition and contextual conditions associated with poor signal quality, and multiple binary classification approaches were evaluated using both performance and interpretability criteria. SHAP and LIME were used to examine the features associated with individual model classifications, while DiCE was used to generate model counterfactuals presenting feasible input changes capable of shifting predicted signal quality. This approach supported the identification of models and features that were not only predictive but also more suitable for developing actionable, user-facing guidance.

The resulting model counterfactuals were subsequently incorporated into an LLM-assisted prompt-generation pipeline informed by the Fogg Behavior Model. Technical model outputs were translated into short, model-informed facilitator prompts that linked potentially modifiable measurement conditions associated with bad-quality predictions to practical corrective actions. A single reviewer then refined the generated messages, removing unsupported interpretations, improving alignment between model-identified conditions and recommended actions, and ensuring that the final prompts were appropriate for presentation on a wearable interface under the relevant movement conditions.

To evaluate the transferability of this corrective feedback approach to an interactive wearable measurement context, the resulting prompts were implemented in a Figma prototype of the wearable interface using unsuccessful smartwatch electrocardiogram (ECG) measurement scenarios. Twenty-eight participants interacted with prototype smartwatch interfaces and evaluated the clarity, usefulness, confidence, information needs, and intended actions associated with the proposed and existing interfaces.

Participants generally valued the proposed interface's ability to explain a measurement condition likely contributing to unsuccessful measurements, provide corrective action, and communicate clear next steps. Reported retry intention varied across scenarios, and the proposed interface did not produce a statistically significant overall increase in reported retry intention relative to the existing interface. These findings indicate that the value of explanatory prompts extends beyond encouraging immediate repetition of a measurement and includes supporting users' understanding, confidence, and context-sensitive decision-making about whether to retry, wait, adjust the device, or seek additional information.

Based on the technical and human-centered findings, this thesis proposes ACT-AI, a design framework for integrating explainability, actionable counterfactual guidance, LLM-assisted interpretation, and behavior-informed interface design into highly automated digital health technologies. The framework advances the design of AI-enabled systems by contributing an approach for making system limitations and model-identified factors more accessible to users while supporting calibrated human involvement when corrective action is possible.

Keywords:

Artificial Intelligence, Automation, Health Care, Human Factors Engineering, Human-AI Interaction, Wearables

Acknowledgments

Research is not undertaken merely to acquire knowledge, but to contribute meaningfully to society. Throughout this degree, I have come to understand that contribution and learning are deeply intertwined. As a professor once noted, graduate school teaches you about the things you did not realize you did not know. Over the past two years, through the guidance and support of the people I have had the privilege of working with, I have come to appreciate the profound truth of those words.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Catherine Burns, for her guidance, support, and exceptional mentorship. Her insights and encouragement shaped the direction of this thesis and contributed significantly to my growth as a researcher. I am also grateful to the members of my examining committee: Dr. Moojan Ghafurian and Dr. Amir-Hossein Karimi, for their time, expertise, and thoughtful feedback on this work.

I sincerely appreciate the financial assistance provided by the CALDO-IFARHU-SENACYT Scholarship Program of Panama during this degree. Additionally, the study presented in Chapter 4 was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was also undertaken, in part, thanks to funding from the Canada Research Chairs Program, held by Dr. Catherine Burns. The presentation of its findings at the Supporting Health by Technology Conference in Groningen, The Netherlands, was supported by the GSPA's Graduate Student Research Dissemination Award and the Erasmus+ K171 Program.

To my colleagues and friends in Canada and abroad, you have made this degree truly unmatched: thank you for your encouragement, collaboration, and stimulating discussions.

Finally, to my dearest mom, dad, and brother: ¡Gracias! Thank you for your unconditional love, support, and unwavering belief in me. I am extraordinarily lucky to have had you by my side throughout this journey. The best is yet to come!

Land Acknowledgment

I acknowledge that much of the work presented in this thesis took place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. The University of Waterloo's main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River. Our active work toward reconciliation takes place across our campuses through research, learning, teaching, and community building, and is coordinated within the Office of Indigenous Relations.

Dedication

To my mom, dad, and brother

whose love and unwavering support were of the utmost importance through this degree.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract vi

Acknowledgments viii

Land Acknowledgment ix

Dedication x

List of Figures xvii

List of Tables xix

List of Abbreviations xx

Artificial Intelligence Disclosure Statement xxii

Graphical Abstract xxiii

Quote xxiv

Chapter 1 Introduction 1

1.1 Research Questions 6

1.2 Thesis Outline 7

Chapter 2 Background 9

2.1 The Human-Centered Problem of Wearable Automated Systems for Hypertension Screening 9

2.1.1 Health Technologies in Hypertension Care 9

2.1.2 Artificial Intelligence-Enabled Wearable Technologies for Hypertension Care 11

2.1.3 AI Approaches Toward Improving PPG Measurements 13

2.2 Human Factors and Risk Assessment Capabilities of Smartwatches 14

2.2.1 Human Interaction with Automation 14

2.2.2 Omission Errors and Trust in Digital Health Technologies for Hypertension Care 14
2.3 Adherence and Persuasive Design for Digital Health Technologies 15
2.3.1 Fogg Behavior Model 16
2.4 AI Limitations and the Need for Counterfactual Explanations 17
2.4.1 Counterfactual Algorithms 17
2.4.2 Counterfactual Applications and LLM Integrations 18
2.5 Knowledge Gap 20
Chapter 3 Designing Prompts to Assist Users in Taking Corrective Steps Toward Enhancing PPG Signal Quality 22
3.1 Overview of the AI Pipeline 23
3.2 The Dataset 23
3.2.1 Data Missingness 25
3.2.2 Correlation Analysis 25
3.2.3 Statistical Analysis 28
3.3 Methods 29
3.3.1 Machine Learning Pipeline 32
3.3.2 Explainable Artificial Intelligence Methods 35
3.3.3 Evaluation Metrics 36
3.3.4 Balanced Motion-Only Experiment 38
3.4 Results 38
3.4.1 Experiment 1: Models Trained with Full Tabular Features 39
3.4.2 Experiment 2: Motion-Only Features 39
3.4.3 Experiment 3: Balanced Motion-Only Models 40
3.5 Best Model Selection 41

3.6 Using the Fogg Behavior Model for LLM Prompt Design 43
3.7 Prompt Generation and Evaluation 45
3.8 Prompts Generated by the LLMs 50
3.8.1 Prompt Evaluation 50
3.8.2 Revised Version of the Prompts 51
3.9 Code Availability 52
Chapter 4 A User-Centered Approach to Evaluating Counterfactual Explanations Aimed at Prompting User Behavior to Perform Corrective Actions 53
4.1 Research Question and Hypothesis 53
4.2 Methods 54
4.2.1 Recruitment 54
4.2.2 Study Design 55
4.2.3 Prototype Development 55
4.2.4 Study Protocol 58
4.3 Scenario Design 60
4.3.1 Coughing 60
4.3.2 Talking 61
4.3.3 Running 62
4.3.4 Walking 63
4.3.5 Changing Light 64
4.3.6 Exploratory Multi-Factor Guidance in the Changing Light Scenarios 65
4.4 Data Analysis Plan 65
4.4.1 Order, Period, Learning, and Primacy 69
4.4.2 Prior Experience, Observed Behavior, and Other Confounders 69

4.5 Results 70
4.5.1 Participant Demographics 70
4.5.2 Scenario Interpretation Before Interface Interaction 71
4.5.3 Reported action after the interaction with the interface (RQ3.1 and RQ3.4) 73
4.5.4 Interface Preference and Perceived Helpfulness (RQ3.2) 78
4.5.5 AI-generated Videos in User-Centered Study Knowledge Tasks (RQ3.5) 80
4.5.6 Order, Period, Scenario, and Confounding Factors 81
4.6 Ethics Clearance 81
Chapter 5 Discussion 82
5.1 Developing AI models for Digital Health Technologies 82
5.1.1 Dataset Limitations 82
5.1.2 False Warnings 82
5.1.3 Model Training Effect of Motion-Only Features 83
5.1.4 Model Training Effect of Dataset Balancing 84
5.1.5 The Trade-off of Actionability 84
5.1.6 Dimensionality Reduction 85
5.1.7 Best Performing Model 85
5.2 Design of prompts to guide users in corrective actions for improving PPG signal quality 87
5.3 Human Validation of Model-Informed Counterfactual Prompts for Supporting Reported Corrective-Action Intentions 88
5.3.1 Scenario-level Analysis 89
5.3.2 Confidence and Ability 90
5.3.3 User Understanding of the System's Error State 91
5.3.4 Information Needs 91

5.3.5 AI-Generated Videos for Knowledge Tasks in Human-Centered Studies....92
5.4 ACT-AI Framework: Designing Human-AI Interaction....93
5.4.1 Core Design Principle....93
5.4.2 Mandatory Automation Level 9 Requirement....93
5.4.3 Framework....94
5.5 Research Questions....98
5.5.1 How can explainable AI techniques be used to evaluate and compare artificial intelligence models for binary PPG signal-quality classification in terms of accuracy and interpretability?....98
5.5.2 How effectively can a large language model transform technical, explainable AI outputs into actionable, user-facing facilitator prompts?....99
5.5.3 Do facilitator prompts increase the likelihood of users performing corrective actions to address system limitations as measured by participants' reported intention to record a single-lead ECG when presented after an unsuccessful measurement notification?....100
5.5.4 How can high-level automations, such as an AI-enabled PPG-based smartwatch hypertension screening system with a limited user-facing interface, be designed to communicate system limitations and support corrective user action?....101
5.6 Contributions....103
5.7 Limitations....107
5.7.1 Artificial Intelligence Pipeline....107
5.7.2 Generating the User-facing Prompts....107
5.7.3 Human Validation of the Prototype....109
5.8 Future Work....111
5.9 Conclusion....113
References....115
Appendix A LLM Evaluation Report....130

Appendix B Chapter 4 Study Materials 136
Appendix B.1 Recruitment Materials 136
Appendix B.1.1 Recruitment Email 137
Appendix B.1.2 Recruitment Poster 140
Appendix B.1.3 Recruitment LinkedIn Post 141
Appendix B.2 Pre-Study Session Materials 142
Appendix B.2.1 Information Consent Letter 142
Appendix B.2.2 Consent Form & Secondary Use of Data 146
Appendix B.2.3 Demographics and Screening Questionnaire 149
Appendix B.2.4 Study Session Booking Site 153
Appendix B.2.5 Reminder of the Booking Session 155
Appendix B.3 Study Session Materials 157
Appendix B.3.1 Researcher Study Guide 157
Appendix B.3.2 Participant Study Platform 164
Appendix B.3.3 Data Collection Form 168
Appendix B.4 Post-Study Session Materials 180
Appendix B.4.1 Feedback and Appreciation Letter 180
Appendix B.4.2 Participant Contact Details Form for Remuneration 183
Appendix B.5 Ethics Clearance Certificate 184

List of Figures

Figure 1. Graphical Abstract xxiii
Figure 2. Hypertension Notification Algorithm Flow Diagram 13
Figure 3. Correlation Matrix 26
Figure 4. DiCE counterfactual feature ranking for MLP [RandomizedSearchCV] 42
Figure 5. Interfaces in the Intervention-Group Scenarios 57
Figure 6. Existing interface 58
Figure 7. Aspects of the intervention that improved the participants' decision-making process 79
Figure 8. Participants' Perceptions of AI-Generated Videos 80
Figure 9. Recruitment Email 139
Figure 10. Recruitment Poster 140
Figure 11. Recruitment LinkedIn Post 141
Figure 12. Information Consent Letter 145
Figure 13. Consent Form 146
Figure 14. Declined Consent Message 147
Figure 15. Secondary Use of Data 148
Figure 16. Demographics and Screening Questionnaire - Demographics Questions 149
Figure 17. Demographics and Screening Questionnaire - Screening Question 150
Figure 18. Study Ineligibility Notification 150
Figure 19. Demographics and Screening Questionnaire - eHEALS Scale 151
Figure 20. Demographics and Screening Questionnaire - End of the Questionnaire 152
Figure 21. Study Session Booking Site 153
Figure 22. Booking Confirmation Message 154
Figure 23. Study Session Booking Reminder Email 156
Figure 24. Participant Study Platform - Welcome Screen 164
Figure 25. Participant Study Platform - Start of Scenario 164
Figure 26. Participant Study Platform - Video and Description of the Scenario 165
Figure 27. Participant Study Platform - Redirection to the data collection form 165
Figure 28.Participant Study Platform - Redirection to the prototype 166
Figure 29. Participant Study Platform - Redirection to the data collection form 166

Figure 30. Participant Study Platform - End of the Study 167
Figure 31. Feedback and Appreciation Email.....182
Figure 32. Participant Contact Details Form for Remuneration 183
Figure 33. Ethics Clearance Certificate.....185

List of Tables

Table 1. Health Technologies in Hypertension Care 10
Table 2. Cohort Characteristics 24
Table 3. Signal-Quality Distribution by Motion Condition 24
Table 4. Experiment 1. Predictive performance of the top full-feature models 41
Table 5. Structure of the Facilitator Prompt 45
Table 6. LLM Prompt 45
Table 7. Facilitator Prompts Generated by the Large Language Models 50
Table 8. Final Version of the Facilitator Prompts 52
Table 9. Participants' Perceived Causes of The Unsuccessful ECG Measurements 71
Table 10. Correct Identification of The Likely Contributing Factor in the Unsuccessful Measurement per Condition 72
Table 11. Immediate Intended Action before Interface Interaction 73
Table 12. Evaluation of the LLM-Generated Prompts following the QUEST Principles 132

List of Abbreviations

Area Under the Receiver Operating Characteristic Curve (AUROC)

Artificial Intelligence (AI)

Atrial Fibrillation (AFib)

Cross-Validation (CV)

Deep Learning (DL)

Digital Health Technology (DHT)

Digital Health Technologies (DHTs)

Diverse Counterfactual Explanations (DiCE)

Electrocardiogram (ECG)

Explainable Artificial Intelligence (XAI)

Fogg Behavior Model (FBM)

Generative Pre-trained Transformer (GPT)

Health Information Technology (HIT)

Heart Rate Variability (HRV)

K-Nearest Neighbors (KNN)

Linear Discriminant Analysis (LDA)

Local Interpretable Model-Agnostic Explanations (LIME)

Machine Learning (ML)

Mobile Health (mHealth)

Photoplethysmography (PPG)

Precision–Recall Area Under the Curve (PR-AUC)

Principal Component Analysis (PCA)

Radial Basis Function (RBF)

Receiver Operating Characteristic (ROC)

Retrieval-Augmented Generation (RAG)

SHapley Additive exPlanations (SHAP)

Support Vector Machine (SVM)

Volume of Oxygen (VO_2)

Artificial Intelligence Disclosure Statement

AID Statement: Artificial Intelligence Tools: All tools listed in this statement were used between November 2025 and June 2026. Tools used included Gemini in Google Colab by Google; Grammarly's AI writing assistance by Grammarly; Veo 3.1 by Google; ChatGPT Plus using GPT-5.5 by OpenAI; Llama 3.1:8B by Meta.

Methodology: Llama 3.1:8B and GPT-5.5 were used to generate and compare prompts using the approach described in the methodology. Videos generated with Veo 3.1 were also used as part of the methodological approach described in the methods;

Execution: Gemini in Google Colab was used to assist with debugging, support, and script refinement;

Visualization: ChatGPT Plus using GPT-5.5 in Thinking mode was used to generate images for visual materials included in or associated with the project, as disclosed on an individual basis;

Writing - Review & Editing: Grammarly's AI writing assistance was used in Microsoft Word to review and edit the thesis for grammar, clarity, conciseness, and academic tone.

The author reviewed and edited the AI-generated modifications as needed and takes full responsibility for the content of this thesis.

Graphical Abstract

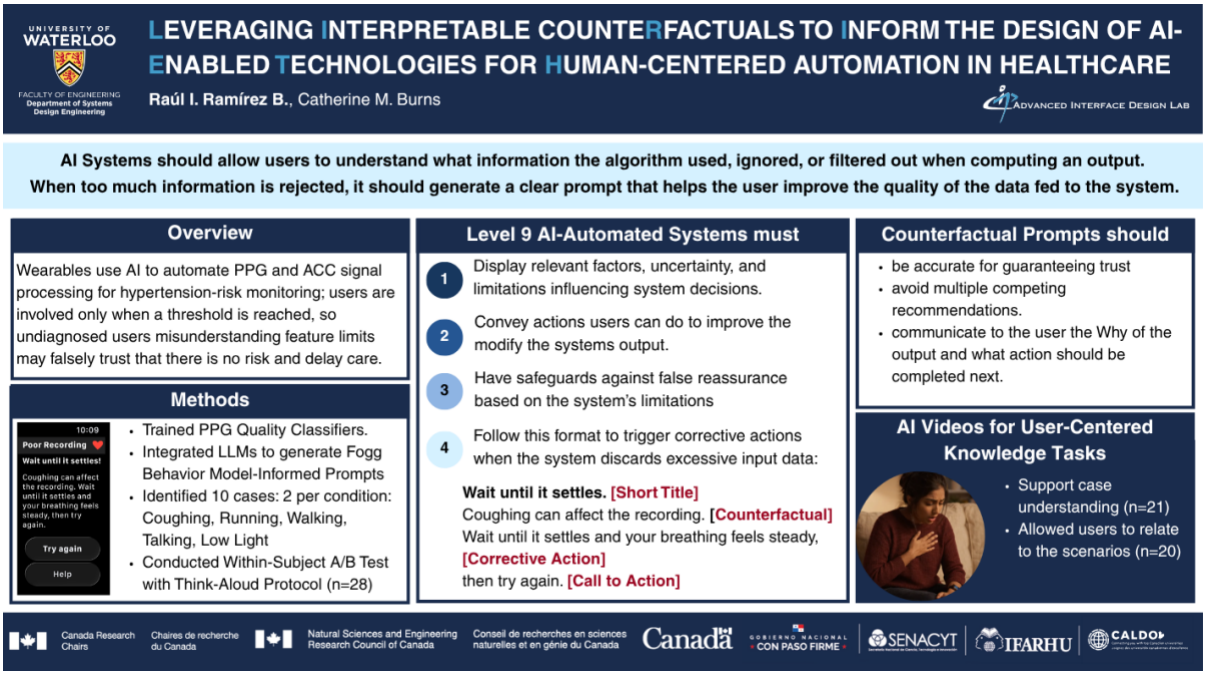


Figure 1. Graphical Abstract

Quote

"Ethical discernment cannot be limited to asking whether we are using a system for good or bad purposes; it must also examine how that system is designed and what vision of the human person and society is embedded in the data and models that guide it" (Leo XIV, 2026, para. 104).

Chapter 1
Introduction

Cardiovascular care encompasses the prevention, diagnosis, treatment, and long-term management of medical conditions affecting the heart and blood vessels (World Health Organization, 2025). Unhealthy diet, limited physical activity, tobacco, alcohol, hypertension, diabetes, and high blood lipids are factors that contribute to the risk of developing cardiovascular disease (World Health Organization, 2025).

Hypertension, a condition in which blood pressure exceeds 130/80 mmHg (Goupil et al., 2025), is particularly significant because it is often asymptomatic, undiagnosed, or undertreated (Ali et al., 2021). In Canada, 1 in 4 adults has been diagnosed with hypertension (Statistics Canada, 2024b), while approximately 1 in 5 adults with hypertension is unaware of their condition (Statistics Canada, 2024a). These challenges are exacerbated by pressures within the health care system, including a shortage of medical professionals responsible for the prevention, screening, and treatment of hypertension (Averbuch et al., 2024). Health Canada indicates that to address the demands of the health care system, the number of primary care providers must increase by 49% (Health Canada, 2025).

Blood pressure can vary due to factors such as stress, activity, and nutrition. International hypertension guidance recommends at-home monitoring (Goupil et al., 2025; Jones et al., 2025; McEvoy et al., 2024), as in-office measurements may lead to white-coat hypertension, in which blood pressure is high at the clinician's office but lower elsewhere (Franklin et al., 2013), or masked hypertension, in which readings are normal in the clinic but elevated outside the clinic (Pickering et al., 2007). These limitations have created a need for approaches that support continuous, context-sensitive cardiovascular monitoring outside traditional clinical settings.

Technology has been identified as a means of addressing these gaps and reducing the burden of the cardiovascular care system (Averbuch et al., 2024; Buyting et al., 2022; Zwack et al., 2023). However, high levels of automation may affect how users interpret the outputs of these systems. This thesis investigates how explainable AI (XAI), counterfactual explanations, large language models (LLMs), and behavior-informed interface design can be integrated to support more transparent and actionable communication between highly automated health systems and their users.

Health Information Technology (HIT) is defined as “the application of information processing involving both computer hardware and software that deals with storage, retrieval, sharing, and use of health care information, data, and knowledge, for communication and decision making” (Brailer, 2004, p.38). Recent technological advancements have necessitated a broader umbrella to reflect the expansion from the original standard of care, where clinicians had full control over the patient’s health, to a different paradigm with increased patient participation in their health care (Mesko et al., 2025). Digital Health Technologies (DHTs) encompass all systems that aim to improve users' health and wellness (Mangion & Piller, 2025). These include mobile health (mHealth), health information technology (HIT), wearable devices, telehealth and telemedicine, and personalized medicine (Center for Devices and Radiological Health, 2024).

Digital health technologies have enhanced patient care by making data more accessible, particularly in supporting wellness and cardiovascular health. For instance, smartwatches contribute substantial data that clinicians and patients can generate, interpret, and integrate during clinical visits (Mak et al., 2025). They enable continuous monitoring of metrics such as heart rate, irregular rhythm, VO_2 max/cardio fitness, ECG/AFib detection, activity minutes, sleep, and Heart Rate Variability (HRV) outside an in-office context (A. Hughes et al., 2023) by leveraging algorithmic models to detect patterns that might be relevant to health. This is especially important for hypertension, which is often asymptomatic, potentially deadly, and requires long-term monitoring.

The increase in the volume of patients' health data since the internet became a source of health information has significantly changed how doctor-patient interactions occur (Mesko et al., 2025; Sieck & Huerta, 2019). For patients, this shift has enabled better understanding and management of health conditions (McMullan, 2006; Tan & Goonawardene, 2017). For clinicians, however, it has increased the complexity of cases during medical consultations as they address the health information patients bring. Specifically, the literature indicates that during consultations, clinicians face time pressure, must correct misinformation, may have difficulty assessing the reliability of patient-provided sources (Lu & Schulz, 2024), and may fear appearing dismissive or insufficiently informed (Ahluwalia et al., 2010). The integration of artificial intelligence (AI) into DHTs continues to transform these technologies into AI-enabled systems that support monitoring, screening, risk prediction, and decision-making, thereby altering how clinicians engage with their patients (Chen, 2025; A. M. Hughes et al., 2026).

In the health care context, the implementation of AI-enhanced systems has been constrained by limitations in transparency, explainability, equity, and fairness (Shevtsova et al., 2024). In particular, algorithmic limitations can affect understanding, trust, and acceptance of these technologies, as data used to train the models may vary across populations, physiological conditions, devices, and contexts of use, potentially exacerbating disparities (Dankwa-Mullan, 2024). These concerns about AI implementation are further heightened by the complex decision-making processes of the algorithms, which users may find difficult to understand.

When these systems deliver results without explanations, they create additional risks, such as users relying too much on the output, misinterpreting its significance, or believing that the absence of an alert indicates no health risk. Although there have been advancements aimed at explaining how algorithms produce their outputs through approaches such as explainable AI (XAI), these methods often fail to provide actionable information to support appropriate understanding, use, and decision-making (Hassan et al., 2024).

This is particularly important for AI-enabled systems, where users may lack the clinical or technical expertise to interpret physiological values or algorithmic outputs. Whereas traditional DHT interfaces were designed to support the technology's operation, AI-enabled systems must adapt their interfaces to help users interpret and act on probabilistic outputs. In patient care, this transition marks a shift from patients examining and comparing multiple online resources and asking clinicians questions to receiving synthesized answers that appear coherent, personalized, and medically confident.

These challenges are especially important in wearable and remote health-monitoring systems, where AI performance relies not only on the algorithm but also on the device's sensors, signal quality, the environment, and user behavior. Photoplethysmography, commonly referred to as PPG, is an optical sensing technique that detects changes in blood volume in the microvasculature beneath the skin by directing light onto the skin surface and using photodetectors to measure the reflected light (Tamura et al., 2014). Nevertheless, its signal quality can be compromised by several factors, including motion (running, walking, talking, coughing), user error in device placement, and skin tone (Bent et al., 2020; Singh et al., 2024).

These factors can create out-of-distribution conditions: real-world data points that differ from the data on which the model was trained or validated. Signal-quality classification is therefore an important technical step, because the reliability of any PPG-enabled feature depends directly on whether the physiological data is sufficiently reliable to enter the processing pipeline.

A major breakthrough in hypertension care occurred in September 2025, when Apple launched a Hypertension Notification feature that passively analyzes randomly sampled PPG signals over a 30-day period to identify patterns that may indicate a high risk of hypertension. If signs indicative of hypertension are detected, an algorithm subsequently alerts the user. However, this feature does not display any feedback to the user, including, but not limited to, the amount of data the model uses to produce the outcome or the systolic or diastolic blood pressure readings, unlike traditional cuff-based devices (Apple Inc., 2025; U.S. Food and Drug Administration, 2025a). This represents level 9 of automation in the Parasuraman et al. (2000) model, where the system only informs the human if it (the system) chooses to do so.

The Human Validation Study for this feature, submitted by Apple to the U.S. Food and Drug Administration (FDA), indicates that 58.8% of individuals with undiagnosed hypertension did not receive the notification (U.S. Food and Drug Administration, 2025b). The report further indicates that the feature is intended to prompt users to seek clinical validation; however, this represents a human-AI interaction problem, as high levels of automation can lead to misuse of technology (Parasuraman & Riley, 1997). This feature has the potential to falsely reassure individuals with undiagnosed hypertension, who can then be discouraged from seeking care due to a lack of any feedback until a decision is made by the algorithm, potentially affecting the cardiovascular health of undiagnosed patients (Cohen et al., 2026).

Therefore, in this thesis, we aim to evaluate user-facing prompts generated from model counterfactual explanations to prevent data produced under out-of-distribution conditions from entering the AI pipelines of user-facing AI-enabled systems by prompting users to take corrective action before the data is processed. PPG-based hypertension risk screening is used as a case example to propose a human-centered framework for designing and developing AI-enabled health care systems, emphasizing that, for health systems or device features such as hypertension screening, automation levels should not eliminate users' access to data and the decision-making process.

To prompt corrective actions, persuasive design literature is leveraged, particularly the Fogg Behavior Model, which states that for a behavior to occur, the user must be motivated, have the ability, and be prompted to perform the intended behavior (Fogg, 2009). Within this model, a facilitator prompt is appropriate when the user is already motivated but has limited ability to perform the behavior.

Because PPG features are passively collected, ECG recordings are used as a practical proxy for studying a user-initiated behavior and responsiveness to prompts. A user may be motivated to record an ECG because they feel symptoms, want reassurance, or believe the situation is important to their health. However, their ability to act may be limited if they do not understand why the measurement failed, how to fix the problem, whether to try again immediately, wait, rest, adjust the watch, or seek help. For this reason, this study structures the smartwatch message as a user-facing facilitator prompt: a prompt shown directly to the user, designed to make the target behavior easier to perform.

High levels of automation introduce additional barriers for patients, making it harder to make informed health decisions in consultation with health care providers. Instead, AI-enabled systems should facilitate user-system interaction by enabling users to verify detected patterns in their signal data. Additionally, when data is compromised for a prolonged period, the system should clearly notify users and offer actionable prompts to encourage behavior or measurement adjustments.

To support this approach, this thesis proposes using a large language model (LLM) orchestrator to identify factors affecting signal quality and translate them into user-facing facilitator prompts. The thesis also examines whether the redesigned smartwatch ECG interface supports user engagement and intended action after unsuccessful measurements. Engagement is defined here as noticing the notification, understanding the reason for the failed attempt, understanding the recommended next step, and feeling capable of acting on that recommendation. DiCE outputs are treated as model-relative counterfactuals. They identify feature changes that would alter a classifier's prediction based on learned associations; they do not, by themselves, establish that a feature caused signal degradation or that changing it will improve the actual signal. Causal language is reserved for relationships supported by independent domain evidence, causal analysis, or deliberately manipulated study conditions.

1.1 Research Questions

The primary goal of this thesis is to evaluate whether AI-generated facilitator prompts can improve users' intended corrective actions to address certain measurement limitations of a highly automated system that decides when to bring a human in the loop. To achieve this, the first part of this thesis presents a PPG-based modeling study using explainable AI to evaluate and interpret binary PPG signal quality classification models, then integrates a large language model to transform the model counterfactuals into condition-linked explanations embedded in actionable facilitator prompts that can be understood by the user. The second part is an ECG-based human evaluation study that evaluates the transferability of the interaction and corrective-feedback design by assessing whether these prompts increase the likelihood that users intend to perform a corrective action, operationalized as recording a single-lead ECG after an unsuccessful measurement notification. To achieve this goal, the following research questions will be addressed throughout the thesis.

Research Question 1 (RQ1): How can explainable AI techniques be used to evaluate and compare artificial intelligence models for binary PPG signal-quality classification in terms of accuracy and interpretability?

Research Question 2 (RQ2): How effectively can a large language model transform technical, explainable AI outputs into actionable, user-facing facilitator prompts?

Research Question 3 (RQ3): Do facilitator prompts increase the likelihood of users performing corrective actions to address system limitations as measured by participants' reported intention to record a single-lead ECG when presented after an unsuccessful measurement notification?

Research Question 4 (RQ4): How can high-level automations, such as an AI-enabled PPG-based smartwatch hypertension screening system with a limited user-facing interface, be designed to communicate system limitations and support corrective user action?

1.2 Thesis Outline

The thesis is organized into five chapters. After the introduction:

Chapter 2 reviews the background relevant to this thesis and identifies the knowledge gap addressed by the thesis. It discusses health technologies, AI-enabled wearables, and the use of PPG measurements for hypertension risk assessment. The chapter also examines user interaction with automated health systems, focusing on trust, omission errors, and persuasive design. Finally, it considers the limitations of AI in health care and explores how counterfactual explanations and large language models may support transparency and user understanding.

Chapter 3 addresses RQ1. The chapter presents an explainable AI pipeline designed to improve PPG signal quality through user corrective actions. Within the pipeline, PPG quality classifiers are trained and evaluated using SHAP, LIME, and DiCE to examine recording conditions associated with the model classifications and generate counterfactuals identifying input changes predicted to shift the classified signal quality.

The chapter also addresses RQ2. Llama 3.1:8B is integrated into the pipeline as a translation layer, converting these technical counterfactual outputs into clear, user-facing facilitator prompts. Guided by the Fogg Behavior Model, these prompts are designed to encourage specific, feasible, and actionable user responses. The chapter presents the generated prompts and evaluates them using the QUEST principles for human evaluation of LLMs in health care. Furthermore, the outputs are compared with GPT-5.5 across five PPG acquisition conditions: walking, coughing, changing-light, talking, and running. Finally, this chapter presents the revised facilitator prompts based on model-informed counterfactual explanations selected for inclusion in the smartwatch interface and for subsequent evaluation in Chapter 4.

Chapter 4 addresses RQ3 by presenting the ECG-based human validation study conducted to assess the transferability of the interaction and corrective-feedback design developed in Chapter 3. The model-informed facilitator prompts were incorporated into the proposed Apple Watch-style Figma interface, which was compared with the existing ECG interface across ten unsuccessful-measurement scenarios. In this counterbalanced, within-participant mixed-methods study with 28 participants, the primary questionnaire outcome was reported intended action, operationalized as reported retry intention.

Finally, Chapter 5 concludes the thesis by discussing the findings of Chapters 3 and 4, addressing the four research questions, and presenting the ACT-AI framework in response to RQ4. Using AI-enabled, PPG-based smartwatch hypertension screening as a motivating example, the framework guides the design of highly automated AI-enabled systems that decide when to involve users by communicating system limitations that may arise from out-of-distribution conditions affecting their AI pipelines. The chapter also presents the thesis's contributions, limitations, and directions for future research.

Chapter 2
Background

This chapter examines the existing literature and provides the essential background for understanding the research reported in this thesis. It begins by reviewing the role of health technologies and AI-enabled wearables in hypertension care, with particular attention to the use of photoplethysmography (PPG) measurements in hypertension risk scoring. The chapter then explores how users interact with automation, the risks of omission errors, and the role of trust in digital health technologies. It also examines persuasive design through the Fogg Behavior Model, reviews the limitations of AI in health care, and introduces causal explanations, counterfactual algorithms, and their possible integration with large language models to improve transparency, interpretability, and user understanding of AI.

2.1 The Human-Centered Problem of Wearable Automated Systems for Hypertension Screening

2.1.1 Health Technologies in Hypertension Care

To enhance hypertension care, multiple health technologies have been designed for automated home blood pressure measurements, blood pressure tracking, ambulatory blood pressure monitoring, remote control of hypertension treatment, and alerts for high blood pressure readings (Parati et al., 2023; Shimbo et al., 2020). Table 1 summarizes the health technologies available for hypertension care, the information each system presents to the user, and the limitations of the technology.

Table 1. Health Technologies in Hypertension Care

Technology	Information Presented to the User	Limitations
Upper-arm home blood pressure monitors[1]	Systolic Blood Pressure, Diastolic Blood Pressure, Pulse.	Accuracy depends on device, correct cuff size, and positioning.
Ambulatory blood pressure monitoring[2]	- No information is presented to the patient. - The device repeatedly measures the blood pressure over 24 hours and presents it to the clinician.	Limited availability and it might disrupt daily activities.
Mobile health applications for hypertension[3]	Blood pressure logs, trends, and medication reminders.	- Require regular, long-term use by patients. - Patients might misinterpret isolated readings.
Digital lifestyle and education tools[4]	Sodium intake, weight, physical activity, medication reminders, lifestyle goals and educational guidance for controlling blood pressure.	- User engagement decreases over time. - The tools usually require feedback from a medical professional to increase their effectiveness.
CureApp HT Digital Therapeutic[5]	Structured lifestyle guidance, behavior-change tasks, and hypertension self-management support.	It increases clinicians' monitoring and data-management burden as they have to monitor the patients' progress.

[1] (American Medical Association, 2026; George & MacDonald, 2015)
[2] (Q.-F. Huang et al., 2021)
[3] (Joshee et al., 2026)
[4] (Alnooh et al., 2022)
[5] (Kario et al., 2021)

2.1.2 Artificial Intelligence-Enabled Wearable Technologies for Hypertension Care

Recent advancements in AI have led to new approaches that do not rely on the inflatable cuff to support hypertension management. To better understand what these approaches entail and their limitations, it is helpful to review some key technical concepts, starting with the definition of artificial intelligence.

Russell and Norvig (2003) define AI as the field that studies agents that perform actions based on the mapping of percept sequences. McCarthy (2007) defines it as the science and engineering that focuses on making intelligent machines, especially computer programs, without being limited to biologically observable methods. And Bergmann (2025) defines it as a set of computational methods that learn from data to solve core mathematical problems by mapping inputs to outputs through a mathematical model.

A key distinction to highlight is among the concepts of artificial intelligence (AI), machine learning (ML), and deep learning (DL), which are often used interchangeably. AI is a broad concept encompassing techniques that enable machines to mimic human behavior. ML is a subset of AI that is based on the idea that systems can learn from data by identifying patterns and making decisions with minimal human intervention. DL is a subset of machine learning that, inspired by the structure of the human brain, uses artificial neural networks to identify complex patterns in data (Kufel et al., 2023; Sarker, 2021a).

Depending on the learning paradigm used, machine-learning approaches can be categorized as supervised, unsupervised, or reinforcement learning. Supervised learning refers to the training of a model on data with known inputs and labels, or on data with expected outcomes. Unsupervised learning, on the other hand, applies computational methods to find patterns, dependencies, and correlations in the data. When models are trained to evaluate the environment and take the action that yields the greatest reward, this is considered reinforcement learning (Bergmann, 2025).

Another way to categorize AI models is by the core mathematical problems they aim to solve: classification, regression, clustering, dimensionality reduction, anomaly detection, optimization and search, decision-making, and generative modeling (Sarker, 2021b).

Wearable health devices have become a vast and increasingly valuable source of health data, enabling continuous, non-invasive collection of physiological data. Photoplethysmography (PPG) is one of the most widely used technologies enabling continuous monitoring of cardiovascular health metrics, including heart rate and blood oxygen saturation (Almarshad et al., 2022).

PPG accomplishes this by detecting changes in blood volume in the microvasculature beneath the skin using optical techniques: light is directed onto the skin surface, and photodetectors detect the reflected light (Tamura et al., 2014). This technology offers significant clinical benefits by facilitating remote patient monitoring, continuous data collection, early identification of cardiovascular irregularities, and chronic disease management.

These advantages are especially valuable both for patients who face barriers to receiving conventional care and for health care providers who face limited resources to meet the patient's needs. However, the reliability of PPG signals remains a major limitation to their application in health contexts (Almarshad et al., 2022; Bayoumy et al., 2021). Signal quality can be compromised by several factors, including motion artifacts, user error in device placement, and skin tone, leading to unreliable health insights that can affect clinical decision-making (Bent et al., 2020; Singh et al., 2024).

Such limitations can adversely affect patient outcomes, particularly in settings where these devices are used for chronic disease monitoring or risk assessment (Almarshad et al., 2022; Weng et al., 2024; Yu et al., 2022). Apple's PPG-based hypertension notification feature exemplifies a risk assessment feature in hypertension care. Unlike competitors such as Huawei and Samsung, which have developed blood pressure measurement models, the "hypertension notification" is a machine-learning-based feature that identifies patterns in PPG recordings over 30-day windows.

As shown in Figure 2, this feature operates through a three-step process: first, features are extracted into generalizable characteristics using a DL algorithm that was not trained on any blood pressure measurements or hypertension-related data. Second, it uses a machine learning algorithm to score the risk of hypertension based on features extracted by the DL algorithm. Finally, a separate algorithm determines if the notification should be presented to the user. This last algorithm excludes data collected during sleep and aggregates and averages the remaining scores, which are then compared against a predetermined threshold to determine whether to send the notification (Apple Inc., 2025).

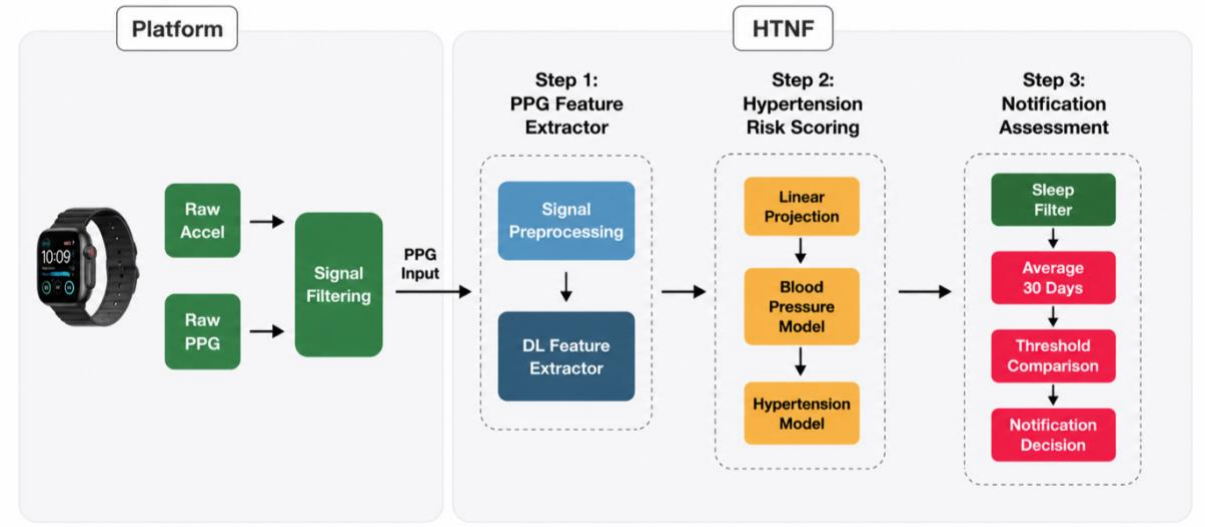


Figure 2. Hypertension Notification Algorithm Flow Diagram

However, this feature reportedly failed to notify 59% of people with hypertension, whereas 8% of individuals in the study without hypertension received a notification of possible hypertension (Cohen et al., 2026). Although the manufacturer states that the point on the receiver operating characteristic (ROC) curve was selected to prioritize high specificity, it remains unclear whether the relatively low sensitivity is due to limitations of the algorithm itself, difficulties in accurately recording usable data from users, the inherent difficulty of inferring hypertension risk from the opportunistic way the data was collected and preprocessed, or characteristics in the data that distinguish false-negative cases from correctly detected cases.

2.1.3 AI Approaches Toward Improving PPG Measurements

AI is being implemented to address limitations that degrade the quality of PPG measurements, such as motion, weak sensor-skin contact, ambient light interference, or skin tone. Algorithms are developed to improve PPG measurement quality using unusable-segment detection (Argüello-Prada & Castillo García, 2024; Goh et al., 2020), signal quality classification (n.d.), motion suppression (M. Lee et al., 2025; Sawangjai et al., 2025), and denoising (Kwon et al., 2022).

2.2 Human Factors and Risk Assessment Capabilities of Smartwatches

2.2.1 Human Interaction with Automation

Automation can be defined as a system that performs information acquisition, analysis, decision, and action selection, and/or action implementation, tasks that were previously performed by a human (Parasuraman & Riley, 1997). Based on this definition, automation can be categorized into 10 levels according to the degree of human involvement it relinquishes, with level 1 representing low automation and level 10, high automation (Parasuraman et al., 2000).

Level 1 implies that the automation provides no assistance and that the human must make all decisions. If the automation offers a complete set of decision or action options that the human must consider, then it is a level 2. If the automation further narrows the options for the human and presents a few options (level 3) or only one alternative (level 4), it increases the system's involvement in the decision-making process.

When the system performs the action with human approval, it is classified as level 5. If the system decides to perform the action but allows the human to veto it within a restricted time before execution, it is classified as level 6 automation. Level 7 includes systems that execute actions automatically and then inform the human after execution. Level 8 informs the human that the automatic action has been executed only if the system is explicitly asked to do so. Level 9 implies that the system performs the function and decides whether the human should be informed. Finally, level 10 is an automation that spans the entire continuum of information processing; the system does everything, keeping the human out of the loop.

2.2.2 Omission Errors and Trust in Digital Health Technologies for Hypertension Care

Level 9 automated systems, like the Apple hypertension notification feature, decide whether to notify the user when a threshold is reached; if not, the user remains unnotified. This can lead to omission errors, where users fail to act because the automation did not alert them (Parasuraman & Manzey, 2010).

Endsley and Kiris (1995) explained that high levels of automation, like this one, might lead to a weakening of the user's ability to detect failures in the system, understand the current state, and take over effectively when the automation fails. In hypertension management, because missed notifications can discourage care-seeking, the feature requires a high level of sensitivity.

Health care systems must be designed and evaluated to balance sensitivity, the proportion of people with the condition who test positive, and specificity, the proportion of people without the condition who test negative (Asan et al., 2020).

Trust is defined as the user's perception that an automated agent will help achieve task goals under conditions of uncertainty and vulnerability. When the system is authorized to make or implement decisions with limited direct human involvement, it is directly influenced by that balance, and it contributes to the use and adoption of the technology by the users (J. D. Lee & See, 2004).

2.3 Adherence and Persuasive Design for Digital Health Technologies

In addition to trust, adherence is another factor that significantly affects the effectiveness of digital health technologies, like smartwatches. Although smartwatches show promise in improving users' health, their effectiveness still depends on users' willingness to consistently engage with and use them (Mercer et al., 2016). This is particularly important in hypertension care, as it is a condition that requires early detection and long-term management.

Smartwatches use prompts, reminders, and haptic cues to trigger behavior at specific moments. When worn, these cues can be delivered in a timely manner without requiring further user intervention. To better understand how smartwatches should be designed to influence users' attitudes or actions without coercion, this section reviews models of persuasive design and behavior change.

The health belief model explains why a user may value a health behavior by focusing on perceived risks, benefits, barriers, and cues to action (Alyafei & Easton-Carr, 2024); it does not address how design can lead to action. The theory of planned behavior offers an explanation of users' intentions through attitudes, subjective norms, and perceived behavioral control (Ajzen, 1991), but it fails to address situations in which users may already intend to behave in a certain way yet fail to act unless the circumstances are timely and easy to act on.

The persuasive systems design model provides a broader framework for classifying features into primary task, dialogue, credibility, and social support (Oinas-Kukkonen & Harjumaa, 2009). COM-B and the Behavior Change Wheel extend this analysis by examining whether interventions consider the user's capability, opportunity, and motivation for sustained behavior change (Michie et al., 2011). Other models, such as self-determination theory (Deci & Ryan, 1985) and nudge theory (Thaler & Sunstein, 2021), further explain how systems can influence behavior by shaping motivation or simplifying the decision-making process.

2.3.1 Fogg Behavior Model

As the literature presents conflicting views on the effectiveness of digital health technologies (Golbus et al., 2026; Sun et al., 2024), the Fogg Behavior Model was chosen because it not only explains how behavior can be influenced but also why the trigger for a specific behavior (a prompt) can succeed or fail. The Fogg Behavior Model shows that a user should have sufficient motivation and ability and receive an effective prompt for a behavior to occur.

The model considers three core motivators, each with two sides: pleasure/pain, hope/fear, and social acceptance/rejection. The second factor in the model is ability, which includes six elements: time, money, physical effort, brain cycles, social deviance, and non-routine. If any of these six elements is required and missing for a behavior to occur, the ability to perform that behavior is compromised.

The third factor is the prompt. The model defines three types of prompts: facilitator, when the user has high motivation but lacks the ability to perform a target behavior; spark, when there is high ability to perform a behavior, but there is low motivation for the user; and signal, where both ability and motivation are sufficient for a behavior to happen (Fogg, 2009).

2.4 AI Limitations and the Need for Counterfactual Explanations

As AI systems became increasingly available, it became challenging to understand how their outputs were computed, making it difficult to contest or act on them, thereby hindering the acceptance of these technologies in complex domains like health care. Explainable AI techniques like SHAP and LIME often describe the features that influence the prediction, but their outputs are difficult for lay users to understand. Model counterfactuals aim to address this gap by showing the minimal input changes that would have led to a different prediction, but they do not by themselves establish causal explanations.

Causal counterfactual reasoning has been studied since 1748. The first definition was offered by David Hume, who tried to define causation as "if the first object had not been, the second never had existed." This later became an early formulation of counterfactual reasoning. In 1973 and 1986, David Lewis proposed the first elaborated counterfactual theory of causation, in which he argued that causal dependence can be understood by considering the possibility that an event would have occurred in the closest possible "world" in which the event is assumed not to have occurred.

These conceptions of counterfactual reasoning were later translated into the context of explainable AI, offering a practical method for explaining black-box automated decisions. Wachter et al. (2017) aimed to explain why the output was reached, how to contest the decision if the outcome is undesired, and what change is needed in order to receive a desired result in the future, without explaining how the algorithmic function worked.

2.4.1 Counterfactual Algorithms

Guidotti (2024) benchmarked the counterfactual algorithms based on the strategy they adopted to produce the counterfactuals: Brute Force (BF), Optimization (OPT), Heuristic Search Strategy (HSS), Instance-Based (IB), and Decision Tree-based (DT). Using minimality, actionability, stability, diversity, discriminative power, and running time (Mothilal et al., 2020; Verma et al., 2020) as evaluation criteria, the review found that no counterfactual explainers could guarantee all of these properties at once.

DiCE model counterfactuals are examined further in this thesis because the algorithm's approach to counterfactuals satisfied most of the properties listed above, provided diverse explanations, and was particularly effective with tabular data. The main risk associated with this approach to generating counterfactuals is that generating diverse counterfactuals does not necessarily guarantee feasibility, and the counterfactuals do not automatically have causal interpretations. LLMs have been proposed as a way to address challenges related to the understandability and practical feasibility of implementing counterfactuals.

2.4.2 Counterfactual Applications and LLM Integrations

Studies in which counterfactuals are presented to human participants suggest that these explanations can improve aspects of user understanding, perceived usefulness of the AI system, and decision quality. They show that benefits vary case by case and depend on explanation format, task domain, decision stakes, user expertise, AI correctness, and whether explanations support corrective actions rather than presenting minimal changes (Celar & Byrne, 2023; Le et al., 2023; M. H. Lee & Chew, 2023).

In general, for AI-assisted decision-making tasks, Wang and Yin (2021) found that counterfactuals improved objective understanding of the AI output, but only on tasks where people had some domain expertise. A separate study found that counterfactual explanations improved participants' prediction accuracy compared with no explanation, and that explanations involving categorical features were easier to understand than those involving continuous features (Warren et al., 2022). Similarly, Celar and Byrne (2023) found that counterfactual reasoning was viewed as more helpful than causal explanations, but it did not consistently improve users' ability to predict AI decisions.

In clinical decision support, M. H. Lee and Chew (2023) found that counterfactual explanations reduced overreliance on incorrect AI outputs by 21%, while Le et al. (2023) found that counterfactual explanations of model confidence improved users' understanding of and trust in the AI model. The findings also show that counterfactual explanations work best when they are actionable, feasible, and easy to process.

In consumer-facing and recourse settings, Ramon et al. (2021) found that users generally did not prefer counterfactual explanations, except when outcomes were negative, such as loan denial. However, Cau and Spano (2026) found that counterfactual explanations were perceived as less understandable, even though they improved accuracy and reduced cognitive load when the models' predictions were correct.

Recent studies on LLM-based and conversational counterfactual explanations suggest that these systems may make the outputs more personalized, interactive, and accessible, but they also introduce new risks. Fredes and Vitria (2024) showed that LLMs can translate sets of counterfactual examples into plain-language action explanations, although they noted the need for further datasets and human evaluations. Narendra and Chatterjee (2026) proposed the CLEaR system for dynamic, personalized counterfactual generation based on user constraints and preferences. Finally, Tešić and Hahn (2022) showed that counterfactual explanations can lead users to infer real-world causality from AI-used features, although warnings about correlation versus causation reduced this effect.

2.5 Knowledge Gap

This chapter examined how AI-enabled automation is reshaping human involvement in health-related decision-making, using Apple's hypertension notification feature as a case study. In this feature, the decision to alert the user to a potential health concern is largely delegated to an algorithm, raising important questions about whether individuals with undiagnosed hypertension may be falsely reassured by the absence of a notification and, as a result, less likely to seek medical care.

The chapter first considered the technical foundations and limitations of the photoplethysmography (PPG) technology on which the hypertension notification feature is built. It discussed how algorithmic approaches are used to improve signal quality and compensate for the constraints of wearable sensing, while also highlighting the limited role currently assigned to human action in addressing these limitations.

Persuasive design theory was then reviewed, as the effectiveness of smartwatch-based health features depends not only on technical performance but also on user adoption, adherence, and sustained engagement. The broader limitations of AI systems and how these limitations may affect users' willingness and ability to adopt and adhere to wearable health technologies were also examined.

This discussion established the relevance of model-informed counterfactual explanations for improving users' understanding of AI-driven health features. Counterfactual algorithms, their application areas, and recent advances were reviewed, including the use of large language models to address challenges related to feasibility, interpretability, and the translation of model-informed explanations into lay language.

However, a knowledge gap remains in understanding how counterfactual explanations for AI-enabled wearable health systems can be generated and communicated in ways that are feasible, actionable, and understandable to users. Existing counterfactual explanation research has shown that model counterfactuals can improve users' understanding, trust, satisfaction, and decision-making in some contexts, but that their effects depend on presentation format, domain familiarity, user expertise, and explanation complexity. Moreover, counterfactual explanations may pose risks if users misinterpret algorithmic associations as real-world causal relationships or place unwarranted trust in explanations that appear meaningful yet do not accurately reflect the system's limitations.

This gap is particularly important in health contexts in which AI-enabled systems are being implemented, and users may need to respond to uncertain or limited AI outputs without having the technical knowledge required to understand why a system failed, why a notification was not generated, or what corrective action is possible. In the case of Apple's hypertension notification feature, users with undiagnosed hypertension may be falsely reassured by the absence of a notification, especially if they do not understand the limitations of photoplethysmography-based sensing, algorithmic classification, and passive monitoring. Therefore, the challenge is not only to generate model-informed explanations from model counterfactuals but also to translate them into language that supports appropriate interpretation and action selection.

Recent work suggests that large language models may help address this problem by interpreting model counterfactuals, summarizing sets of counterfactual examples, and supporting conversational or personalized recourse. However, LLM-mediated, model-informed explanations also pose risks, including hallucinations, bias, overreliance, and the potential to produce explanations that are not faithful to the underlying system. For this reason, behavior models may be needed to guide the design of counterfactual explanations that are technically valid and aligned with users' motivation, ability, and reported corrective-action intentions.

Accordingly, this thesis addresses the gap between model counterfactuals and understandable, actionable, model-informed outputs for users of wearable health technologies. It investigates how system limitations can be communicated through actionable, counterfactual-based prompts that support effective error recovery, improve understanding of AI-driven health features, and guide users toward appropriate corrective action when system limitations or errors arise. This aligns with the findings of Abdelaal et al. (2024) that identified the need for user evaluation of explanations to develop transparent and reliable AI systems in wearable health contexts.

Chapter 3
Designing Prompts to Assist Users in Taking Corrective Steps Toward Enhancing PPG Signal Quality

The reliability of photoplethysmography (PPG)-based measurement depends heavily on the conditions under which the signal is acquired. Motion, unstable sensor contact, sensor pressure variations, user behavior, and environmental factors can compromise PPG waveform quality and reduce the reliability of downstream AI predictions. For this reason, poor acquisition conditions can compromise system performance before an algorithm produces a classification.

The challenge, therefore, extends beyond detecting whether a PPG signal is of good or bad quality. An AI-enabled system should also help identify which aspects of the recording process contribute to poor signal quality and whether those aspects can be modified. In this chapter, we propose a pipeline that uses DiCE counterfactual explanations to support corrective user actions when factors related to PPG signal recording result in bad-quality recordings. The approach generates facilitator prompts that translate model outputs and counterfactuals into condition-linked feedback for users. By embedding the Fogg Behavior Model within the design of the prompts, it frames causal explanations as specific, feasible actions that may change the model's output during data acquisition.

To implement this approach, Llama 3.1:8B, a large language model, was integrated as an orchestration layer within a pipeline of PPG quality classifiers. Its role is to convert model-derived counterfactuals into user-facing causal explanations that communicate which input changes would shift the model toward a good-quality prediction and possible corrective actions. The resulting explanations were evaluated using the QUEST principles for human evaluation of LLMs in health care. Pipeline outputs were also compared with GPT-5.5 responses to the same prompt across 10 cases representing five acquisition conditions: walking, coughing, changing light, talking, and running, with running cases derived from the walking condition.

3.1 Overview of the AI Pipeline

The AI pipeline uses supervised learning to classify tabular data associated with expert-labeled PPG recordings as either good or bad quality and to explain bad-quality classifications through counterfactual analysis. Rather than evaluating models only by their ability to distinguish between acceptable and unreliable recordings, the pipeline examines which signal features would need to change for a bad-quality prediction to shift toward a good-quality prediction.

This design supports two related objectives. First, it enables the comparison of PPG quality classifiers based on both predictive performance and explanatory usefulness. Second, it links bad-quality classifications to interpretable acquisition-related conditions, such as motion, sensor pressure, lighting variation, or user behavior, rather than treating them as unexplained system failures. In the context of smartwatch-based health features, such as PPG-derived hypertension notifications, this explanatory layer is important because unreliable recordings can undermine the trustworthiness of downstream outputs. By connecting model classifications to modifiable recording conditions, the pipeline contributes to the broader goal of designing highly automated systems that provide safer, more actionable, and more user-centered feedback.

3.2 The Dataset

This pipeline uses version 2.0 of the Brno University of Technology Smartphone PPG Database (BUT PPG) (Nemcova et al., 2024) to train binary PPG quality classifiers. The dataset includes 3,888 ten-second PPG signals recorded with Xiaomi Mi9 and Huawei P20 Pro smartphones, along with corresponding ECG recordings. Most signals are accompanied by accelerometric (ACC) data and annotations for blood pressure, blood oxygen saturation, and glycemia. As presented in Table 2, a total of 50 subjects (25 female and 25 male, ages 19-76) participated in the recordings. Each participant was recorded multiple times under varying conditions, such as rest and movement, and with either a finger or an ear PPG sensor.

Table 2. Cohort Characteristics

Cohort Characteristics	
n	50
Sex	
Female	25
Male	25
Age (mean (SD))	33.14 years (16.74)
Good Quality	830
Bad Quality	3,058

Table 3 shows the distribution of signal-quality labels across the derived motion-condition variables. The largest group of recordings occurred during rest, with 2,724 total cases, including 2,059 bad-quality and 665 good-quality cases. The next-largest condition was changing light, with 288 total cases, of which 256 were bad quality and 32 were good quality. Moving finger or ear on lens and walking each had 156 total cases, with 134 and 137 bad-quality cases, respectively. The remaining conditions, higher pressure, coughing, laughing, and talking, each had 144 total cases, with bad-quality percentages ranging from 79.86% for higher pressure to 86.81% for coughing. No motion condition had missing quality labels. The number of unique participants varied by condition, with rest, moving_finger_ear_on_lens, and walking including all 50 participants, while changing_light, higher_pressure, coughing, laughing, and talking each included 38 participants.

Table 3. Signal-Quality Distribution by Motion Condition

Motion condition	Bad-quality cases	Good-quality cases	Total cases	Bad quality % among labeled	Unique participants
rest	2,059	665	2,724	75.59	50
changing_light	256	32	288	88.89	38
moving_finger_ear_on_lens	134	22	156	85.90	50
walking	137	19	156	87.82	50
higher_pressure	115	29	144	79.86	38
coughing	125	19	144	86.81	38
laughing	122	22	144	84.72	38
talking	121	23	144	84.03	38

3.2.1 Data Missingness

An analysis of data missingness identified 192 missing values (0.49% of the dataset) across five features: height, blood pressure, glycemia, oxygen saturation (SpO_2), and ACC data. However, this data is structurally missing, not missing at random. As this dataset (BUT PPG v2.0.0) is an updated version with additional variables collected (e.g., accelerometric (ACC) data, oxygen saturation, blood pressure (systolic and diastolic), and glycemia), the 12 participants (4 recordings per participant) carried over from the original (v1.0.0) dataset are missing these values. To address this, recordings from v 1.0 were excluded from model training. The classifier dataset therefore contained 3,840 recordings from 38 participants: 3 045 bad-quality and 795 good-quality recordings.

3.2.2 Correlation Analysis

Pearson correlations were calculated among age, height, weight, measurement site (ear or finger), glycemia, SpO_2, signal quality, heart rate, systolic and diastolic blood pressure, and the binary acquisition-condition indicators for rest, higher pressure, finger or ear movement on the lens, walking, coughing, laughing, changing light, and talking. Figure 3 presents these correlations, which were examined to identify clinically relevant associations and potential modeling concerns, including shortcut learning, data leakage, feature redundancy, and confounding.

But more importantly, the correlation matrix is used to detect possible sources of bias, shortcut learning, data leakage, feature redundancy, and confounding effects that could affect model training and evaluation. Statistical significance was evaluated using a threshold of $p < 0.05$. Pearson correlation coefficients were reported, and statistically significant correlations were marked after False Discovery Rate correction to reduce false-positive findings caused by multiple pairwise comparisons.

The strongest correlations among the demographic, anthropometric, and physiological variables involved weight, height, and blood pressure. Weight was strongly correlated with systolic blood pressure ($r = 0.78$) and moderately correlated with diastolic blood pressure ($r = 0.45$). Height was also correlated with systolic blood pressure ($r = 0.57$) and diastolic blood pressure ($r = 0.37$). These relationships are medically plausible because body mass is associated with elevated blood pressure and an increased risk of hypertension (Landi et al., 2018).

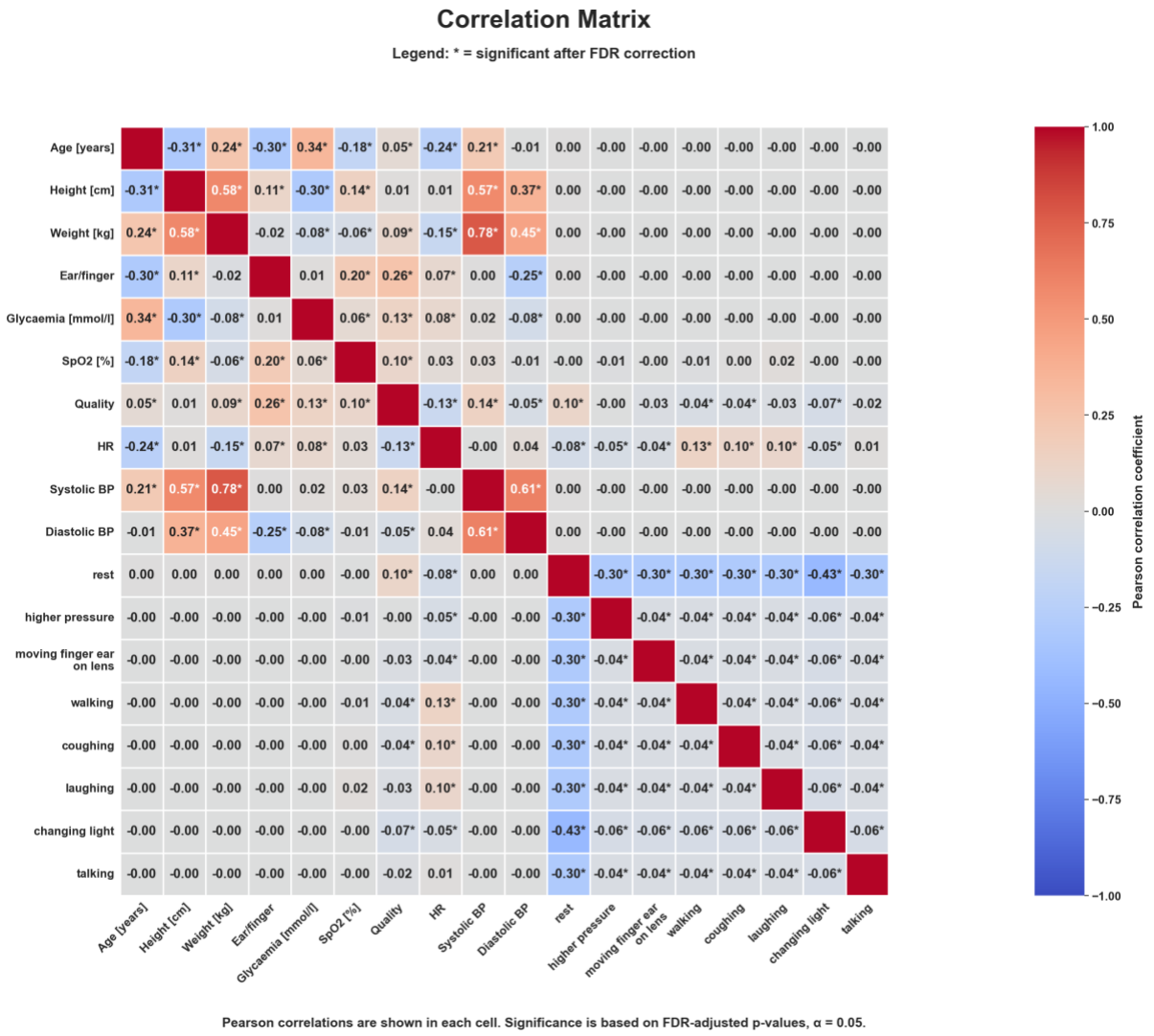


Figure 3. Correlation Matrix

However, from an AI pipeline perspective, these variables may also represent shortcut features. If metadata such as weight, height, blood pressure, or gender are included as model inputs, the model may learn demographic or anthropometric patterns that might not be easily changed.

Systolic and diastolic blood pressure were positively correlated (r = 0.61). This supports internal consistency within the physiological measurements, since both variables represent related components of arterial blood pressure (Gavish, 2010).

For the AI pipeline, this correlation indicates feature redundancy. If both systolic and diastolic blood pressure are included as inputs, they may provide overlapping information and should therefore be handled carefully during feature selection, regularization, or model interpretation.

Age was positively correlated with glycemia (r = 0.34) and systolic blood pressure (r = 0.21). Age was negatively correlated with heart rate (r =- 0.24), SpO2 (r =- 0.18), and height (r =- 0.31). These associations indicate that age is linked to several physiological variables in the dataset. For the AI pipeline, this means age may influence model behavior even if age is not explicitly used as an input, because age-related physiological patterns may still be present in the PPG signal.

The measurement spot was coded as 0 for the ear and 1 for the finger. Therefore, positive correlations with the ear/finger variable indicate higher values for finger recordings, while negative correlations indicate higher values for ear recordings. The ear/finger variable was positively correlated with PPG quality (r = 0.26), indicating that finger recordings had higher PPG quality than ear recordings in this dataset. It was also positively correlated with SpO_2 (r = 0.20), suggesting that finger recordings had higher oxygen saturation values than ear recordings. This is important for the AI pipeline because the model could learn sensor-location-specific signal characteristics rather than generalizable PPG quality features that can be improved through user action.

The ear/finger variable was negatively correlated with age (r =- 0.30) and diastolic blood pressure (r = - 0.25). Since the ear was coded as 0 and the finger as 1, these negative correlations indicate that older participants and participants with higher diastolic blood pressure were more associated with ear recordings. For the AI pipeline, this introduces a potential confounding pathway: the model may learn age- or blood-pressure-related differences from sensor-location artifacts.

PPG quality showed weak correlations with most physiological and task-related variables. Quality was correlated with measurement spot, r = 0.26; heart rate, r = -0.13; SpO_2, r = 0.10; rest, r = 0.10; and changing light, r = -0.07. These results suggest that signal quality is affected by acquisition conditions, but no single variable strongly explains PPG quality on its own.

Motion or acquisition-condition variables were also relevant to the AI pipeline. The condition variable was coded as 0 for rest and 1 for higher pressure, with the remaining task conditions represented separately in the correlation matrix.

The task variables included rest, higher pressure, moving the finger/ear on the lens, walking, coughing, laughing, changing the light, and talking. Several of these task variables were negatively correlated with one another, especially rest compared with active or altered-acquisition conditions. These negative correlations are likely due to the experimental design.

Walking was positively correlated with heart rate (r = 0.13), whereas rest was negatively correlated (r =- 0.08). This is physiologically plausible because physical activity generally increases heart rate compared with resting conditions.

Altogether, the correlation matrix indicates that the dataset contains several non-random structures that could affect AI model development. These include gender-related physiological differences, body size and blood pressure clustering, measurement-spot effects, age-related physiological patterns, and condition effects. These structures are important because an AI model may exploit them as shortcuts rather than learning robust features.

To address this risk, gender, age, measurement site, and task condition should be treated as potential confounders. Measurement sites should be monitored closely because finger recordings showed higher PPG quality than ear recordings, while ear recordings were more associated with older participants and higher diastolic blood pressure.

3.2.3 Statistical Analysis

Based on the correlation matrix, a Generalized Linear Mixed Model was conducted in JASP to evaluate whether selected tasks and acquisition conditions significantly affected PPG quality. This was the only inferential statistical analysis conducted. Because PPG quality was modeled as a binomial outcome, a binomial family with a logit link function was used. The number of trials was included as the binomial denominator, and participant ID was included as a random-effects grouping factor to account for repeated measurements from the same participant. The fixed effects included walking, coughing, talking, and changing the light. Model terms were tested using likelihood-ratio tests with Type III sums of squares.

The results showed that all four fixed effects were statistically significant predictors of PPG quality. Walking had a significant effect on PPG quality, χ^2 (1) = 17.22, $p < .001$. Coughing also had a significant effect, χ^2 (1) = 18.55, $p < .001$. Talking had a smaller but still statistically significant effect, χ^2 (1) = 5.249, $p = .022$. Changing light also significantly affected PPG quality, χ^2 (1) = 14.42, $p < .001$.

Despite their statistical significance, these factors were experimental variables in the dataset and one-hot encoded for this research project. In a different implementation, however, these factors would not necessarily be manually labeled. An AI model would need to detect or infer them from available sensor data, contextual signals, or learned patterns. But for the purposes of this study's AI pipeline, these results confirm that acquisition and task conditions are associated with PPG quality.

Walking, coughing, talking, and changes in environmental lighting should therefore not be treated as random noise. Instead, they represent structured acquisition conditions that can affect signal quality and model behavior. These conditions should be preserved across training, validation, and test sets, and model performance should be evaluated under each condition when possible.

3.3 Methods

Following the correlation and statistical analyses of the previous sections, three experiments were conducted; at no point were PPG, ECG, or ACC data processed for inclusion in any of them. In Experiment 1, models were trained using the full tabular feature set. The tabular data were preprocessed by merging subject-level information with the signal-quality and heart-rate annotations.

Blood pressure was split into systolic and diastolic components, gender was one-hot encoded, and motion labels were converted into binary one-hot indicators for rest, higher pressure, moving the finger or ear on the lens, walking, coughing, laughing, changing light, and talking. After preprocessing, the full tabular model used 18 numeric feature columns. It was trained with data from 38 participants.

In Experiment 2, the models' input was restricted to the motion-only features. The same prediction target was used, but the feature set was limited to the eight binary movement-condition variables. This experiment tested whether movement context alone was sufficient to identify bad-quality PPG. In Experiment 3, the same motion-only feature set was used, but the training data were balanced by undersampling the majority class within the training split.

The balanced training set contained 611 bad-quality observations and 611 good-quality observations. The test set was not balanced, so performance was still evaluated on the original held-out distribution.

The main goal was to evaluate whether machine-learning models could classify PPG signal quality and whether model explanations could support user-facing corrective feedback. In all experiments, the target variable was Quality, where 0 represented bad-quality PPG, and 1 represented good-quality PPG. The practical focus was therefore on identifying the bad-quality class, without mislabeling a significant number of good-quality measurements.

The same models were evaluated across the experiments: Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors, XGBoost, LightGBM, Multilayer Perceptron, and Multilayer Perceptron without early stopping.

Each model was evaluated with no dimensionality reduction, PCA, and LDA. These three feature representations were evaluated to determine whether performance on unseen participants was best supported by retaining all predictors, reducing correlated and redundant information while preserving 95% of the variance with PCA, or emphasizing the single class-discriminative dimension with LDA.

Hyperparameters were tuned using both GridSearchCV and RandomizedSearchCV, with the F1-Score for the bad-quality class as the primary tuning objective. Next, 20 cases with bad-quality outcomes, eight Walking cases, four Coughing cases, four Changing Light cases, and four Talking cases, were randomly selected to evaluate the explainable AI techniques (SHAP, LIME, and DiCE) across models.

Because of the class imbalance, accuracy and bad-quality F1-score were carefully analyzed. A model could obtain high accuracy and a high bad-quality F1-score simply by predicting most or all observations as bad quality. Therefore, balanced accuracy, AUROC, class-specific F1-scores, confusion matrix, and explainability outputs were also used to compare model quality.

Counterfactual analysis was used as an additional model-selection and interpretation step after the main predictive evaluation. The purpose was to determine not only which model performed best, but also which model could provide concise and meaningful explanations for how bad-quality PPG signals could be shifted toward good-quality classifications.

The counterfactual analysis was applied to the full tabular-feature experiment because this experiment produced the strongest overall predictive performance. Twenty true bad-quality test cases were selected from the Experiment 1 test set. For each trained model, a DiCE bad-to-good counterfactual search was performed. A valid counterfactual was defined as a modified case that satisfied two conditions. First, the model prediction had to change from bad quality to good quality. Second, the counterfactual was allowed to alter no more than two input features. This maximum-two-feature constraint was used to keep the explanations concise, interpretable, and relevant for identifying possible intervention points.

For each model, the number of valid counterfactuals was recorded across the 20 selected cases. In addition, the features changed in the valid counterfactuals were counted and ranked. This DiCE counterfactual feature rank was used to identify which variables most frequently acted as decision-boundary levers. Model selection was therefore based on a combined criterion: predictive performance, counterfactual coverage, and interpretability of the ranked counterfactual features.

The best model was selected using two main criteria: predictive performance and model counterfactuals. Predictive performance was assessed primarily using balanced accuracy and supported by PR-AUC, overall accuracy, bad-quality recall, and bad-quality precision. Model counterfactuals were evaluated by the number of selected bad-quality cases for which the model generated valid bad-to-good counterfactuals.

After the best model was identified, 10 of the 20 selected test cases were randomly retained for deeper case-level interpretation: four Walking cases, two Coughing cases, two Changing Light cases, and two Talking cases. SHAP, LIME, DiCE, and corrective-action outputs were compared for each case. The 20-case analysis was used for model comparison, whereas the 10-case analysis was used to evaluate case-level outputs and generate the prompts used in the subsequent interface study.

Llama 3.1:8B was then prompted through the Ollama generation API to generate a facilitator prompt based on the Fogg Behavior Model, using the AI model's output and explainability techniques, with a particular focus on counterfactuals that require the minimum single change in the motion feature to produce high-quality output. Lastly, GPT-5.5 was then prompted with Llama's prompt and the same data outputs from the 10 selected cases of the best-performing model across all 3 experiments.

The QUEST principles for human evaluation of LLMs in health care were used to compare the LLM outputs. The higher-ranking prompts from this comparison were then used to create smartwatch interfaces in Figma for an A/B study comparing them to the existing interface.

3.3.1 Machine Learning Pipeline

3.3.1.1 Preprocessing

Before model fitting, the dataset was restricted to numeric tabular predictors. The identifier column, participant/group identifier, and target label were excluded from the feature matrix. Missing numeric feature values were replaced using median imputation. For models that are sensitive to feature scale, including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Multi-Layer Perceptron, the predictors were standardized using z-score scaling. Tree-based models did not require scaling unless dimensionality reduction was applied. To avoid data leakage between participants, training and testing were performed using a group-aware split based on person_id. The same participant could not appear in both the training and test sets. A stratified group k-fold procedure used five folds, a random state of 42, and the first fold as the held-out test fold. Each model was evaluated under three feature-representation conditions. In the first condition, the predictors were used without dimensionality reduction.

In the second condition, Principal Component Analysis was applied to retain 95% of the variance. In the third condition, Linear Discriminant Analysis (LDA) was applied, retaining one discriminant component because the task was binary.

3.3.1.2 Logistic Regression

Logistic Regression classified signal quality by learning a weighted linear combination of the input features. For each sample, the model computes a linear score from the predictors and then transforms it into a probability using the logistic sigmoid function. If the estimated probability was above the decision threshold, the sample was assigned to the corresponding class. This model assumes that the log-odds of the outcome are linearly related to the predictors. In this analysis, Logistic Regression was configured with a maximum of 5,000 iterations, balanced class weights, the lbfgs solver, and random state 42.

Balanced class weights were used so that the minority class contributed more strongly to the optimization objective. The model minimized regularized binary cross-entropy loss, and optimization was performed using the limited-memory BFGS algorithm.

3.3.1.3 Support Vector Machine

The Support Vector Machine classified data by finding a separating boundary between bad-quality and good-quality observations. Because a radial basis function kernel was used, the classifier could learn nonlinear boundaries rather than only a straight linear separation. The model works by identifying support vectors, which are training samples near the class boundary that determine the final decision surface. For a new sample, the SVM computes its similarity to support vectors using the radial basis function kernel. The sample is then classified according to which side of the learned boundary it falls on. In this analysis, the SVM used an RBF kernel (C = 1.0), gamma = scale, enabled probability estimates, balanced class weights, and a random state of 42.

Probability estimates were enabled because AUROC, SHAP, LIME, and DiCE required continuous probability outputs. The model optimized a margin-based hinge-loss objective with regularization, where (C) controlled the trade-off between margin width and classification errors.

3.3.1.4 Random Forest

Random Forest classified signal quality using an ensemble of decision trees. Each tree learned a sequence of threshold-based rules, such as whether a feature value was above or below a selected cutoff. A single tree classified a sample by applying these rules until it reached a terminal node.

The Random Forest combines many such trees and makes the final classification by averaging class probabilities or by majority voting. This approach can capture nonlinear relationships and interactions between features without requiring feature scaling. In this analysis, the Random Forest used 500 trees, a minimum leaf size of 2, balanced subsampling for class weighting, parallel processing with n_jobs = -1, and random state 42. Each tree was trained using greedy recursive splitting, where the selected split was the one that most improved node purity. The optimization was therefore not gradient-based; it was a greedy search over candidate feature thresholds.

3.3.1.5 K-Nearest Neighbors

K-Nearest Neighbors classified each test sample by comparing it to the training samples and selecting the k most similar observations. Similarity was determined using a distance metric in feature space. The model then looked at the labels of the nearest training samples and predicted the class by voting.

In this analysis, KNN used 15 neighbors and distance-weighted voting. This means that closer neighbors had more influence on the final prediction than farther neighbors. Unlike Logistic Regression or neural networks, KNN does not learn model coefficients during training. Instead, training consisted of storing the labeled training data, and classification is performed at prediction time by searching for the nearest neighbors.

3.3.1.6 Multi-Layer Perceptron

The Multi-Layer Perceptron classified signal quality using a feed-forward artificial neural network. The input features were passed through hidden layers, where each layer computed weighted combinations of the previous layer followed by nonlinear activation functions. These hidden layers allowed the model to learn nonlinear relationships between predictors and signal quality.

The default neural network architecture contained two hidden layers with 100 and 50 neurons. The maximum number of training iterations was 500, the random state was 42, and the validation fraction for early stopping was 0.2. Two MLP variants were evaluated: one with early stopping enabled and one without. The MLP minimized binary cross-entropy loss with L2 regularization. Optimization was performed using Adam, a stochastic gradient-based optimizer, in the scikit-learn implementation.

3.3.1.7 XGBoost

XGBoost classified samples using gradient-boosted decision trees. Instead of fitting all trees independently, XGBoost builds trees sequentially. Each new tree was trained to address the mistakes of the previous ensemble. The final prediction was the sum of all tree outputs, transformed into a probability using the logistic function.

The default XGBoost model used 500 trees, a learning rate of 0.05, a maximum tree depth of 3, subsampling of 0.8, column subsampling of 0.8, the binary logistic objective, AUCPR as the evaluation metric, the histogram tree method, random state 42, and parallel processing with n_jobs = -1.

Class imbalance was handled through scale_pos_weight, calculated from the class distribution. XGBoost optimized a regularized logistic-loss objective using gradient boosting with first- and second-order derivative information.

3.3.1.8 LightGBM

LightGBM was also used as a gradient-boosted decision-tree classifier. Like XGBoost, it built an additive ensemble of trees, in which each tree corrected the residuals of the current model. The final raw model score was transformed into a probability for binary classification.

The default LightGBM model used 500 trees, a learning rate of 0.05, 31 leaves, subsampling of 0.8, column subsampling of 0.8, class imbalance weighting through scale_pos_weight, random state 42, parallel processing with n_jobs = -1, and suppressed verbosity. LightGBM optimized binary log loss using gradient boosting, with efficient histogram-based split finding and leaf-wise tree growth.

3.3.1.9 Hyperparameter tuning

Each classifier was evaluated under three hyperparameter conditions: no tuning, GridSearchCV, and RandomizedSearchCV. The no-tuning condition used the default parameters described in the subsections above. GridSearchCV exhaustively evaluated all combinations in a predefined parameter grid. RandomizedSearchCV sampled up to 20 parameter combinations from predefined candidate sets.

Both tuning methods used stratified group k-fold cross-validation with three folds. The tuning objective was the F1-score for the bad-quality class (Quality = 0). This ensured that model selection was aligned with the primary objective of detecting bad-quality signal segments. For PCA-based models, the retained variance was also tuned using candidate values of 0.90, 0.95, and 0.99.

3.3.2 Explainable Artificial Intelligence Methods

SHAP, LIME, and DiCE were used to interpret model classifications. SHAP estimated the contribution of each feature to the predicted probability of poor quality. Kernel SHAP was used as a model-agnostic method to generate explanations for all classifiers. The SHAP background sample size was 100, the explanation sample size was 20, the number of SHAP samples was 100, and the top 20 features were retained for global importance. LIME generated local explanations for individual classifications by perturbing the input features around a selected sample and fitting a simple interpretable model locally.

LIME explanations were generated for 20 samples, and the top 15 local feature conditions were reported. DiCE generated counterfactual explanations. DiCE tries to identify feature changes that would move a prediction from bad quality to good quality. One counterfactual was generated per explained instance. In experiment 1, DiCE was allowed to vary all model features, rather than being restricted to motion-related features (as in experiments 2 and 3).

3.3.3 Evaluation Metrics

3.3.3.1 Confusion Matrix

The confusion matrix was interpreted using the class order [0, 1], corresponding to bad-quality and good-quality PPG, respectively. Because class 0 was the class of interest, the terms true positive and false negative below refer to bad-quality detection.

$$CM = \begin{bmatrix} TP_0 & FN_0 \\ FP_0 & TN_0 \end{bmatrix}$$

$$TP_0 = actual\ bad\ quality, predicted\ bad\ quality$$

$$FN_0 = actual\ bad\ quality, predicted\ good\ quality$$

$$FP_0 = actual\ good\ quality, predicted\ bad\ quality$$

$$TN_0 = actual\ good\ quality, predicted\ good\ quality$$

3.3.3.2 Accuracy

Accuracy measures the proportion of all observations correctly classified.

$$Accuracy = \frac{TP_0 + TN_0}{TP_0 + TN_0 + FP_0 + FN_0}$$

3.3.3.3 Precision for Bad Quality

Precision for class 0 (bad quality) measures how many observations predicted as bad quality were actually bad quality.

$$Precision_0 = \frac{TP_0}{TP_0 + FP_0}$$

3.3.3.4 Recall for Bad Quality

Recall for class 0 (bad quality) measures how many actual bad-quality observations the model detected.

$$Recall_0 = \frac{TP_0}{TP_0 + FN_0}$$

3.3.3.5 F1-Score and F0.5-Score

The F1-score is the harmonic mean of precision and recall. It was used as the main optimization criterion for hyperparameter tuning.

$$F1_0 = \frac{2(Precision_0)(Recall_0)}{Precision_0 + Recall_0}$$

The F0.5-score gives more weight to precision than recall. It was computed as:

$$F_{0.5} = (1 + 0.5^2)\frac{(Precision_0)(Recall_0)}{0.5^2 Precision_0 + Recall_0}$$

3.3.3.6 Balanced Accuracy

Balanced accuracy averages recall across both classes and is useful when classes are imbalanced.

$$Balanced\ Accuracy = \frac{Recall_0 + Recall_1}{2}$$

3.3.3.7 AUROC

Area Under the Receiver Operating Characteristic Curve (AUROC) assesses the model based on predicted scores rather than just binary class labels. It determines whether the model assigns higher scores to genuinely bad-quality samples compared to good-quality ones. In this project, the AUROC score is derived from the predicted probability or decision score for class 0, which indicates bad quality.

$$FPR_0 = \left(\frac{FP_0}{FP_0 + TN_0}\right) \qquad TPR_0 = \left(\frac{TP_0}{TP_0 + FN_0}\right)$$

$$AUROC = \int_0^1 TPR_0\ (FPR_0)\ dFPR_0$$

3.3.4 Balanced Motion-Only Experiment

The balanced motion-only experiment (Experiment 3) was designed to evaluate whether class imbalance affected model learning when using only movement-condition features. The training data were balanced by undersampling the majority class. If N_0 is the number of bad-quality training samples and N_1 is the number of good-quality training samples, the balanced training size per class was:

$$n = min(N_0, N_1)$$

$$N_{balanced} = 2n$$

This balancing was applied only to the training data. The held-out test set remained unchanged, so final performance reflected the original evaluation distribution.

3.4 Results

The dataset was highly imbalanced. In total, it had 3,045 bad-quality records and 795 good-quality ones. Bad-quality PPG accounted for 79.3% of the full dataset and 78.4% of the test set. The test set contained 852 recordings: 668 bad-quality PPG-labeled signals (78.4%) and 184 good-quality ones (21.6%). As a result, some models predicted all samples as bad quality, achieving an apparently strong accuracy and a strong bad-quality F1-score without distinguishing between good-quality and bad-quality PPG.

Three experiments were evaluated to determine which modeling strategy provided the most reliable classification of PPG signal quality. The classification task was binary, with class 0 representing bad-quality signals and class 1 representing good-quality signals. For the purposes of this thesis, balanced accuracy, bad-class precision, bad-class recall, and the pattern of false positives and false negatives were used to interpret the results.

3.4.1 Experiment 1: Models Trained with Full Tabular Features

Experiment 1 used the full tabular feature set and produced the strongest overall results. The best practical model was the MLP model optimized with RandomizedSearchCV. This model achieved an accuracy of 0.781, a balanced accuracy of 0.657, a bad-class precision of 0.850, a bad-class recall of 0.874, and a bad-class F1-Score of 0.862. It correctly detected 584 out of 668 bad-quality cases and missed 84 bad-quality cases. It also correctly identified 81 of 184 good-quality cases but incorrectly classified 103 as bad. This result shows that the full tabular feature set provided useful information for distinguishing between bad- and good-quality PPG signals. Although the MLP with PCA achieved a slightly higher bad-class F1-Score of 0.876, it had a balanced accuracy of only 0.497 and failed to correctly identify any good-quality samples. Therefore, the MLP without PCA was the stronger and more reliable model for practical use.

3.4.2 Experiment 2: Motion-Only Features

Even though the MLP without PCA was the best-performing model, the explainability analysis showed that it was learning from non-changeable features such as age, gender, and blood pressure. Experiment 2 evaluated models trained only on motion and condition features, including rest, higher pressure, moving the finger or ear on the lens, walking, coughing, laughing, changing light, and talking, in the hope that a model that recommends changes in motion activity to the user could be used. Several models in this experiment reached a bad-class F1-Score of approximately 0.879.

However, this result was misleading because these models often predicted all or nearly all test samples as bad quality. For example, the SVM model optimized with RandomizedSearchCV achieved a bad-class recall of 1.000 and a bad-class F1-Score of 0.879, but its balanced accuracy was only 0.500 because it classified all 184 good-quality test samples as bad.

This means the model did not truly distinguish between the two classes. The models with better balanced accuracy reached only around 0.608, but they detected far fewer bad-quality samples. For instance, Logistic Regression optimized with RandomizedSearchCV achieved a balanced accuracy of 0.608 and a high bad-class precision of 0.906, but its bad-class recall was only 0.347, meaning it missed 436 out of 668 bad-quality cases.

These results indicate that motion-only features were insufficient for robust signal-quality classification, but they did generate counterfactuals suggesting modifications to the prediction that the user could make.

3.4.3 Experiment 3: Balanced Motion-Only Models

Therefore, Experiment 3 repeated the motion-only approach but used a balanced training set with 611 bad-quality and 611 good-quality samples. This was intended to reduce the effect of class imbalance during training. However, balancing the training data did not substantially improve test performance. Some models again achieved a bad-class F1-Score of approximately 0.879 by predicting all samples as bad quality. For example, the MLP with LDA optimized using RandomizedSearchCV achieved a bad-class recall of 1.000 and a bad-class F1-Score of 0.879, but its balanced accuracy remained 0.500 because it failed to correctly identify any good-quality samples. The best balanced-accuracy results in this experiment were approximately 0.609, which was almost identical to the best balanced performance in Experiment 2. Logistic Regression optimized with GridSearchCV achieved a balanced accuracy of 0.609, a bad-class precision of 0.924, a bad-class recall of 0.310, and a bad-class F1-Score of 0.464. This model correctly identified most of the good-quality cases but missed 461 out of 668 bad-quality cases. Therefore, balancing the training data reduced some bias but did not solve the core limitation of using motion-only features.

3.5 Best Model Selection

As shown in Table 4, MLP [RandomizedSearchCV] was selected as the best model because it combined the strongest predictive performance with high counterfactual coverage. It achieved the highest balanced accuracy (0.657), with PR-AUC of 0.882, overall accuracy of 0.781, bad-quality recall of 0.874, and bad-quality precision of 0.850. It also generated valid bad-to-good counterfactuals for 19 of the 20 selected bad-quality cases. Compared with other competitive models, this provided the best balance between classification performance and counterfactual usefulness; some models produced more counterfactuals, but their balanced accuracy was close to 0.50, while others had reasonable performance but lower counterfactual coverage.

Table 4. Experiment 1. Predictive performance of the top full-feature models

Model	Balanced Accuracy	PR-AUC	Accuracy	Bad-Quality Recall
MLP [RandomizedSearchCV]	0.657	0.882	0.781	0.874
MLP + LDA [GridSearchCV]	0.613	0.869	0.708	0.780
LightGBM + LDA [RandomizedSearchCV]	0.608	0.861	0.623	0.635
MLP no early stop + LDA [GridSearchCV]	0.607	0.869	0.714	0.795
KNN + LDA [GridSearchCV]	0.605	0.860	0.704	0.780
KNN + LDA [RandomizedSearchCV]	0.605	0.860	0.704	0.780
MLP no early stop + PCA [RandomizedSearchCV]	0.605	0.859	0.754	0.867
SVM + LDA [RandomizedSearchCV]	0.596	0.869	0.481	0.394
Logistic Regression [GridSearchCV]	0.595	0.887	0.448	0.337
Logistic Regression [RandomizedSearchCV]	0.595	0.887	0.448	0.337

The DiCE feature ranking for MLP [RandomizedSearchCV], shown in Figure 4, indicated that Ear/finger was the dominant changed feature, appearing in 14 valid counterfactuals. This was followed by HR in 4, Gender in 2, and SpO2 and Height in 1 each. These results suggest that the model's bad-to-good decision boundary was mainly influenced by sensor-site and physiological features. Importantly, Ear/finger is the most actionable of these features because it can be related to sensor placement, measurement site, or contact quality during PPG acquisition.

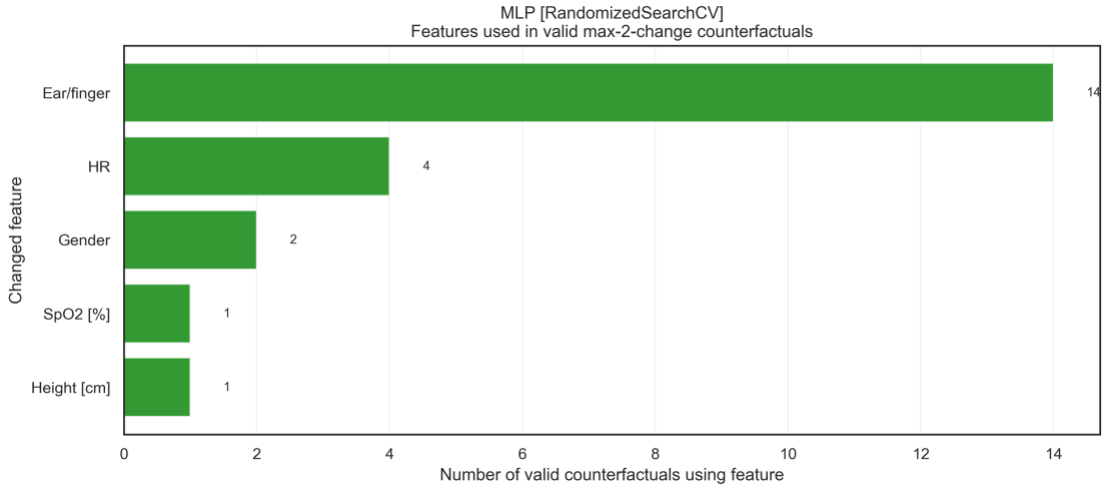


Figure 4. DiCE counterfactual feature ranking for MLP [RandomizedSearchCV]

The 10-case interpretation run further showed that the selected model could generate case-level feedback for most individual bad-quality recordings. Out of the 10 cases, 9 produced valid counterfactuals, while only 1 case did not generate a valid bad-to-good counterfactual. The generated outputs identified actionable features associated with the model's classification, such as higher pressure, coughing, finger or ear movement on the lens, walking, changing light, and talking. These were translated into practical corrective prompts such as applying lighter and steadier contact, waiting until coughing settles, keeping the measurement site still, stopping walking, improving lighting conditions, and staying quiet during measurement. Therefore, the counterfactual analysis supported the selection of MLP [RandomizedSearchCV] and showed that its explanations could be translated into human-centered corrective feedback, while still requiring caution for non-actionable features such as gender, height, and other physiological or demographic variables.

3.6 Using the Fogg Behavior Model for LLM Prompt Design

The Fogg Behavior Model was used as a behavior framework for prompt construction. According to this model, behavior occurs when motivation, ability, and a prompt converge simultaneously. For this project, the target behavior was defined as the user successfully retrying a smartwatch-based physiological measurement under conditions that allow the system to obtain sufficient signal quality.

This approach assumes that users who have already attempted a measurement are motivated to complete the action. Therefore, the primary design challenge is not to increase motivation, but to increase ability. The prompts were consequently designed as facilitator prompts. A facilitator prompt is appropriate when the user is motivated, but the behavior is difficult to perform. In this context, difficulty may arise from physical movement, speaking, coughing, running, walking, poor lighting, or other factors that reduce signal quality.

The LLM was instructed to use the Fogg Behavior Model as a reasoning framework. Specifically, it was prompted to identify the ability barrier preventing a successful or interpretable measurement, translate that barrier into a small corrective action based on the counterfactuals, and generate a concise prompt that supports immediate retry or appropriate help-seeking. The prompt was not intended to persuade the user or assign blame, but to enable corrective action that would allow the measurement to succeed.

3.6.1.1 Inputs

The prompt-generation pipeline receives several types of evidence from the underlying machine-learning system. These include the model output, signal-quality indicators, SHAP explanations, LIME explanations, and DiCE counterfactuals. The output indicates whether the system produced a valid recording, an unsuccessful measurement, an inconclusive result, or no reliable decision. Signal-quality metrics provide information about measurement reliability and may include indicators such as motion variables, measurement spot, gender, age, height, weight, glycemia, SpO2, Systolic_BP, Diastolic_BP, and the pressure applied to the sensor.

SHAP and LIME are used to identify features that contributed to the model's bad-quality classification. DiCE counterfactuals are used to identify feature changes that would likely move the model toward a good-quality prediction, such as a successful measurement or sufficient signal quality.

Together, these methods provide complementary evidence: SHAP and LIME indicate which features influenced the current result, while DiCE suggests which feature changes may lead to a different result.

3.6.1.2 Feature-Actionability Layer

A static, RAG-inspired feature-actionability policy was embedded directly in the prompt to provide the LLM with a feature-actionability policy. This layer serves as a domain-specific knowledge source that classifies features as changeable, partially changeable, non-changeable, sensitive, or physiological, though no runtime retrieval was conducted. The purpose of this layer is to prevent the LLM from generating prompts based on unfair, unsafe, non-actionable, or misleading features. Changeable features included user behavior, device positioning, and environmental conditions that the user could reasonably adjust. Partially changeable features included states such as elevated breathing after activity, heart rate after exertion, or skin moisture after sweating, where the user cannot directly change the physiological state but can rest, wait, dry the wrist, or try again later. Non-changeable or sensitive features included age, gender, and blood pressure. These features were excluded from user-facing corrective prompts.

3.6.1.3 Prompting and Inference Strategy

The model was instructed through the prompt to apply a triage process before producing a facilitator prompt. First, it determined whether a valid decision had been made. Second, it assessed whether the model-associated factor of the unsuccessful or inconclusive measurement was identifiable. Third, it compared SHAP, LIME, and DiCE to determine whether the explanation methods converged on the same or similar features. Fourth, it checked whether those features were actionable, partially actionable, non-actionable, sensitive, or physiological. Finally, it generated a facilitator prompt. Facilitator prompts were generated when the evidence converged on clear, actionable features.

3.6.1.4 Runtime Prompting

At runtime, the instruction-guided LLM receives each structured, case-specific prompt for each of the ten cases identified in the previous sections. The runtime prompt includes the decision state, signal-quality metrics, SHAP explanations, LIME explanations, DiCE counterfactuals, the prompt-embedded feature-actionability policy, interface constraints, and domain context. The runtime prompt instructs the model to use the Fogg Behavior Model, assuming the user is motivated if they have already attempted the measurement.

The LLM is then directed to focus on increasing the user's ability by identifying actionable barriers and generating the safest, most appropriate prompt. The prompt explicitly instructs the model not to invent causes, not to use protected or sensitive attributes in the user-facing text, not to imply guaranteed success, and not to allow a lack of result to be mistaken for a valid negative result. The model output is structured as a standardized numbered list. Each prompt contains a short title, an explanatory sentence, a corrective-action sentence, and a call-to-action label, following the structure presented in Table 5.

Table 5. Structure of the Facilitator Prompt

Structure of the Facilitator Prompt	
Short title	Maximum 3 words and 2 connectors.
Body	1-3 short sentences. (1) The smallest actionable change that may help change the output; no sensitive attributes; (2) What to do to support the change; (3) Call to Action.
Additional constraints	Do not assume blame

3.7 Prompt Generation and Evaluation

The prompt shown in Table 6 was used to generate counterfactuals with Llama 3.1:8B and GPT-5.5.

Table 6. LLM Prompt

You are an LLM interpreter for smartwatch physiological measurement quality outputs.
Your task is to generate facilitator prompts for exactly the 10 cases provided below.

Use the Fogg Behavior Model as the reasoning framework. The target behavior is that the user successfully retries a smartwatch-based physiological measurement under conditions that allow the system to obtain sufficient signal quality.

Assume the user has already attempted the measurement and is motivated. Therefore, do not increase motivation, persuade, pressure, or emotionally influence the user. Focus only on increasing ability by identifying the most likely practical barrier and translating it into the smallest safe corrective action.

Each generated facilitator prompt must have four user-facing parts:

1. Short Title

- Maximum 3 words.
- Calm, neutral, and supportive.
- Not demanding, bossy, urgent, alarming, or emotionally manipulative.

2. Condition-linked explanation

- One short sentence explaining what likely affected the recording.
- Based only on the provided evidence.
- Describe the measurement condition, not the user.
- Do not mention SHAP, LIME, DiCE, probabilities, model internals, folds, feature names, or technical explanations.
- Do not mention sensitive, protected, or medical attributes.

3. Corrective Action

- One short sentence describing the smallest practical change that may improve signal quality.
- Actionable, transparent, and non-manipulative.
- Focus on increasing ability, not motivation.
- Do not imply guaranteed success.

4. Watch-Button Call to Action

- Separate button label.
- 1 to 3 words.
- Suitable for a smartwatch or measurement-interface button.
- Calm, neutral, optional, and non-coercive.
- Do not imply retrying is mandatory or guaranteed to work.

Apply this feature-actionability policy:

Changeable features may be used for user-facing corrective prompts:

- movement
- running
- walking
- talking
- coughing
- poor lighting
- device position
- wrist position
- measurement spot
- sensor pressure
- environmental conditions
- user behavior during measurement

Partially changeable features may be translated into safe retry guidance:

- elevated breathing after activity
- heart rate after exertion
- sweating
- skin moisture
- recent physical activity

Non-changeable, sensitive, or medical features must not appear in user-facing text:

- age
- gender
- height
- weight
- glycemia
- SpO2
- systolic blood pressure
- diastolic blood pressure
- protected attributes
- medical attributes that could imply diagnosis, blame, risk, or health status

Anti-manipulation rules:

- Do not use subliminal, hidden, manipulative, or deceptive techniques.

- Do not materially distort the user's behavior.
- Do not impair the user's ability to make an informed decision.
- Do not use urgency, fear, guilt, shame, social comparison, overconfidence, or emotional pressure.
- Do not exploit vulnerability, health concern, uncertainty, or lack of result.
- Do not frame the prompt as a medical warning.
- Do not suggest that the user has a health condition.
- Do not suggest that retrying will guarantee a valid result.
- Do not hide the reason for the recommended action.
- The prompt must be transparent: the user should understand why the suggested action may help the measurement.
- The prompt must preserve user autonomy: retrying should be presented as optional and supportive.

For each case, perform this triage internally:

1. Determine whether a valid decision was made.
2. Determine whether the cause of the unsuccessful, inconclusive, or low-quality result is identifiable.
3. Compare available evidence to determine whether it converges on the same or similar features.
4. Determine whether the convergent features are actionable, partially actionable, non-actionable, sensitive, physiological, or unknown.
5. Generate a facilitator prompt only when the evidence supports a safe actionable or partially actionable barrier.
6. If the evidence is weak, unavailable, conflicting, sensitive, or non-actionable, generate a cautious neutral retry prompt that does not invent a cause.

Rules:

- Do not invent causes.
- Do not blame the user.
- Do not mention SHAP, LIME, DiCE, probabilities, model internals, feature names, folds, or technical explanations in the user-facing prompt.
- Do not use sensitive, protected, non-actionable, or medical attributes in the user-facing prompt.
- Do not imply guaranteed success.
- Do not treat an inconclusive, failed, missing, or unreliable result as a valid negative result.

- Keep the user-facing text short, supportive, and action-oriented.
- The watch-button call to action must be a separate button label.
- The watch-button call to action must be suitable for a smartwatch or measurement-interface button.
- Do not include internal reasoning in the output.
- Prompts for cases with the same selected condition, candidate barrier, actionability, evidence convergence, and action hint must use consistent wording.
- Cases 3 and 4 must refer to running, not walking.

Output format:

Return only the facilitator prompts as a normal readable numbered list.

Do not return JSON.

Do not include metadata, evidence labels, safety notes, model names, experiment names, actionability classes, feature names, or technical explanations.

Return exactly 10 numbered prompts, in the same order as the provided cases.

Each prompt must use this exact format:

Short Title

Model-informed counterfactual explanation.

Corrective action sentence.

CTA: Button Label

Now process only the following 10 cases:

[Insert the cases here]

3.8 Prompts Generated by the LLMs

Table 7 presents the facilitator prompts generated by the two LLMs per condition.

Table 7. Facilitator Prompts Generated by the Large Language Models

Condition	Llama 3.1:8B	GPT-5.5
Coughing	Wait It Out Coughing during rest likely affected the recording. Wait until coughing settles before retrying. CTA: Retry Measurement	Rest Still Coughing can affect the recording. Stay still in a resting position. CTA: Try Again
Talking	Stay Quiet Talking during the measurement likely caused issues with the signal quality. Stay quiet during the measurement. CTA: Try Again	Stay Quiet Talking can affect the recording. Stay quiet during the measurement. CTA: New Reading
Running	Stop and Hold Running may have caused movement that affected the recording. Stop running and keep the wrist still. CTA: Retry Measurement	Hold Still Running can make the signal harder to read. Stop running and keep your wrist still. CTA: Try Again
Walking	Stop and Hold Walking may have caused movement that affected the recording. Stop walking and keep the wrist still. CTA: Retry Measurement	Hold Still Walking can make the signal harder to read. Stop walking and keep your wrist still. CTA: Try Again
Changing-light Environment	Steady Lighting Changing lighting conditions likely affected the recording. Move to steadier lighting and keep the sensor clean and dry. CTA: Try Again	Rest Still Changing light can affect the recording. Stay still in a resting position. CTA: Try Again

3.8.1 Prompt Evaluation

Both sets of prompts were evaluated by a single reviewer following a variation of the QUEST principles proposed by Tam et al. (2024) for human evaluation of large language models in health care. The complete evaluation is presented in Appendix A. The QUEST-based evaluation showed that GPT-5.5 outperformed Llama 3.1:8B.

GPT-5.5 passed all mandatory gate criteria, including the required four-part structure, title and button length constraints, absence of technical leakage, absence of sensitive or medical content, and avoidance of manipulative language. Its prompts were generally concise, neutral, and well aligned with the evidence provided for each case. GPT-5.5 also maintained consistent wording across repeated cases with the same selected condition, candidate barrier, evidence convergence, and action hint, which improved reliability and reduced unnecessary variation across the prompt set.

Llama 3.1:8B performed substantially worse, primarily because it introduced several unsupported causes and corrective actions that were not present in the case evidence, including laughing, finger movement on an ear lens, sensor pressure, sensor cleanliness, and skin dryness.

It also failed a mandatory formatting requirement by treating one case as not requiring a prompt, despite the instruction to produce exactly 10 facilitator prompts. Although Llama 3.1:8B generally avoided protected attributes, medical claims, and overtly manipulative language, its alignment with unsupported model counterfactuals reduced accuracy, relevance, evidence handling, transparency, and reliability. Overall, GPT-5.5 was judged to be the stronger prompt set because it better preserved evidence alignment, user autonomy, neutral tone, and smartwatch-appropriate ability-focused guidance.

Nonetheless, both model-generated prompt sets required human revision before inclusion in the smartwatch interfaces. The human revisions corrected mismatches between the identified barrier and the suggested action, removed unsupported or fabricated causes, standardized wording across repeated cases, and ensured that each final prompt followed the required structure.

The final human-edited prompts presented in the next subsection, therefore, retained the useful ability-focused elements of the model outputs while improving evidence alignment, interface consistency, and suitability for deployment.

3.8.2 Revised Version of the Prompts

After comparing the prompts generated by GPT-5.5 and Llama 3.1:8B, as shown in Table 8, the prompts were modified to address limitations in the LLM outputs before they were evaluated in the human validation study.

Table 8. Final Version of the Facilitator Prompts

Condition	Prompt
Coughing	Wait until it settles! Coughing can affect the recording. Wait for it to settle and your breathing feel steady, then try again. CTA: Try Again
Talking	Settle in for the ECG Speaking can affect the ECG signal. Start again and avoid speaking until the recording is complete. CTA: Try Again
Running	Rest before the ECG Running can affect the ECG signal. Stop safely and wait until your breathing slows before trying again. CTA: Try Again
Walking	Pause safely Walking makes it harder to get a clear reading. Stop in a safe place, keep still, then try again. CTA: Try Again
Changing-light Environment	Improve your ECG The recording was not clear. Sit still in a well-lit place, rest your arms, and relax. Avoid speaking or moving and wait until coughing or heavy breathing settles. Keep your watch snug on clean, dry skin and move away from plugged-in electronics. CTA: Try Again

Note. The Changing Light Environment prompt intentionally combines guidance for multiple possible recording conditions and was treated as an exploratory multi-factor prompt; the other prompts target one primary condition

3.9 Code Availability

The code for the pipeline presented in this chapter is available on this GitHub repository: https://github.com/raulramirez17/MAScSYDEThesis

Chapter 4
A User-Centered Approach to Evaluating Counterfactual Explanations Aimed at Prompting User Behavior to Perform Corrective Actions

This chapter presents the results of a human validation study evaluating the effect of the counterfactual-based facilitator prompts generated in the previous chapter. The ECG interface from the Apple Watch was used as a proxy for the evaluation because, like the PPG-based hypertension feature, it depends on adequate signal quality and fails under unsuitable recording conditions. This chapter has three objectives: (1) to measure differences in users' reported intention to record an ECG after viewing the existing and proposed unsuccessful-measurement notifications; (2) to assess the understandability and perceived appropriateness of the facilitator prompts; and (3) to identify design considerations for Level 9 automated systems under the classification proposed by Parasuraman et al. (2000).

The aim is to guide the design of AI systems that decide when to involve humans in decision-making and action selection. This is achieved by using model counterfactual explanations to overcome data limitations that impact the effectiveness of these systems' underlying algorithms, thereby prompting corrective actions. An A/B study was conducted to evaluate whether presenting the prompts was associated with greater reported intention to retry an ECG after an initial failed measurement. The study further explores how much information users need before they take action. Additionally, this study incorporates AI-generated videos to present the case scenarios to participants.

4.1 Research Question and Hypothesis

This chapter addresses RQ3 through the following questions and hypotheses:

- **RQ3.1:** Do the proposed model-informed counterfactual facilitator prompt shown in the proposed interface increase retry intention after an unsuccessful ECG measurement compared with the existing interface?
    - H1: Across the ten scenarios, the proposed interface will produce higher odds and a higher adjusted probability of retry intention than the existing interface.

- **RQ3.2:** Does the proposed interface improve attention, clarity, motivation-to-retry, post-block motivation and confidence, and reduce the need for additional information?
    - H2: The proposed interface will yield more favorable scores on these secondary outcomes than the existing interface.
- **RQ3.3:** Are pre-to-post changes in motivation and confidence greater after the proposed interface than after the existing interface?
    - H3: The proposed interface will produce more positive pre-to-post changes in motivation and confidence.
- **RQ3.4** (Exploratory): Do interface effects vary by scenario, condition, or interface sequence?
- **RQ3.5** (Descriptive): How clear, helpful, relatable, and realistic are the AI-generated scenario videos?

4.2 Methods

This section presents the methods used in this human validation study. It describes the recruitment strategy, study design, prototype development, participant and researcher procedures, data-analysis plan, and exclusion process.

4.2.1 Recruitment

Participants were recruited from May 1st to May 29th, 2026, until the computed a priori sample size was reached. The recruitment strategy included email campaigns (Appendix B.1.1) to departments in the faculties of health, science, and engineering, wall posters at the University of Waterloo main campus (Appendix B.1.2), LinkedIn posts (Appendix B.1.3), and snowball sampling, where participants were asked to share the study recruitment poster with people they knew.

Eligible participants were adults aged 18 to 65 years living in Canada, the United States, Ireland, the European Union, or Panama who had a basic level of eHealth literacy, as measured by the eHealth Literacy Questionnaire (eHEALS). This questionnaire measures comfort and confidence in finding and using health information from an online source.

4.2.2 Study Design

The study used a two-period, two-sequence, counterbalanced crossover mixed-methods design. Every participant completed the ten scenarios in the same fixed numerical order. The interface sequence, however, was counterbalanced: participants assigned to sequence AB used the proposed interface for scenarios 1-5 and the existing interface for scenarios 6-10, whereas participants assigned to sequence BA received the reverse interface order. Interface was therefore a within-participant factor, while sequence was a between-participant factor.

After recruitment, participants read the information consent letter (Appendix B.2.1) and signed the consent form (Appendix B.2.2). After consent was obtained, participants completed the demographic and screening questionnaire (Appendix B.2.3). If the participants were eligible, they used a Microsoft Booking Site to book the study session at their convenience (Appendix B.2.4).

Study sessions consisted of 45- to 75-minute semi-structured interviews. During the study session, the researcher followed the study guide presented in Appendix B.3.1 . The participants were asked to share their screens, and the session was recorded, including audio, video, and screen. Participants interacted with a study platform (Appendix B.3.2) in Figma and a data-collection questionnaire (Appendix B.3.3) in Qualtrics. The development of the interfaces and the general study protocol will be discussed in the next subsection. After the study session, participants received an email containing the feedback and appreciation letter (Appendix B.4.1) and were invited to complete a contact form for remuneration (Appendix B.4.2).

4.2.3 Prototype Development

The studied hypertension notification feature uses PPG signals recorded in the background, without human intervention, to power its automated, AI-enabled system. Although users do not directly control PPG acquisition, their behavior over time can influence the quality of the recorded PPG, thereby affecting the availability, reliability, and validity of the hypertension feature. To address this limitation, a user-facing interface was developed to provide clear error messages and facilitator prompts when recording conditions prevent the system from collecting PPG data of sufficient quality.

The main purpose of this interface is to promote human corrective behavior. In other words, when the system identifies that user behavior is associated with its poor signal quality classification, the interface should help users understand what went wrong and how they can improve future recordings. Because the feature is AI-enabled, it is also important to evaluate how users respond when the system provides an incorrect or misleading recommendation. For example, the algorithm may mistakenly infer that a user's behavior is preventing adequate signal collection and present a counterfactual prompt that does not accurately reflect the true cause of the problem, when in fact the behavior has a limited effect on how the measurement is recorded. Evaluating this possibility is also essential because users may rely on these prompts to decide how to adjust their behavior or whether to seek further medical advice.

To test the effect of counterfactual prompts in a realistic smartwatch environment, we selected an existing heart-related feature with similar usability and communication challenges: the Apple Watch single-lead ECG feature. As with the proposed PPG-based hypertension feature, the ECG feature depends on adequate signal quality and may fail under unsuitable recording conditions. It also requires users to make corrective changes, such as adjusting how they position or interact with the watch, in order to obtain a valid recording. These characteristics made it an appropriate testing scenario for evaluating whether counterfactual-based causal prompts can support user understanding and encourage corrective behavior in a watch-based digital health interface.

The Apple Watch ECG feature is used by a population that overlaps with the intended users of a hypertension notification feature. Users may attempt to record an ECG when experiencing symptoms such as chest pain, which may raise broader concerns about cardiovascular health and prompt users to consider whether they should consult a health care professional. Therefore, testing the prompts in the context of the ECG feature provides a practical way to evaluate how users interpret system limitations, respond to signal-quality problems, and engage in corrective behavior within a controlled smartwatch environment where users have high motivation but limited ability to perform the behavior.

Chapter 3 identified four PPG acquisition conditions associated with classifier output: talking, walking, coughing, and changing light. Running was added as a higher-motion variation of walking, producing five conditions in total. Ten facilitator prompts were generated from these conditions, as described in Section 3.8.2, and displayed in Apple Watch-style Figma interfaces representing unsuccessful measurements.

As this was a within-subjects design study, half of the scenarios presented the proposed interface, and the other half presented the existing interface, which lacks the counterfactual prompt that we are validating.

Figure 5 shows the interfaces shown in the intervention scenarios. Interface 1 is intended to be the main screen of the ECG feature. To begin the recording, users clicked the digital crown on the right side of the watch. An information icon provided guidance on recording an ECG correctly. Selecting the icon opened Interface 4, which was also accessible through the Help call-to-action on Interface 3, which has a text that reads:

> Sit still in a well-lit place. Rest your arms, relax, and avoid speaking or moving. Wait until coughing or heavy breathing settles. Keep your watch snug on clean, dry skin and move away from plugged-in electronics.

After clicking the crown, interface 2 is shown to users to simulate the recording. After 1500ms, the recording fails, showing interface 3. Interface 3 displays a one-line banner indicating that the recording was unsuccessful. It then presents a suggested action, followed by up to three short sentences containing an explanation, a corrective action, and a call-to-action button leading to the next step.



Figure 5. Interfaces in the Intervention-Group Scenarios

The existing interface, showing a poor ECG recording, is shown in Figure 6.



Figure 6. Existing interface

4.2.4 Study Protocol

Participants were assigned to one of two interface sequences. Participants assigned to Sequence AB interacted with the proposed interface in Scenarios 1–5 and the existing interface in Scenarios 6–10. Participants assigned to Sequence BA received the interfaces in the reverse order. Participants were asked to answer the questions and interact with the prototype as if they were in the situations described in each scenario. Before beginning with the scenarios, the researcher asked participants (1) the model of the smartwatch that they currently use (if any), (2) how long they have been using the watch, (3) how many times they have recorded an ECG with the smartwatch (if ever), and (4) how confident they were in recording an ECG with a smartwatch.

For each scenario, participants watched an AI-generated video depicting an unsuccessful ECG measurement and read the corresponding scenario text. Before receiving device guidance, they were asked: (1) what they believed caused the unsuccessful reading; (2) what they would do immediately after the unsuccessful measurement; (3) how motivated they were to record an ECG in the scenario; (4) how important the situation would be if it happened to them; (5) how easy or difficult it would be to record an ECG under the depicted conditions; and (6) how confident they would be in recording the ECG correctly after the initial unsuccessful measurement.

After answering these questions, the participants were shown a second version of the scenario in which the cause of the unsuccessful measurement was disclosed, and then they were guided to the prototype representing the person's watch. They were asked to think aloud as they interacted with it. The navigation path was annotated by the researcher, and follow-up questions were asked to explore themes that might come up from the think-aloud comments. After the participants interacted with the prototype, they indicated whether they intended to retry the measurement and whether they needed additional information to make the decision. When the reported action was inconsistent with the action performed on the prototype, the researcher asked a follow-up question about the reason for the discrepancy.

This process was repeated across all ten scenarios. After each interface block, participants rated: (1) how strongly the unsuccessful-measurement notification captured their attention; (2) how clearly the corrective action was presented; (3) how much the interface content made them want to retry the ECG measurement; (4) what information or design elements were missing; (5) how motivated they felt to record an ECG in a similar situation; and (6) how confident they felt about recording an ECG correctly.

At the end of the study, participants were asked: (1) what influenced their decision to retry an ECG after an unsuccessful measurement; (2) what influenced their decision not to retry; (3) how confident they felt about recording an ECG after interacting with the interfaces; (4) in which scenarios the corrective action was most difficult to follow; (5) whether the existing or proposed interface was more motivating or helpful; (6) which features of the proposed interface helped them decide what to do more quickly or clearly; (7) what their strongest reason was for retrying after an unsuccessful measurement; (8) what their main barrier was when they chose not to retry after an unsuccessful measurement; and (9) what changes would make smartwatch health notifications clearer or more useful.

Finally, participants were asked about the AI-generated videos that were used in the study protocol: (1) how realistic they felt, (2) how helpful they were to understand the scenario, (3) how easy it was to place themselves in the conditions presented in the scenarios, and (4) how clear they were in conveying the scenario.

4.3 Scenario Design

The following scenarios were developed for this human validation study. Each scenario presented a variation to assess whether participants understood the cause of the unsuccessful measurement as presented in the case. After measuring this, a second version of the scenarios was presented to users to explain the cause of the unsuccessful measurement before interacting with the smartwatch prototype.

The scenarios were derived from the acquisition conditions identified in Chapter 3: coughing, talking, walking, and changing light, with running added as a higher-motion variant of walking. Each condition was represented by one scenario in each block, with different people and contexts used to reduce simple repetition while preserving the intended failure mechanism and response structure. Scenarios 5 and 10 intentionally followed a different structure by presenting multi-factor guidance, and for Scenario 10, an AI-video hallucination. Results from these scenarios were therefore treated as exploratory rather than clean estimates of the interface effect.

4.3.1 Coughing

Scenario 1

> Ms. Nora Whitcombe is a 36-year-old woman who is experiencing chest pain while seated in a living room. In the video, she appears physically uncomfortable and attempts to record an ECG using her smartwatch. The ECG recording is unsuccessful.

> Ms. Nora Whitcombe, a 36-year-old woman, is experiencing chest pain while coughing in her living room. She attempts to record an ECG using her smartwatch. The ECG recording is unsuccessful.

Scenario 6

> Mr. Marcus Hill is a 55-year-old man who is experiencing chest pain near a bathroom sink. In the video, he appears physically uncomfortable and attempts to record an ECG using his smartwatch. The ECG recording is unsuccessful.

Mr. Marcus Hill is a 55-year-old man who is experiencing chest pain while coughing near a bathroom sink. He attempts to record an ECG using his smartwatch. The ECG recording is unsuccessful.

4.3.2 Talking

Scenario 2

Ms. Natalie Foster is a 31-year-old woman who is experiencing chest pain in her bedroom. In the video, she is using a phone while attempting to record an ECG using her smartwatch. The ECG recording is unsuccessful.

Ms. Natalie Foster is a 31-year-old woman who is experiencing chest pain while speaking on the phone in her bedroom. She attempts to record an ECG using her smartwatch. The ECG recording is unsuccessful.

Scenario 7

Mr. Thomas Greystone is a 39-year-old man who is experiencing chest pain while seated in a parked car. In the video, he appears to be attempting to record an ECG using his smartwatch. The ECG recording is unsuccessful.

Mr. Thomas Greystone is a 39-year-old man who is experiencing chest pain while talking in a parked car. He attempts to record an ECG using his smartwatch. The ECG recording is unsuccessful.

4.3.3 Running

Scenario 3

Mr. Miles Fairweather is a 33-year-old man experiencing chest discomfort at the gym. In the video, he is on a treadmill while attempting to record an ECG using his smartwatch. The ECG recording is unsuccessful.

Mr. Miles Fairweather is a 33-year-old man experiencing chest discomfort at the gym. In the video, he is running on a treadmill while attempting to record an ECG using his smartwatch. His movement, elevated breathing, and possible sweat from running may interfere with the smartwatch's ability to capture a clear ECG signal. The ECG recording is unsuccessful.

Scenario 8

Ms. Helen Bennett is a 27-year-old woman who is experiencing chest pain on an outdoor path. In the video, she is physically active while attempting to record an ECG using her smartwatch. The ECG recording is unsuccessful.

Ms. Helen Bennett is a 27-year-old woman who is experiencing chest pain while jogging on an outdoor path. She attempts to record an ECG using her smartwatch while still moving. The ECG recording is unsuccessful.

4.3.4 Walking

Scenario 4

Ms. Elena Brooks is a 39-year-old woman who is experiencing chest pain in an indoor hallway. In the video, she moves through the hallway and attempts to record an ECG with her smartwatch. The ECG recording is unsuccessful.

Ms. Elena Brooks is a 39-year-old woman who is experiencing chest pain while walking through an indoor hallway. She attempts to record an ECG using her smartwatch while continuing to walk. The ECG recording is unsuccessful.

Scenario 9

Mr. Daniel Carter is a 32-year-old man who is experiencing chest pain outdoors on a sidewalk. In the video, he is moving along the sidewalk and attempts to record an ECG using his smartwatch. The ECG recording is unsuccessful.

Mr. Daniel Carter is a 32-year-old man who is experiencing chest pain while walking outdoors on a sidewalk. He attempts to record an ECG using his smartwatch while continuing to walk. The ECG recording is unsuccessful.

4.3.5 Changing Light

Scenario 5

Ms. Grace Alderwick is a 35-year-old woman who is experiencing chest pain while lying in bed in the middle of the night. In the video, she attempts to record an ECG using her smartwatch. The ECG recording is unsuccessful.

Ms. Grace Alderwick is a 35-year-old woman who is experiencing chest pain while lying in bed in the middle of the night. In the video, she attempts to record an ECG using her smartwatch while lying down in a dark bedroom. The smartwatch's position, low-light conditions, and her body position contribute to an unsuccessful ECG recording.

Scenario 10

Mr. Nathan Briarwood is a 27-year-old man who is experiencing chest pain while lying in bed in the middle of the night. In the video, he attempts to record an ECG using his smartwatch. The ECG recording is unsuccessful.

Mr. Nathan Briarwood is a 27-year-old man who is experiencing chest pain while lying in bed in the middle of the night. In the video, he attempts to record an ECG using his smartwatch while lying down in a dark bedroom. The smartwatch's position, low-light conditions, and his body position contribute to an unsuccessful ECG recording.

4.3.6 Exploratory Multi-Factor Guidance in the Changing Light Scenarios

The prompts for Scenarios 5 and 10 were intentionally modified to include guidance addressing multiple factors that might affect the recording. These scenarios were treated as exploratory multi-factor cases and were therefore not directly equivalent to the single-condition scenarios. In addition, the AI-generated video for Scenario 10 contained a visual hallucination, allowing its potential influence on participants' interpretation of the scenario to be examined

4.4 Data Analysis Plan

The goal of this analysis plan is to enable a comparison among pre-interface scenario intentions, post-prototype reported intentions, block-level interface evaluations, and pre-post changes in motivation and confidence. Pre-interface scenario data describe participants' interpretations of unsuccessful measurement situations. Post-prototype reported intention to retry the ECG measurement serves as the primary outcome of whether participants would act in response to smartwatch guidance.

Block-level and pre-post analyses assess whether the proposed interface is associated with differences in motivation, confidence, clarity, and willingness to retry compared with the existing interface. Because each scenario is viewed only once by each participant, scenario-level analyses account for differences across the ten scenario contexts, and block-level analyses are interpreted in relation to the counterbalanced interface order. Final survey items and open-ended responses are summarized descriptively to contextualize participants' decisions and identify notification features that may support clearer, safer, and more actionable smartwatch guidance. The data collected in this study were analyzed to assess how the proposed interface influenced participants' willingness to record an ECG after a failed measurement and whether it was associated with differences in motivation and confidence to attempt the measurement again.

In total, 209 responses were received to the screening and demographic questionnaire; of these, 37 people booked the study session and are referred to as participants from this point onwards. Two participants were excluded from the analysis because their responses during the study session indicated careless responding (J. L. Huang & Wang, 2021). Four participants were excluded based on the normalized eHEALS score by level of education, as suggested by Mialhe et al. (2023).

One participant canceled the study session. Of the remaining 30 participants, only the first 28 were included in the analysis because the final two were assigned the same condition order.

The data were tabulated with each row representing one participant's response to one scenario under one interface version. Variables included masked participant ID, scenario, condition, interface version, interface order, retry decision, need for additional information, motivation pre-interface, motivation post-interface, confidence pre-interface, confidence post-interface, perceived health relevance, ease of recording the ECG, and confidence in correctly recording an ECG. The interface version was coded as the proposed interface or the existing interface. Scenario represented the ten individual scenario contexts, and condition represented the five broader measurement-failure categories: coughing, talking, running, walking, and change of light. The final analysis contained all 28 participants, all 280 scenario observations, and all 56 interface-block observations.

Interface was coded as a categorical factor with levels A (proposed) and B (existing). Sequence was coded as AB or BA, period as first or second block, condition as one of the five failure categories, and scenario as one of ten contexts. Retry intention was coded as a binary outcome. The need for additional information was retained as a separate binary outcome and was not combined with retry intention.

Participant demographics were aggregated and reported descriptively, including age group, highest level of education completed, ownership of a smartwatch capable of ECG recordings, duration of smartwatch use, number of previous ECG recordings, baseline confidence in recording an ECG, and eHEALS score. For each scenario, participants first viewed an AI-generated video, read the scenario text, and responded to questions before receiving interface guidance. These pre-interface responses were analyzed as measures of scenario interpretation.

Participants were assigned to one of two counterbalanced interface orders. Half of the participants completed scenarios 1-5 with the existing interface and scenarios 6-10 with the proposed interface. The remaining participants completed scenarios 1-5 with the proposed interface and scenarios 6-10 with the existing interface. Therefore, each participant experienced both interface types, but each scenario was viewed only once. Interface type was treated as a within-subject factor at the participant level, and interface order was treated as a between-subjects factor. The scenario was included in scenario-level analyses to account for differences across the ten scenario contexts.

Responses about the perceived cause of the unsuccessful measurement were summarized using frequencies and percentages. For each scenario, the number and percentage of participants who selected the intended cause of failure were reported. Correct identification was also summarized by condition to examine whether some types of unsuccessful measurement were easier to interpret than others. Participants' immediate intended action after the unsuccessful measurement, before seeing guidance from the device, was also summarized descriptively.

Pre-interface items assessed motivation to record the ECG, perceived importance of the scenario, perceived ease or difficulty of recording the ECG in the scenario, and confidence in recording the ECG correctly after the initial unsuccessful measurement. These items were used to describe how participants interpreted the scenarios before receiving interface guidance and to support baseline checks of the scenarios. Because these items were collected before participants interacted with the smartwatch interface, they were interpreted as measures of scenario appraisal rather than as effects of the interface.

After completing the pre-interface questions, participants viewed a second version of the scenario in which the cause of the unsuccessful measurement was disclosed and then interacted with the smartwatch prototype. The primary reported intended-action outcome was post-prototype retry intention. These variables represented participants' reported intention to retry or record the ECG; it did not measure a completed real-world ECG recording. Participants' responses were recoded into a binary retry-intention variable indicating whether they intended to retry or record the ECG after the unsuccessful measurement notification. Frequencies and percentages were reported by interface type and scenario. The need for additional information was analyzed separately as a secondary binary outcome rather than combined with the retry-intention variable.

To evaluate whether the proposed interface was associated with a higher likelihood of retrying the measurement, paired analyses were conducted to compare outcomes between Interface A (the proposed interface) and Interface B (the existing interface). These comparisons included the primary retry-intention outcome, exact binary retry-intention rates, secondary post-interface ratings, baseline-scenario checks, and pre- and post-change scores for motivation and confidence. Because scenario-level sample sizes were small, scenario-level results were interpreted primarily using descriptive percentages and percentage-point differences. The analyses were conducted both overall and separately by condition or scenario to examine whether the effect of the proposed interface differed across unsuccessful-measurement contexts.

After each block of scenarios, participants completed interface-level evaluations. These items assessed how much the unsuccessful-measurement notification grabbed their attention, how clearly the corrective action was displayed, how much the interface content made them want to retry the ECG measurement, what aspects of the interface were lacking, how motivated they felt to record an ECG in similar situations after interacting with the interface, and how confident they felt in recording an ECG correctly after interacting with the interface. Descriptive statistics were calculated for each item by interface type.

To examine changes from before to after interface exposure, pre-post comparisons were conducted for constructs measured at both points. For each scenario and interface type, motivation and confidence change scores were calculated by subtracting the pre-interface rating from the post-interface rating. Mean change scores were then summarized and compared descriptively between the proposed and existing interfaces. Because each participant viewed each scenario only once, these analyses were interpreted as scenario- and block-level pre-post comparisons rather than repeated evaluations of identical scenarios.

Responses that addressed which aspects of the interfaces were lacking or what would have made the interfaces easier to understand, easier to follow, or more useful were summarized descriptively. These responses were used to contextualize participants' interface-level ratings and retry decisions, but no formal qualitative analysis was conducted.

End-of-study responses were also summarized descriptively. These items examined what influenced participants' decisions to retry or not retry an ECG measurement after an unsuccessful attempt, their confidence in recording an ECG after interacting with the interfaces, the scenarios in which corrective actions were most difficult to follow, and whether the existing or proposed interfaces were perceived as more motivating or helpful. Responses regarding proposed interface features that the users found helpful, reasons for retrying, barriers to retrying, and suggested improvements to smartwatch health notifications were reported descriptively to provide additional context for participants' decision-making and to identify practical considerations for future smartwatch notification systems.

Participants' perceptions of the AI-generated videos used in the study protocol were analyzed as secondary measures of scenario presentation quality. Four Likert-scale items assessed how realistic the videos felt, how helpful they were for understanding the scenario, how easy it was for participants to place themselves in the conditions presented, and how clearly the videos conveyed the scenario.

These items were summarized using means and standard deviations. These analyses were used to determine whether the AI-generated videos were perceived as clear and useful for supporting participants' understanding of the unsuccessful-measurement scenarios, while also identifying limitations related to realism and ecological validity.

The required sample size was based on the primary within-subject comparison between the existing and proposed interface conditions. Assuming a two-tailed alpha level of .05, statistical power of .80, and an expected standardized within-subject effect size of $dz = 0.55$, the required sample size was 28 participants to detect a moderate within-subject intervention effect. Although participants were divided into two counterbalanced order groups, each participant completed both the existing and proposed interface conditions. The final analyses were interpreted with caution because scenario-level comparisons involved fewer observations, and the observed effects varied by scenario.

4.4.1 Order, Period, Learning, and Primacy

Potential learning, fatigue, and first-option preference were evaluated in four ways: the period term in the primary model, an interface-by-sequence sensitivity model, adjusted second-versus-first-period models for the block outcomes, and first period-only comparisons between participants who received A first and those who received B first. Interface preference was also reported separately for the two sequence groups. Because only 14 participants were assigned to each sequence, these analyses were treated as sensitivity checks rather than definitive carryover tests.

4.4.2 Prior Experience, Observed Behavior, and Other Confounders

Pre-study smartwatch and ECG experience was summarized descriptively. However, the linked participant-level analytic export did not contain a usable pre-exposure watch-experience variable; consequently, no experience-adjusted model or sequence-balance analysis for prior experience could be estimated. The post-study ECG-confidence item was not substituted as a baseline covariate. This limitation is reported explicitly because prior experience may affect both task performance and interface preference.

Prototype actions were annotated during the sessions, but the observed action was not digitized as a participant-by-scenario variable in the analytic file. Formal agreement analyses between observed behavior and stated retry intention were therefore not estimable. The principal outcome in this chapter is reported retry intention, not a completed real-world ECG recording.

4.5 Results

4.5.1 Participant Demographics

4.5.1.1 Age Group

Of the 28 participants included in the data analysis, 7 were between 18 and 24 years of age, 20 were between 25 and 34 years of age, and 1 was between 45 and 54 years of age.

4.5.1.2 Level of Education

2 participants had started their bachelor's degree but had not received a degree up to the point of the study, 18 participants had obtained a university bachelor's degree, and 8 participants had obtained a graduate or professional degree after a bachelor's.

4.5.1.3 Prior Experience with ECG

Participants' prior experience with ECG functionality varied. A total of 10 participants (36%) reported owning a smartwatch capable of recording single-lead ECGs. Among participants who reported prior smartwatch use, duration of use varied: three participants reported using a smartwatch for more than 3 years (n = 3, 30.0%), two reported using one for less than 6 months (n = 2, 20.0%), two reported using one for more than 1 year (n = 2, 20.0%), two reported using one for more than 5 years (n = 2, 20.0%), and one reported using one for 6 months up to a year (n = 1, 10.0%).

Three participants reported never having recorded an ECG (n = 3, 30.0%), two reported recording an ECG fewer than five times (n = 2, 20.0%), two reported recording an ECG between 5 and 10 times (n = 2, 20.0%), one reported recording an ECG daily (n = 1, 10.0%), one weekly (n = 1, 10.0%), and one monthly (n = 1, 10.0%).

Baseline confidence was measured on a 0–10 scale. Using NPS-style cutoffs, three participants were classified as detractors (0–6), 15 as passives (7–8), and 10 as promoters (9–10). The proportion classified as promoters exceeded the proportion classified as detractors by 25.0 percentage points. Because the item measured confidence rather than likelihood to recommend, this value should not be interpreted as a conventional Net Promoter Score.

4.5.1.4 eHealth Literacy (eHEALS Score)

The mean eHEALS score was M = 32.68 (SD = 4.53), with scores ranging from 24 to 40. eHEALS scores were classified using the following thresholds: low digital health literacy = 8-23, moderate digital health literacy = 24-32, and high digital health literacy = 33-40. Based on these thresholds, 15 were classified as having moderate digital health literacy (n = 15, 53.6%), and 13 as having high digital health literacy (n = 13, 46.4%). Based on the exclusion criteria, no participants with low digital health literacy were included in this study's analysis.

4.5.2 Scenario Interpretation Before Interface Interaction

Across the 10 scenarios, the percentage of participants who selected the intended cause ranged from 35.71% to 92.86%. Participants' perceived causes of the unsuccessful ECG measurements are summarized by scenario in Table 9.

Table 9. Participants' Perceived Causes of The Unsuccessful ECG Measurements

Potential Reason	Scenarios									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Expected response	22	18	25	20	10	11	24	26	26	11
The person was moving	21	8	17	20	1	11	5	16	19	1
The person had recently exercised or was still physically active	0	0	24	3	0	0	0	25	5	0
The watch was loose or not positioned correctly	6	8	7	3	15	4	3	6	4	21
The person may not have been holding still	15	12	13	17	4	15	10	12	21	4
The smartwatch or sensor may not have worked properly	9	11	9	5	17	10	6	6	5	13
The person's symptoms or discomfort may have affected the attempt	9	3	2	13	1	18	9	6	3	3

Responses were classified as selection of the intended cause of failure or not, per condition, as summarized in Table 10. Across the two coughing scenarios, 33 of 56 scenario responses (58.93%) identified coughing as the intended contributing condition. Identification decreased from 22 of 28 participants (78.57%) in Scenario 1 to 11 of 28 participants (39.29%) in Scenario 6. Because no inferential comparison between these scenarios was reported, this difference should be interpreted descriptively.

Besides that, changes in light scenarios had the lowest rate of correct identification of the likely contributing factor in the unsuccessful measurement. These results indicate that some unsuccessful measurement contexts were interpreted more accurately than others before participants received smartwatch guidance.

Table 10. Correct Identification of The Likely Contributing Factor in the Unsuccessful Measurement per Condition

Condition	Mean n per scenario	%
Coughing	16.5	58.93%
Talking	21	75.00%
Running	25.5	91.07%
Walking	23	82.14%
Changing Light	10.5	37.50%

Note. Each condition was represented by two scenarios with 28 responses per scenario. Mean n is the mean number of participants per scenario who selected the intended contributing condition; percentages are based on 28 participants per scenario.

Participants' immediate intended actions after the unsuccessful measurement, but before seeing interface guidance, are summarized in Table 11. The most frequently reported immediate action was to stop moving and try again, selected 79 times by participants across 10 scenarios. The next most common response was "Sit down or rest first, then try again," selected 58 times across the ten scenarios. These responses suggest that prior to interface guidance, participants identified motion as the likely contributing factor in the unsuccessful measurement.

Table 11. Immediate Intended Action before Interface Interaction

Scenario	1	2	3	4	5	6	7	8	9	10
Try recording the ECG again immediately	8	3	1	2	8	1	7	2	1	3
Stop moving and try again	8	6	11	15	1	3	7	12	14	2
Sit down or rest first, then try again	4	5	8	7	1	6	7	13	7	0
Adjust or tighten the watch, then try again	2	7	4	0	10	0	2	0	0	11
Wait and try again later	1	0	2	0	5	2	2	1	2	4
Ask someone nearby for help	2	3	0	1	2	4	0	0	2	0
Contact a health care professional	1	2	0	1	0	5	0	0	0	0
Seek emergency help	1	2	1	2	1	7	1	0	0	2
Do not try again	0	0	1	0	0	0	2	0	2	6
I am not sure	1	0	0	0	0	0	0	0	0	0

4.5.3 Reported action after the interaction with the interface (RQ3.1 and RQ3.4)

Paired analyses were conducted to compare outcomes between Interface A (proposed interface) and Interface B (existing interface). The comparisons included the primary retry-intention outcome, exact binary retry-intention rates, secondary post-interface ratings, baseline-scenario checks, and pre- and post-change scores for motivation and confidence. These outcomes were examined both overall and separately by condition/scenario. Across outcomes, there was no statistically significant evidence that Interface A outperformed Interface B using one-sided t-tests.

A few outcomes showed small descriptive advantages for A, including higher attention ratings and greater change in confidence in the walking condition, but these effects were not statistically significant. The primary “retry” intention more often descriptively favored B or showed no difference, particularly for the talking condition and the overall participant-level retry measure. A condition-level analysis is presented below to examine how the pattern of results differed across scenarios.

4.5.3.1 Scenario 1 - Coughing

Scenario 1 showed a small descriptive advantage for the proposed interface. Retry intention was 78.6% under the proposed interface and 71.4% under the existing interface, a 7.1 percentage-point difference favoring the proposed interface. Participants were also less likely to report needing additional information under A than under B, 7.1% versus 14.3%. Motivation change was similar across versions, with the proposed interface showing a mean increase of 0.43 and the existing interface showing a mean increase of 0.36. The clearest descriptive difference was for confidence change: confidence increased under A (MΔ = 0.64) but decreased under B (MΔ= -0.36). Thus, Scenario 1 provides one of the more favorable descriptive profiles for the proposed interface, particularly for improvement in participants' confidence about recording an ECG correctly.

4.5.3.2 Scenario 2 - Talking

Scenario 2 showed a mixed pattern. Retry intention was somewhat higher under the existing interface, 78.6%, than under the proposed interface, 71.4%, a 7.1 percentage-point difference favoring the existing interface. Participants were also more likely to report needing additional information under the proposed interface than under the existing interface, 28.6% versus 14.3%, suggesting that the existing interface may have been slightly more appropriate in this scenario.

However, the change in confidence favored the proposed interface: confidence increased by 0.43 points under A, compared with 0.14 points under the existing interface. Motivation declined slightly under both versions, with a marginally larger decline under the proposed interface. Participants' comments suggested that they expected smartwatch measurements to fail occasionally and were willing to retry once without further explanation. Accordingly, the higher retry intention and lower information needs observed with the existing interface should not be interpreted as evidence that it supported retrying more effectively. The proposed interface showed a descriptive advantage in confidence change

4.5.3.3 Scenario 3 - Running

Scenario 3 favored the proposed interface on retry intention but showed a mixed attitudinal pattern. Retry intention was 71.4% under the proposed interface and 57.1% under the existing interface, a 14.3 percentage-point difference favoring the proposed interface. The proportion of participants needing additional information was identical across versions, 21.4% under both the proposed and existing interface. Confidence changes also favored the proposed interface, with confidence increasing by 0.86 points under the proposed interface compared with 0.43 points under the existing interface.

However, motivation declined more under the proposed interface (MΔ= -0.86), than under the existing interface (MΔ = -0.50). Thus, Scenario 3 suggests a possible trade-off: the proposed interface was associated with higher intention to retry and greater confidence improvement, but also with a larger decline in motivation.

4.5.3.4 Scenario 4 - Walking

Scenario 4 showed one of the clearest descriptive trade-off patterns. Retry intention favored the existing interface, with 71.4% of participants intending to retry under the existing interface compared with 64.3% under the proposed interface. Participants were also less likely to report needing additional information under the existing interface than under the proposed interface, 7.1% versus 21.4%. However, confidence changed strongly in favor of the proposed interface: confidence increased by 0.29 points under the proposed interface but decreased by 0.86 points under the existing interface. This was one of the largest differences in scenario-level confidence change, favoring the proposed interface. Motivation declined under both versions, with similar decreases in both the proposed and existing interfaces. Thus, Scenario 4 descriptively favored the existing interface for retry intention and information sufficiency, while the proposed interface showed the larger improvement in confidence.

4.5.3.5 Scenario 5 - Changing Light

Scenario 5 showed identical reported retry intention across interfaces. Retry intention was 78.6% for both the proposed and existing interfaces. However, participants were more likely to indicate that they needed additional information under the proposed interface than under the existing one (28.6% versus 7.1%), suggesting that the existing interface may have been clearer or more complete in this scenario. Motivation and confidence changes slightly favored the proposed interface: motivation declined by 0.29 points under the proposed interface compared with 0.43 points under the existing interface, and confidence declined by 0.14 points under the proposed interface compared with 0.36 points under the existing interface. Overall, Scenario 5 suggests no difference in retry intention, with the existing interface appearing clearer, while the proposed interface shows slightly more favorable attitudinal change under the changing-light condition.

4.5.3.6 Scenario 6 - Coughing

Scenario 6, the second scenario within the coughing condition, also showed a small descriptive retry advantage for the proposed interface. Retry intention was 71.4% under the proposed interface and 64.3% under the existing interface, a 7.1 percentage-point difference favoring the proposed interface. The need for additional information was also lower under the proposed interface than under the existing interface, 7.1% versus 14.3%. However, the pre-post change outcomes favored the existing interface. Motivation declined less under the existing interface ($M\Delta$= -0.64) than under the proposed interface ($M\Delta$ = -0.86), and confidence increased more under the existing interface ($M\Delta$= 1.43) than under the proposed interface ($M\Delta$ = 0.71). Thus, Scenario 6 suggests that the proposed interface was associated with higher reported retry intention and reduced information needs, whereas the existing interface may have produced stronger affective gains.

4.5.3.7 Scenario 7 - Talking

Scenario 7 appears to be the strongest contributor to the overall advantage of the existing interface in the talking condition for retry intention. Retry intention was 50.0% under the proposed interface and 85.7% under the existing interface, a 35.7 percentage-point difference favoring the existing interface. This was the largest scenario-level difference in retry intention in the dataset. However, the difference was not statistically significant using Fisher's exact test, so it should be interpreted as a descriptive pattern rather than reliable evidence of an interface effect.

The need for additional information was identical across versions: 14.3% for both the proposed and existing interfaces. Motivation change, however, favored the proposed interface: motivation was nearly stable under the proposed interface (MΔ = -0.07), but declined more under the existing interface (MΔ = -0.57). Confidence change was identical across versions, with both the proposed and existing interfaces showing a mean increase of 0.57. Overall, Scenario 7 suggests that the existing interface was descriptively stronger for retry intention, whereas the proposed interface appeared somewhat better at preserving motivation.

4.5.3.8 Scenario 8 - Running

Scenario 8 showed a more consistently favorable descriptive pattern for the existing interface. Retry intention was 71.4% under the existing interface and 57.1% under the proposed interface, a 14.3 percentage-point difference favoring the existing interface. The need for additional information was identical across versions: 7.1% for both the proposed and existing interfaces. Motivation declined less under the existing interface (MΔ= -0.14) than under the proposed interface (MΔ= -0.43). Confidence also increased more under the existing interface (MΔ= 1.00) than under the proposed interface (MΔ= 0.36). Scenario 8, therefore, provides one of the clearer descriptive profiles favoring the existing interface across retry intention, motivation change, and confidence change.

4.5.3.9 Scenario 9 - Walking

Scenario 9 also favored the existing interface descriptively. Retry intention was 85.7% under the existing interface and 71.4% under the proposed interface, a 14.3 percentage-point difference favoring the existing interface. The need for additional information was identical across versions: 7.1% for both the proposed and existing interfaces. Motivation declined less under the existing interface ($M\Delta = -0.14$) than under the proposed interface ($M\Delta = -0.43$). Confidence change was positive for both versions, but slightly larger under the existing interface ($M\Delta = 0.64$) than under the proposed interface ($M\Delta = 0.50$). Together with Scenario 4, this suggests that the walking condition's retry pattern was generally more favorable for the existing interface, although Scenario 4 showed a separate confidence-change advantage for the proposed interface.

4.5.3.10 Scenario 10 - Changing Light

Scenario 10 showed no difference in retry intention between interfaces. Retry intention was 64.3% under both the proposed and existing interfaces, and the need for additional information was identical at 14.3% under both versions. However, motivation and confidence change favored the existing interface. Motivation increased by 0.43 points under the existing interface, compared with 0.07 points under the proposed interface, and confidence increased by 1 point under the existing interface, compared with 0.07 points under the proposed interface. Thus, Scenario 10 showed no difference in reported retry intention but was descriptively more favorable for the existing interface on motivation and confidence change.

4.5.4 Interface Preference and Perceived Helpfulness (RQ3.2)

Participants were asked which interface they preferred for motivation and helpfulness when deciding what to do after an unsuccessful measurement. Despite the absence of a measurable retry advantage, 23 of 28 participants (82.1%) identified the proposed interface as the more motivating or helpful option. Three participants considered the interfaces equally helpful, one preferred the existing interface, and one was unsure. Preference was similar across sequences.

Among participants assigned to Sequence AB, 12 preferred the proposed interface, one considered both interfaces equally helpful, and one preferred the existing interface. Among participants assigned to Sequence BA, 11 preferred the proposed interface, two considered both interfaces equally helpful, and one was unsure.

Furthermore, participants were asked about the specific aspects of the proposed interface that they considered improved their decision-making process. As shown in Figure 7, the most frequently selected helpful aspects were that the proposed interface explained what action to take after the unsuccessful measurement (25 participants), it was perceived as providing a possible explanation of why the measurement may have failed (25) and provided clear next steps (19).

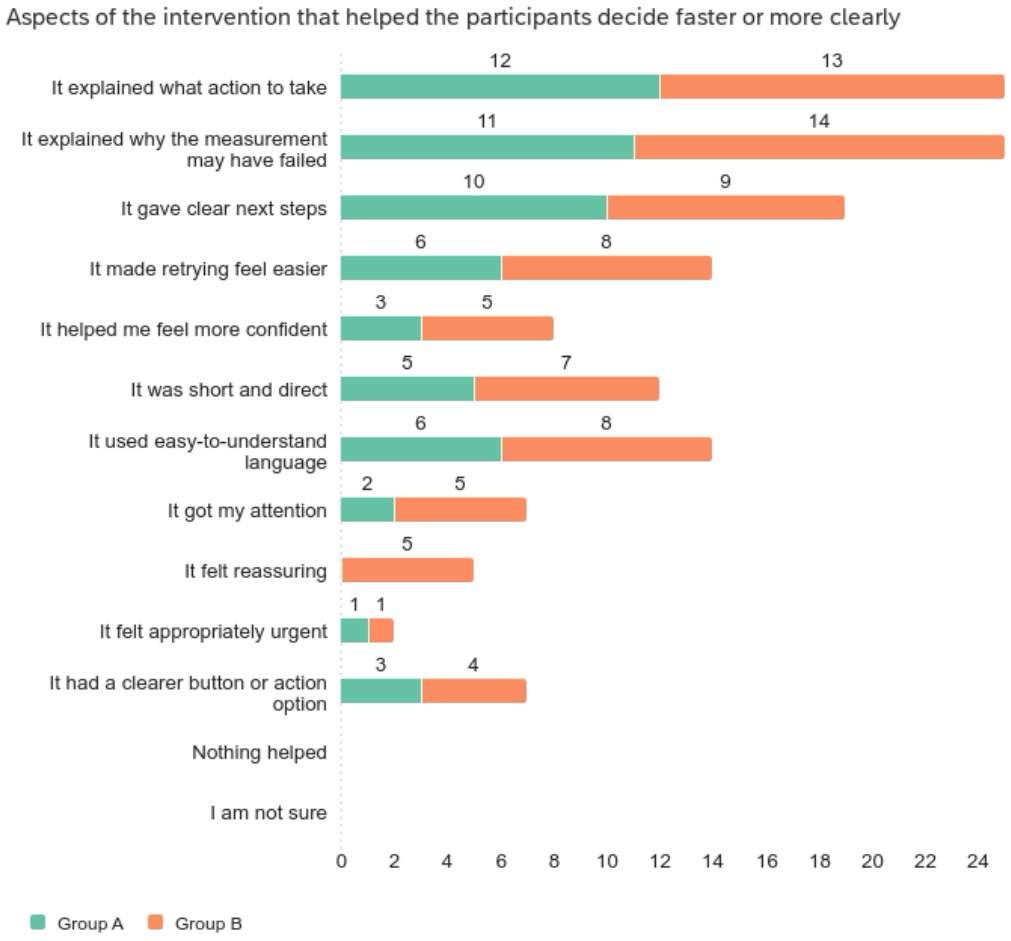


Figure 7. Aspects of the intervention that improved the participants' decision-making process

4.5.5 AI-generated Videos in User-Centered Study Knowledge Tasks (RQ3.5)

Participants' ratings of the AI-generated videos are summarized in Figure 8. The videos were rated as moderate in realism, M = 3.11, SD =0.94. Participants rated the videos as high in helpfulness for understanding the scenario, M = 3.93, SD = 1.03, ease of placing themselves in the scenario, M = 3.68, SD = 0.97, and clarity of scenario communication, M = 4.0, SD = 0.96. These results suggest that the AI-generated videos were moderately effective in supporting participants' understanding of the unsuccessful-measurement scenarios.

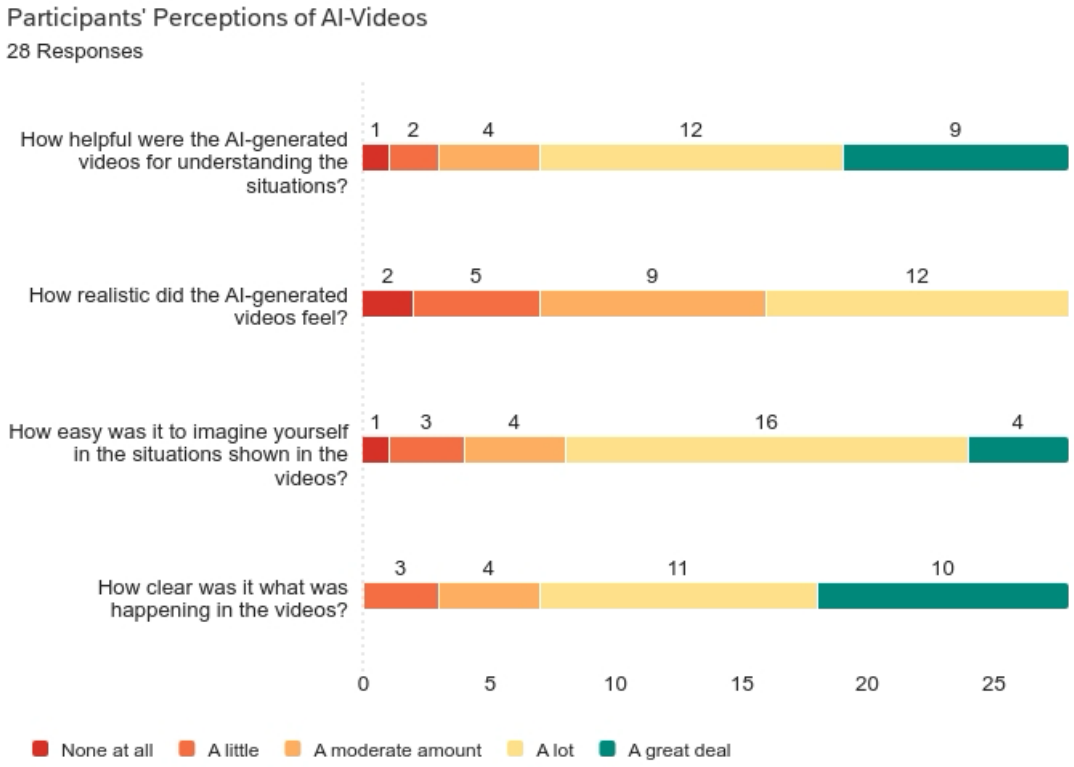


Figure 8. Participants' Perceptions of AI-Generated Videos

4.5.6 Order, Period, Scenario, and Confounding Factors

The adjusted primary model found no evidence of a general period effect (second versus first: OR = 0.86, 95% CI [0.52, 1.44], p = .573) or sequence effect (BA versus AB: OR = 0.75, 95% CI [0.45, 1.25], p = .268). The interface-by-sequence interaction was also not significant (interaction OR = 0.72, 95% CI [0.27, 1.93], p = .519). Across eleven block outcomes, no second-versus-first period effect remained significant after Holm correction; all adjusted p values were 1.000.

In the first period, before exposure to the other interface, mean retry intention was 72.9% among participants assigned A first and 71.4% among those assigned B first, a 1.4 percentage-point difference (95% CI [-17.0, 19.9], p = .875). The adjusted first-period model was likewise null (OR = 1.09, 95% CI [0.53, 2.25], p = .816). Replacing period with the ten scenarios yielded an interface OR of 0.80 (95% CI [0.48, 1.34], p = .398), showing that the primary conclusion was not explained by scenario context.

The leave-one-participant-out analysis produced 28 successful refits. Every refit favored Interface B directionally; the OR for A versus B ranged from 0.66 to 0.93, with no direction reversals. This indicates that the negative direction of the adjusted interface estimate was not caused by a single participant. Nevertheless, the small sequence groups and wide interaction interval mean that subtle learning, carryover, or primacy effects cannot be ruled out.

4.6 Ethics Clearance

This study has been reviewed and received ethics clearance through the University of Waterloo Research Ethics Board (REB 47991). The certificate of clearance is presented in Appendix B.5.

Chapter 5
Discussion

This chapter provides a comprehensive overview of the prototype design process, which uses AI classifiers for PPG quality assessment, explainable AI techniques, counterfactuals, and large language models to prompt user actions that mitigate issues in sensor-recorded signals affecting data quality. It then discusses the results of the human validation study of this prototype. Finally, it addresses the research questions, contributions, limitations, and suggests directions for future research to build upon these findings.

5.1 Developing AI models for Digital Health Technologies

5.1.1 Dataset Limitations

Although the PPG signals in this dataset were recorded with smartphones, it was selected for the project because it includes high-quality annotations from experts in heart rate estimation. Each PPG recording was independently assessed by three to five experts, who estimated heart rate from the PPG signal alone, without access to the corresponding ECG. The expert estimates were then compared with the reference heart rate derived from the ECG. A signal was considered "good" only when the estimated heart rate differed from the ECG reference by less than or equal to 5 bpm. A recording was labeled as good quality, or 1, only if the majority of annotators satisfied this condition, meaning at least 2 out of 3 annotators or 3 out of 5 annotators provided "good" annotations. If this condition was not met, the recording was labeled as poor quality (0).

5.1.2 False Warnings

The full tabular experiment had the strongest overall classifier. The best practical model was the MLP after hyperparameter optimization with RandomizedSearchCV, which achieved the highest balanced accuracy and AUROC while maintaining high bad-quality recall. This suggests that richer tabular information helped the model distinguish between bad-quality and good-quality PPG more effectively than motion-condition variables alone.

The full tabular results indicate that predictive information about the PPG quality label extends beyond visible movement and measurement context. Features such as heart rate, blood pressure, measurement site, gender, height, weight, and glycemia were associated with the quality label and contributed useful information for classification. This is consistent with the full tabular model's higher balanced accuracy and AUROC compared with the motion-only models. However, these findings demonstrate predictive associations and do not establish that these features causally influence PPG quality.

However, from the digital health technology perspective, the full tabular model has a practical limitation. Many of the features that help classification are not directly controllable by the user during measurement. A user cannot immediately change their height, gender, body weight, or baseline physiology. Therefore, the full tabular model is more useful for detecting bad-quality PPG than for generating simple corrective feedback.

From this experiment, it was also evident why a low-quality F1-score should not be interpreted in isolation: the MLP with the PCA model achieved the highest low-quality F1-score, but it failed to identify any good-quality samples. In a real user-facing system, this would produce excessive false alarms because all high-quality measurements would be treated as bad. Therefore, MLP RandomizedSearchCV is better than MLP + PCA RandomizedSearchCV, even though its bad-quality F1-score is slightly lower.

5.1.3 Model Training Effect of Motion-Only Features

Experiment 2 showed that movement and measurement-context variables are relevant but insufficient for robust PPG quality classification. The best balanced accuracy was only about 0.608, and many models collapsed into predicting all samples as bad quality. The main weakness of the best motion-only balanced-accuracy model was low bad-quality recall. Although Logistic Regression with RandomizedSearchCV had high bad-quality precision, it detected only 232 of 668 bad-quality samples. This means that many bad-quality PPG recordings would pass without warning. For a signal-quality control system, this is a grave limitation because missed bad-quality signals could lead to unreliable downstream measurements.

However, the motion-only models remain valuable because their features are directly actionable. If a motion-only explanation identifies walking, coughing, talking, or changing light contributing to poor signal quality, the corrective feedback can be translated into a simple user prompt.

For example, the system could instruct the user to remain still, stop talking, avoid coughing during measurement, or stabilize the lighting condition. Therefore, the motion-only experiment is weaker as a classifier but stronger as a model to generate user-facing prompts with significant safeguards to address the limitations on prediction effectiveness.

5.1.4 Model Training Effect of Dataset Balancing

The balanced motion-only experiment showed that under-sampling the majority class did not substantially improve overall classification. The best balanced accuracy increased only slightly from 0.608 in the unbalanced motion-only experiment to 0.609 in the balanced motion-only experiment.

Balancing the training data reduced the number of all-bad predictors from 24 models to 1 model. This shows that class balancing helped prevent some models from simply predicting the majority class. However, the best balanced models still had bad-quality recall. For example, Logistic Regression GridSearchCV detected only 207 of 668 bad-quality samples and missed 461 bad-quality samples.

This suggests that balancing improved recognition of good-quality cases came at the cost of missing many bad-quality cases. In a PPG signal-quality application, this trade-off is problematic because the system's primary purpose is to detect unreliable signals before they are accepted. The balanced motion-only experiment, therefore, confirms that class imbalance was not the only reason for weak motion-only performance.

5.1.5 The Trade-off of Actionability

Across the three experiments, there is a clear trade-off between predictive strength and actionability. The full tabular model is the strongest option for predicting whether the PPG signal is of bad quality. It achieved the best balanced accuracy, AUROC, and a good-quality F1-Score. However, its explanations may rely on features that are not immediately controllable by the user.

The motion-only models are weaker predictors, but their explanations are easier to convert into corrective feedback. Movement conditions such as walking, coughing, talking, and changing light can be directly translated into user-facing triggers for corrective actions. The balanced motion-only models sit between these two purposes.

They reduce the tendency toward all-bad prediction, but they do not substantially improve predictive performance. They remain useful for actionability but are not strong enough to serve as the main classifier. The results, therefore, suggest that the best practical system should combine the strengths of the experiments. A full tabular model could be used to detect likely bad-quality PPG, while motion-based explanations could be used to generate user-facing corrective prompts when relevant movement or context conditions are present.

5.1.6 Dimensionality Reduction

PCA and LDA did not consistently improve performance. In the full tabular experiment, PCA yielded the highest bad-quality F1-score, but this was due to a model that failed to classify good-quality cases. LDA produced several moderate models but did not outperform the best unreduced MLP. In the motion-only experiments, PCA and LDA had limited value because the feature set was already small, binary, and low-dimensional. Dimensionality reduction did not solve the main problem, which was the limited information in the one-hot-encoded motion-only variables.

5.1.7 Best Performing Model

Across the three experiments, the best-performing practical model was MLP with RandomizedSearchCV hyperparameter tuning from the full tabular-feature experiment. This model was selected because it provided the strongest overall balance between detecting bad-quality PPG signals and avoiding excessive false warnings. It achieved an accuracy of 0.781, balanced accuracy of 0.657, bad-quality precision of 0.850, bad-quality recall of 0.874, bad-quality F1-score of 0.862, good-quality F1-score of 0.464, AUROC of 0.708. It was also not flagged as a potentially overfitted model.

The model's confusion matrix showed that it correctly detected 584 of 668 bad-quality samples, while missing 84 bad-quality samples by classifying them as good quality. It also correctly identified 81 of 184 good-quality samples but incorrectly flagged 103 good-quality samples as bad quality. This means that the model was reasonably strong at detecting unreliable PPG, which was the main practical aim of the study, but it still produced a considerable number of false warnings.

The MLP model was selected over models with slightly higher accuracy or bad-quality F1-score because those models often achieved high scores by over-predicting the majority class. For example, some SVM, KNN, and MLP + PCA models classified nearly all test samples as of poor quality. Since the dataset was imbalanced, with most samples belonging to the bad-quality class, these models appeared strong when judged solely by accuracy or bad-quality F1-Score. However, they failed to recognize good-quality PPG and therefore had poor balanced accuracy and a good-quality F1-score. In contrast, the chosen model provided a more useful trade-off because it detected most bad-quality samples while still identifying some good-quality samples correctly. However, it had two important limitations.

First, the model still produced classification errors, especially for the good-quality class. It correctly detected most bad-quality PPG recordings, but it also incorrectly flagged many good-quality recordings as bad quality. This means that, in a real user-facing system, the model could generate unnecessary warnings and ask users to repeat measurements even when the PPG signal was acceptable. At the same time, the model also missed some bad-quality recordings by classifying them as good quality. Therefore, although it was the strongest model in this study, its performance was not reliable enough for deployment without further improvement and validation.

Second, the model had limitations in terms of explainability and corrective feedback. Because it was trained using the full tabular feature set, the model relied not only on actionable movement variables but also on demographic and physiological features such as gender, height, weight, blood pressure, glycemia, heart rate, and measurement site.

These features may help the model predict PPG quality, but many of them cannot be changed by the user during measurement. This limits the usefulness of the explanations for generating real-time corrective feedback. In addition, the DiCE counterfactual analysis showed that changing the movement condition to rest was often insufficient to shift the model's prediction from bad quality to good quality.

This means that counterfactual explanations could not always be found for the best full tabular model, because the model's decision was often driven by non-actionable features rather than by movement alone.

5.2 Design of prompts to guide users in corrective actions for improving PPG signal quality

The findings suggest that LLM-generated facilitator prompts can support the transformation of technical explainable AI outputs into actionable user-facing guidance, but only to a moderate extent. GPT-5.5 produced clearer and more consistent prompt drafts than Llama 3.1:8B, especially for conditions such as coughing, talking, running, walking, poor sensor contact, and changing light environments. These outputs show that LLMs can act as a useful translation layer between XAI methods and users by converting technical evidence, such as feature-level explanations, probabilities, SHAP, LIME, or DiCE outputs, into simpler guidance about which features influenced the model's classification and what the user might try next.

When evaluated through the QUEST health care principles, GPT-5.5 showed stronger performance in clarity, relevance, information quality, and actionability. Its prompts generally connected the identified signal-quality issue to practical corrective actions, such as waiting after coughing, avoiding speech during recording, keeping the wrist still, improving sensor contact, or moving to steadier lighting conditions. However, the QUEST evaluation also showed that the model outputs were not fully reliable. Some GPT-5.5 prompts were vague, generic, or overly directive, while Llama 3.1:8B showed more frequent weaknesses, including missing rendered prompts, technical phrasing, demanding calls to action, unsupported assumptions, and occasional blame-like wording. These issues reduce alignment with QUEST areas, including expression style, safety, harm avoidance, reasoning, and trust.

The need for human revision before translating the prompts into the prototype was therefore essential. Although the LLMs generated useful drafts, the prompts had to be reviewed and rewritten to ensure that they were safe, grounded, non-blaming, non-coercive, and appropriate for real users. This was particularly important because users may interpret a poor-quality physiological recording as a health problem rather than as a recording-quality issue.

Human revision helped clarify that the issue was related to the measurement conditions and added safer, more supportive wording, such as “stop safely,” “wait until your breathing slows,” and “avoid speaking until the recording is complete.” These changes improved the prompts’ alignment with QUEST principles and with the Fogg Behavior Model by helping users retry successfully without creating confusion, pressure, or unnecessary concern.

5.3 Human Validation of Model-Informed Counterfactual Prompts for Supporting Reported Corrective-Action Intentions

The second project in this thesis used a human-validation study to assess whether the generated prompts increased participants’ reported intention to take corrective action. The outcome was operationalized as reported intention to initiate a single-lead ECG recording after an unsuccessful measurement notification. ECG recordings are also highly sensitive to movement, and specific conditions must be met for the reading to succeed. Because the ECG feature requires users to actively engage with the interface and be sufficiently motivated to complete the recording, it provides an appropriate use case for evaluating whether prompts can effectively increase users’ ability and willingness to take corrective action. Overall, the results suggest that the proposed interface was perceived positively by participants and appeared useful for explaining failure, supporting perceived understanding, and, in some scenarios, improving confidence, but these benefits did not consistently translate into higher intention to retry.

The overall effect of the facilitator prompts on reported retry intention was small. Across the ten scenarios, retry intention was approximately 67.9% under the proposed interface and 72.9% under the existing interface. This corresponds to a difference of -5.0 percentage points and Cohen’s h = -0.11. Under conventional interpretation, this is a small descriptive effect favoring the existing interface over the proposed interface. Consequently, H1 for RQ3.1 was not supported. Participants’ comments suggested that they viewed occasional smartwatch measurement failures as normal and were generally willing to make one retry. Their willingness to retry again, however, depended on the context. Furthermore, with the proposed interface, if participants believed there was not a compelling reason to get an ECG and a more pressing action was needed (for example, scenario 9 depicts a person walking late to work), they would not follow the proposed interface’s recommendation.

In some scenarios, being given a more specific explanation may have made users more cautious rather than more likely to retry immediately, for example, when the people in the video were coughing. This distinction is important because, in the context of health-related self-measurement, a successful interface should not only maximize retry behavior. It should also help users decide when retrying is appropriate, when they should adjust their behavior or device, and when they may need more information or support.

5.3.1 Scenario-level Analysis

The scenario-level results show that the condition-level findings mask meaningful descriptive variation across the 10 individual scenarios. The version of the proposed interface appeared strongest in Scenarios 1 and 3, where it showed higher intentions to retry and stronger gains in confidence. The existing interface appeared strongest in Scenarios 7, 8, 9, and 10. Particularly for retry intention in Scenario 7 and for changes in motivation and confidence in Scenarios 8 and 10. Scenarios 5 and 10 showed no difference in retry intention, although they differed in their attitudinal patterns.

The condition-family interpretation is also informative. Coughing was the most favorable condition for the proposed interface on retry intention, because both Scenarios 1 and 6 showed a 7.1 percentage-point retry advantage for the proposed interface. The existing interface's overall descriptive advantage in the talking condition appears to be driven primarily by Scenario 7, where the existing interface exceeded the proposed interface by 35.7 percentage points in terms of retry intention.

Running was scenario-dependent: Scenario 3 favored the proposed interface, whereas Scenario 8 favored the existing interface. This is likely because the person in Scenario 8 seemed to be running in the woods; participants perceived this as a barrier to retrying due to the person's inability to receive help in such a remote area. Walking leaned toward the existing interface on the retry-intention question across both scenarios, although Scenario 4 showed a notable confidence-change advantage for the proposed interface; participants indicated that this was because the person seemed to be running late for work.

The changing light condition showed equal reported retry intention across interfaces, with both Scenarios 5 and 10 showing equal retry intention across interfaces. This was a complex condition, as the participants perceived that the measurement should have worked under the conditions described in the scenario. However, participants indicated that the multi-factor guidance made the system appear unable to identify one specific reason for the unsuccessful measurement; some evaluated each listed condition individually to rule out possible explanations.

5.3.2 Confidence and Ability

Confidence and motivation were measured as pre-post change outcomes. The proposed interface showed a small overall descriptive advantage for confidence change. Across scenarios, confidence increased by approximately 0.43 points with the proposed interface and 0.36 points with the existing interface. This suggests only a small overall difference, but some individual scenarios showed more meaningful descriptive patterns.

For example, Scenario 1 showed a clear descriptive confidence advantage for the proposed interface: confidence increased under the proposed interface but decreased under the existing interface. Scenario 4 also favored the proposed interface for confidence, with confidence increasing slightly under the proposed interface but decreasing under the existing interface. These findings suggest that the proposed facilitator prompts may have helped some participants feel more capable of correcting the failed measurement, even when the prompts did not consistently increase the intention to retry.

This difference between confidence and the intention to retry is important. A participant may feel more confident in understanding the cause of failure but still decide not to retry immediately. In a health context, this may be appropriate. For example, if the user believes they should rest first, adjust the watch, wait, or seek help, then not retrying immediately may represent a better decision rather than poorer usability. Therefore, future studies should distinguish between the intention to retry and the appropriateness of action.

Although the proposed interface showed a small descriptive advantage for confidence in some scenarios, the difference was not consistent across scenarios and was not statistically significant. Therefore, H3 for RQ3.3 was not supported. Nevertheless, the scenario-level patterns suggest that explanatory prompts may influence perceived ability under particular conditions, which should be examined in a larger study.

5.3.3 User Understanding of the System's Error State

The results suggest that users did not always understand why an ECG measurement failed. Before interface interaction, participants often attributed unsuccessful measurement to movement-related causes, such as moving, not holding still, or needing to rest. This indicates a possible motion bias in users' reasoning: when an ECG attempt fails, users may default to assuming they moved too much, that the sensor was not working properly, or that the watch was not positioned correctly, even when the actual factor is different.

This has important design implications. If users misidentify the cause of failure, they may take an ineffective next step. For example, if the actual issue is poor contact or environmental conditions, simply sitting still and retrying may not solve the problem. Repeated unsuccessful attempts could increase frustration, reduce confidence, or cause users to stop using the ECG function.

Therefore, a key design requirement is to help users develop an accurate mental model of ECG measurement failure. The interface should explain what likely went wrong, what the user should do next, and whether the situation requires retrying, waiting, adjusting the device, or seeking help.

5.3.4 Information Needs

The need for additional information was also a binary outcome; descriptively, participants reported needing additional information slightly more often under the proposed interface than under the existing one. Across the scenario-level results, the need for additional information was approximately 15.7% under the proposed interface and 12.1% under the existing interface, a difference of +3.6 percentage points. This corresponds to Cohen's h of approximately 0.11, a small descriptive effect indicating slightly greater information need under the proposed interface.

This finding may seem counterintuitive because the proposed interface was designed to provide more explanation. However, it may indicate that additional explanation can increase users' awareness of uncertainty. When users receive a more specific prompt, they may think more carefully about the cause of failure and may want more information before deciding what to do.

The implication is that information quality matters more than information quantity. Users may not need longer prompts; they may need clearer, more actionable ones. A useful failure notification should answer three questions quickly: what probably went wrong, what I should do now, and when I should be concerned.

Taken together, the interface-preference, understanding, clarity, and information-need findings provide mixed evidence regarding H2 for RQ3.2. Participants' preference for the proposed interface and their positive evaluations of its explanatory features provide descriptive support for H2. However, the lack of consistent statistically significant differences and the slightly greater need for additional information under the proposed interface mean that H2 was not fully supported statistically.

5.3.5 AI-Generated Videos for Knowledge Tasks in Human-Centered Studies

The AI-generated videos appeared to support participants' understanding of the scenarios. Participants rated the videos highly for their helpfulness in understanding the scenario, ease of placing themselves in the situation, and clarity of the scenario's communication. However, realism was rated more moderately. This suggests that the videos were useful as standardized study materials, but they may not have fully reproduced the experience of an actual unsuccessful ECG measurement.

The use of AI-generated videos is, therefore, both a strength and a limitation. As a strength, the videos allowed the study to present consistent scenarios to all participants. This helped ensure that participants responded to the same visual and contextual information. The high clarity and helpfulness ratings suggest that the videos successfully communicated the intended scenarios.

At the same time, the moderate realism rating limits ecological validity. A real failed ECG measurement may involve physical discomfort, concern about symptoms, frustration with the device, or anxiety about health status. These emotional and embodied factors may be especially important for older adults, cardiac patients, or people recovering from heart surgery. Therefore, although the videos were useful for scenario-based evaluation, future work should complement AI-generated scenarios with more realistic testing using actual devices and clinically relevant users.

5.4 ACT-AI Framework: Designing Human-AI Interaction

Building on the findings from the previous sections, we introduce the ACT-AI Framework. This framework is intended for designing AI-enabled systems that automate a substantial portion of decision and action selection, involving a human only when the algorithm deems it necessary for health care applications.

The ACT-AI Framework integrates findings from the classifier and explainability analyses in Sections 3.4 and 3.5, the LLM prompt evaluation in Section 3.8.1, and the human validation findings in Sections 4.5 and 5.3. It includes both components implemented in this thesis and proposed extensions motivated by the findings. In particular, the explainability pipeline and rule-based feature-actionability policy were implemented, whereas RAG, continuous outlier monitoring, and operational deployment safeguards are proposed extensions requiring future evaluation.

5.4.1 Core Design Principle

Use the Fogg Behavior Model in the design process. The system should increase the user's ability, or motivation by giving the right prompt at the right time. A good prompt should be actionable, honest about uncertainty, non-blaming, safe, fair, easy to follow, and based only on factors that are appropriate to communicate.

The human validation findings provide qualified support for this principle. Although 23 of the 28 participants preferred the proposed interface and particularly valued its explanation, corrective action, and next step, the prompts did not consistently increase reported retry intention (Sections 4.5.3 and 4.5.4).

5.4.2 Mandatory Automation Level 9 Requirement

Every Automation Level 9 system must provide access to the algorithmic decision-making process and the factors that influence it. This means the system must be able to expose, for user review, why a decision was made, which features influenced it, which features were excluded, what confidence level was assigned, and how the final user-facing prompt was selected.

Any Automation Level 9 must also detect when a significant number of outliers are entering the system. When outliers are detected, the system should prompt user action only when:

1. The likely reason is identifiable.
2. The reason is changeable or partially changeable.
3. The system has sufficient confidence in the cause.
4. The message can be communicated safely and fairly.

If these conditions are not met, the system should avoid making causal claims and instead use an inconclusive state or a technical support message.

Several findings support the transparency and auditability requirements proposed for Automations with Level 9. Before receiving interface guidance, participants frequently misidentified the cause of measurement failure, while their preference for the proposed interface was largely associated with its explanation of the failure and its clear next steps (Sections 4.5.2, 4.5.4, and 5.3.3). The classifier analysis also showed that models could rely on demographic, physiological, or non-actionable features, and the LLM evaluation showed that unsupported causes could be introduced during prompt generation (Sections 3.5 and 3.8.1). Together, these findings support requiring an auditable decision record and an understandable user-facing explanation, while continuous outlier monitoring remains a proposed safety requirement that was not directly tested.

5.4.3 Framework

Based on the findings of this thesis, the ACT-AI Framework is proposed to guide the design of AI-enabled systems that decide when to involve users in decision-making and action selection.

ACT-AI Framework

The design of an AI-automated system that decides when to include the user in the loop should:

1. Capture the Measurement Outcomes

Start by auditing the AI-enabled system and its output. Identify whether the intended use of the system was successful, failed, inconclusive, or anomalous. Capture supporting signals such as model confidence, uncertainty, signal quality, motion artifact, skin contact, noise, wrist moisture, device position, environmental conditions, and other model outputs.

2. Explain the Model Decision

Design an interface that portrays the system's decision for its users. Use explainability methods to understand why the system reached its decision. Use SHAP for feature attribution, LIME for local explanation, and DiCE for counterfactuals or likely improvements. This explanation layer should convey to the user: "What likely caused the outcome, what influenced the algorithm, and what could change next time?"

3. Incorporate a RAG System

Before generating a user-facing message, build a RAG system to classify relevant features. Changeable features can be directly acted upon by the user, the device, or the environment. Examples from this research include movement, ambient light, or electrical interference. Partially changeable features may be influenced but not fully controlled. For example, elevated heart rate after activity, heavy breathing, sweating, recent movement, or body settling after exercise are actions that take time to occur. Non-changeable or sensitive features should not be used in user-facing prompts. Examples include age, sex, race, ethnicity, disability, permanent medical history, or physiological traits.

4. Detect Outliers and Decide Whether to Intervene

Any automation level 9 should monitor incoming measurements for significant outlier patterns, including repeated failures, abnormal signal-quality distributions, unusual device-environment conditions, or clusters of inconclusive results. When significant outliers are detected, determine whether the model-associated factor is explainable, changeable, and sufficiently confident.

If it is, prompt the user with a clear action. If not, avoid overclaiming and route the case to an inconclusive state or help/support message.

5. Use an LLM for Reasoning and Triage

The LLM reasoning agent should follow a structured decision path:

1. Was a valid decision made?
2. Were significant outliers detected?
3. Is the likely cause identifiable?
4. Do SHAP, LIME, and DiCE agree?
5. Are the relevant features actionable, partially actionable, or non-actionable?
6. Is there enough confidence to prompt user action?
7. Is it fair and safe to communicate the cause?
8. Which prompt type should be used?
9. What final prompt should be generated?

6. The LLM must not invent causes.

When confidence is low, the system should use cautious language and avoid false precision.

7. Generate a Structured Output

Every AI-generated prompt should include Fogg mapping, title, body, call to action, and safety notes.

8. Deliver the Prompt to the User

The final prompt should be concise, supportive, and easy to act on. It should avoid blame and avoid exposing sensitive or non-actionable causes. The user should understand what happened, what they can do next, whether the system is uncertain, and how to retry or get help.

9. Design Considerations

1. Automation Level 9 must allow access to the algorithmic decision-making process and influencing factors.
2. Prompt user action for significant outliers only when the cause is changeable and confidence in the reason is sufficient.
3. Do not expose sensitive or non-actionable features in user-facing prompts.

4. Do not claim certainty when the system is uncertain.
5. Do not blame the user for failed measurements.
6. Do not over-explain technical model behavior.
7. Do give simple next steps when action is possible.
8. Do log reasoning and outcomes for auditability.
9. Do continuously evaluate prompt safety, fairness, and effectiveness.
10. Do not combine multiple explanations for the state of failure; users might lose trust in the system's ability to identify the cause of failure.

Taken together, the findings of this thesis provide the basis for the nine steps of the ACT-AI Framework. Step 1, capturing measurement outcomes, reflects the differences observed among classifiers in predictive performance, error patterns, feature reliance, and counterfactual coverage. Step 2, explaining the model decision, builds on the SHAP, LIME, and DiCE analyses and on participants' frequent misidentification of measurement-failure causes before receiving guidance. Step 3, extending the feature-actionability layer with RAG, is motivated by classifiers relying on some non-actionable features, LLMs occasionally introducing unsupported causes, and participants having scenario-specific information needs; however, RAG represents a proposed expansion of the implemented rule-based policy and was not evaluated in this thesis. Step 4, detecting outliers and deciding whether to intervene, reflects the absence of valid actionable counterfactuals for some cases and the mixed effects of prompts across scenarios. Steps 5 and 6, using an LLM for reasoning and applying a safety and evidence-validation gate, respond to the differences between GPT-5.5 and Llama 3.1:8B in evidence handling, clarity, structure, and reliability. Step 7, generating a structured output, is supported by the greater consistency of prompts containing the required explanation, corrective action, and call to action. Step 8, delivering the prompt to the user, reflects participants' preference for explanations and clear next steps despite the absence of a consistent increase in retry intention. Finally, Step 9 consolidates these findings into design considerations concerning transparency, actionability, uncertainty communication, safety, fairness, and auditability (Sections 3.4-3.8, 4.5.2-4.5.4, and 5.3.3-5.3.4).

5.5 Research Questions

5.5.1 How can explainable AI techniques be used to evaluate and compare artificial intelligence models for binary PPG signal-quality classification in terms of accuracy and interpretability?

This question is answered by the classifiers, model-selection, and explainability analyses in Chapter 3, Sections 3.3 to 3.6, and interpreted in Chapter 5, Section 5.1. Chapter 3 demonstrates how explainable AI techniques can evaluate binary PPG signal quality classification models beyond accuracy alone. While accuracy metrics indicate whether a model can distinguish acceptable from unacceptable signals, they do not reveal why a particular decision was made. Explainable AI adds interpretability by pinpointing which signal features or patterns influenced the model's classification. This enables comparison of models not only by performance but also by the logic of their decision processes and their alignment with domain knowledge about signal quality.

We found that the top-performing model was not always the most reliable, as it sometimes relied on non-modifiable factors like gender, glycemia, blood pressure, and heart rate. Explainable AI served two key roles. First, it aided model evaluation by helping determine if models relied on meaningful, interpretable indicators rather than unstable or uninterpretable patterns. Second, the counterfactual explanations bridged technical interpretation with user communication.

The explanations from the classification pipeline highlighted measurement limitations and influencing factors, which could be translated into user prompts. Thus, explainable AI was not only used to analyze model behavior but also to help develop communication strategies for a highly automated wearable digital health system.

5.5.2 How effectively can a large language model transform technical, explainable AI outputs into actionable, user-facing facilitator prompts?

This research question is addressed by the LLM generation, QUEST evaluation, and prompt-revision work presented in Chapter 3, Sections 3.7 and 3.8, and interpreted in Chapter 5, Section 5.2. In Chapter 3, this thesis shows that an instruction-guided large language model supplied with structured case-specific XAI outputs and a static, prompt-embedded actionability policy can transform technical explainable AI outputs into actionable facilitator prompts with moderate effectiveness, but not reliably enough for use without human validation and revision. A future implementation could replace the static policy used in this study with a retrieval-based actionability system.

In this chapter, Llama 3.1:8B and GPT-5.5 were compared to assess how well each model translated technical explanations of signal quality into short, user-facing guidance. GPT-5.5 produced clearer and more consistent prompt drafts than Llama, especially when translating conditions such as coughing, talking, running, walking, poor sensor contact, and changing-light environments into practical smartwatch guidance.

This suggests that a stronger instruction-following model can act as a useful translation layer between XAI outputs and end users by converting technical evidence, such as SHAP, LIME, DiCE, probabilities, or feature-level explanations, into simpler messages about what may have affected the recording and what the user could try next.

However, both models showed limitations when assessed against the QUEST health care principles. GPT-5.5 generally performed better in clarity, relevance, information quality, and actionability, but some outputs were still too generic, overly directive, or insufficiently precise. Llama was less reliable overall, with issues such as missing rendered prompts, technical phrasing, demanding calls to action, unsupported assumptions, and occasional blame-like wording. These weaknesses reduce performance in QUEST areas such as expression style, safety, harm avoidance, reasoning, and trust. Human validation and rephrasing were therefore necessary to make the final prompts safer, clearer, and more appropriate for smartwatch-based health care communication. The revised prompts improved tone and safety by using language such as “stop safely,” “wait until your breathing slows,” and “avoid speaking until the recording is complete,” rather than pressuring or blaming the user.

Overall, the findings indicate that an LLM can effectively support the drafting and translation of XAI-based facilitator prompts, but the most appropriate workflow remains a controlled pipeline in which the machine-learning model identifies bad-quality recordings, XAI methods provide explanatory evidence, the LLM generates a draft prompt, and human review ensures that the final output is grounded, non-coercive, non-blaming, and suitable for each context of use.

5.5.3 Do facilitator prompts increase the likelihood of users performing corrective actions to address system limitations as measured by participants' reported intention to record a single-lead ECG when presented after an unsuccessful measurement notification?

This research question is answered by the human-validation methods and results reported in Chapter 4, Sections 4.4 and 4.5, and discussed in Chapter 5, Sections 5.2 and 5.3. The results do not provide evidence that the facilitator prompts produced an overall increase in corrective-action intention. Corrective action was operationalized as the intention to retry or record a single-lead ECG after an unsuccessful measurement notification. Across the reported analyses, there was no statistically significant evidence that the proposed interface outperformed the existing interface using one-sided tests. The overall retry-intention pattern was mixed, and in several scenarios, the descriptive data favored the existing interface or showed no difference between interfaces.

However, the results also show that the facilitator prompts' effect was scenario-specific. Some scenarios, such as certain coughing and running scenarios, showed descriptive advantages for the proposed interface, while others, particularly some talking and walking scenarios, favored the existing interface. This means that the facilitator prompts cannot be described as uniformly increasing reported corrective-action intention. Instead, their effects depended on the type of measurement failure and on how participants interpreted the scenario conditions.

The results also suggest that retry intention alone may be an incomplete measure of success in a health-related context. Participants often valued the proposed interface even when it did not increase their intention to retry. This indicates that the prompts may have supported understanding, confidence, or more reflective decision-making, even if they did not always lead users to retry immediately.

In some situations, not retrying immediately may be appropriate if the user needs to rest, adjust the device, seek help, or obtain more information. Therefore, the findings of this thesis answer RQ3 by suggesting that facilitator prompts did not increase corrective-action intention overall, but they did influence participants' interpretations and preferences in ways that suggest potential value for scenario-specific, more context-sensitive prompt designs.

The findings can therefore be linked explicitly to the hypotheses associated with RQ3. As reported in Section 4.5.3, the proposed interface did not significantly increase retry intention; therefore, H1 for RQ3.1 was not supported. The preference and explanatory-feature results reported in Section 4.5.4 provide partial descriptive support for H2 for RQ3.2, although they do not provide confirmatory statistical support. The pre-to-post motivation and confidence findings reported in Section 4.5.3 and discussed in Section 5.3.2 did not demonstrate a consistent advantage for the proposed interface; therefore, H3 for RQ3.3 was not supported. Overall, the results answer RQ3 by showing that the facilitator prompts influenced participants' preferences and interpretations but did not consistently increase their intention to perform the corrective action.

5.5.4 How can high-level automations, such as an AI-enabled PPG-based smartwatch hypertension screening system with a limited user-facing interface, be designed to communicate system limitations and support corrective user action?

This question is synthesized in Chapter 5, Section 5.4, drawing on the classifiers and XAI findings reported in Chapter 3 and the human-validation findings from Chapter 4. This thesis shows that systems operate with Level 9 automation of decision and action selection, in which the system decides when to inform humans, and this decision should be designed to communicate its decision-making process, capabilities, and limitations. This is especially important for a smartwatch hypertension notification feature: the system may only display an interface when a predefined risk threshold is met. If the system fails to detect a substantial proportion of hypertension cases, users who receive no notification may interpret the absence of an alert as reassurance. This presents a design challenge: without an interface for users until the threshold is reached, users might not realize when the system is uncertain, when measurements are limited, or when the absence of notification should not be mistaken for an indication of no risk.

To address this problem, the thesis proposes a framework that uses the Fogg Behavior Model to structure the LLM instructions and facilitator-prompt format in order to generate counterfactual-based outputs. The proposed framework therefore includes a parallel support system that provides users with access to relevant aspects of the system's decision-making process. When substantial volumes of unusable data enter the system, the support system should communicate the factors behind the system's interpretation and suggest an appropriate corrective action, rather than merely stating that the measurements were unsuccessful.

This framework incorporates the Fogg Behavior Model to ensure prompts are both informative and actionable. Designed for situations in which motivation is high, but ability is limited, the prompt supports ability by making corrective actions straightforward, supports motivation by reducing uncertainty, and appears at the relevant decision point. Rather than just automating detection or issuing sporadic alerts, it facilitates calibrated human involvement by communicating relevant limitations, lowering false reassurance, and guiding users on whether to retry, adjust the device, wait, seek help, or seek care.

Overall, the answer to RQ4 is that high-level automated wearable health systems should communicate limitations through transparent, context-specific, and actionable interfaces. The interfaces of these AI-enabled systems should guarantee access to the relevant factors influencing the system's decision, translate technical explanations into user-facing language, and support appropriate corrective action without encouraging blind reliance or unnecessary repeated measurements, in a way consistent with regulations on the acceptable use of AI.

5.6 Contributions

This study contributes to applied research on Design for Human AI-Interaction, LLM-supported counterfactual explanations, consumer health interfaces, and AI-generated scenario materials. First, this study expands applied work on LLM-interpreted counterfactual explanations by extending counterfactual guidance to smartwatch ECG measurement failure. Arefeen et al. (2026) proposed MetaPlate, a counterfactual-guided RAG-LLM system that translates constrained meal modifications into realistic food recommendations for hyperglycemia prevention. Fredes and Vitria (2024) similarly proposed using LLMs to translate sets of counterfactual examples into natural-language explanations for final users. Narendra and Chatterjee (2026) emphasize that LLM-generated counterfactuals should be plausible, causally coherent, and actionable rather than open-ended, fluent "what-if" statements. The present study applies these ideas to a different consumer health context. Instead of using counterfactuals to recommend a meal, explain a tabular classifier, or generate algorithmic recourse, this study uses counterfactuals to provide possible contributing conditions to explain why the model produced a bad-quality output and what the user could do differently. The findings are consistent with the emphasis in these prior studies on actionability and translation: participants preferred the proposed interface because it suggested an explanation of the likely cause of failure based on the model decision boundary, the corrective action, and the next steps. However, the study also adds a limitation: user-facing model-informed counterfactual explanations did not produce uniform effects on reported retry intention across scenarios. This suggests that LLM-interpreted explanations in health interfaces must be evaluated not only for plausibility but also for how users interpret and act on them in context.

Second, this study contributes a design framework for generating user-facing counterfactual prompts for highly automated smartwatch ECG systems. The framework is intended for systems aligned with Level 9 automation, where the system decides when to involve the human rather than continuously asking for human input. In this type of interaction, the system may detect or infer that a measurement has failed, identify likely influencing factors, and decide when to inform the user. However, because this level of automation can reduce the user's visibility into the system's reasoning, the framework emphasizes that the moment of human involvement should provide access to the relevant parts of the algorithmic decision-making process. In this study, this means communicating the likely cause of the bad-quality classification and the factors influencing the notification, such as movement, talking, coughing, recent physical activity, watch position, or environmental conditions.

Third, this framework embeds the Fogg Behavior Model into the causal explanation design process. The Fogg Behavior Model states that behavior occurs when motivation, ability, and a prompt converge at the same moment (Fogg, 2009). In the proposed framework, counterfactual prompt generation begins with a likely cause of measurement failure and constrains the causal explanation to an actionable change the user could realistically make, such as sitting still, stopping talking, resting before retrying, adjusting the watch, or changing the measurement environment. The prompt is then structured to support ability by making the corrective action simple, support motivation by reducing unnecessary uncertainty, and serve as a timely cue at the moment of measurement failure. This is a design contribution rather than a claim that the framework has been fully validated. The results suggest that participants valued prompts that connected cause, action, and next steps, but the scenario-level variation shows that the framework needs further refinement and testing.

Fourth, this study broadens research on model-informed counterfactual explanations, trust, reliance, and user understanding by testing counterfactual prompts in an ambiguous consumer health-device setting. M. H. Lee and Chew (2023) investigated how counterfactual explanations influence human-AI collaborative clinical decisions and found that they can reduce overdependence on incorrect AI results. Similarly, this study found that in changing-light conditions and throughout the experiment, users believed that an inaccurate counterfactual prompt could undermine trust in the system's ability to identify the cause of a measurement failure, leading to under-reliance.

VanNostrand et al. (2024) studied how different types and presentations of counterfactual explanations affect lay users in automated decision scenarios, such as loan denials, to support actionable responses. This research extends these efforts to unsuccessful smartwatch ECG measurements, which are not formal denials but temporary, ambiguous failures that users may interpret in relation to device use, bodily condition, or health concerns. Results indicate that causal explanations should promote calibrated reliance rather than automatic retries. Participants favored the explanatory interface, but their intentions to retry varied across scenarios. Consequently, success should be judged not only by whether users retry but also by whether the explanation helps them decide to retry, wait, adjust the watch, seek help, or consult a medical professional.

Fifth, this study expands work on AI-mediated contextualization of wearable health information. Mahajan and Gilbert (2025) argue that digital health technologies can contribute to information overload, anxiety, fatigue, and misinterpretation when biometric data is insufficiently contextualized. They suggest that AI-enabled health assistants may help by filtering, contextualizing, and personalizing health information. The present study provides evidence from a specific smartwatch ECG failure interaction. Participants valued the proposed interface because it provided contextualized guidance after an unsuccessful measurement. At the same time, the results showed that more explanation did not automatically improve retry intention, and in some scenarios, participants still reported needing additional information. This suggests that AI-mediated contextualization may be useful, but only if the amount, timing, and specificity of the information are carefully designed. In wearable ECG interfaces, the goal is not simply to provide more information, but to provide information that helps users interpret the failure without increasing uncertainty or cognitive burden.

Sixth, this study contributes methodologically by expanding the body of work on AI-generated video scenarios in health research. Chu and Goodell (2024) proposed synthetic patients as a way to use multimodal generative AI to create video-based medical simulations. The present study applies a related approach to evaluating consumer health interfaces by using AI-generated videos to present standardized unsuccessful ECG measurement scenarios. The videos appeared to support participants' understanding of the scenarios: participants rated them highly for helpfulness in understanding the scenario, ease of placing themselves in the situation, and clarity of scenario communication. The videos also provided participants with contextual details they used to decide what the person should do, such as the person's age, possible pre-existing conditions, the surrounding environment, the watch's position, and whether the person appeared able to seek help. At the same time, the videos were rated only moderately realistic. This suggests that AI-generated videos were useful as standardized study materials, but may not fully reproduce the emotional, physical, or clinical experience of an actual unsuccessful ECG measurement. Therefore, future studies should complement AI-generated scenarios with more realistic testing using actual devices and clinically relevant users.

These contributions collectively position the study as an applied extension of recent research on counterfactual health explanations, LLM-supported explanations, human-automation interaction, and AI-mediated health communication. It demonstrates how constrained counterfactuals can be transformed into user-facing guidance for addressing smartwatch ECG failures. The study introduces a framework where the system determines when to involve the human user, providing information that helps them understand the system's decision-making process and the factors influencing the notification. It also emphasizes that counterfactual explanations in health should be evaluated not only for their clarity and confidence but also for the user's calibrated reliance and decision-making thresholds. Furthermore, it underscores that AI-mediated health guidance must be carefully contextualized.

5.7 Limitations

5.7.1 Artificial Intelligence Pipeline

First, the dataset is strongly imbalanced toward bad-quality samples. This makes accuracy and bad-quality F1-score potentially misleading. Balanced accuracy, AUROC, and confusion matrices provide a more reliable interpretation.

Second, many motion-only models predicted all samples as bad quality. Although this gives high bad-quality recall and F1-score, it is not useful for a real system because it cannot identify good-quality measurements.

Third, many full tabular models were flagged as potentially overfit. This suggests that some models may have learned dataset-specific or participant-specific patterns. The group-based split helps reduce this risk, but further validation on external participants would strengthen the findings.

Fourth, the full tabular explanations often relied on non-actionable features. This is a limitation for user-facing feedback. A model can be accurate but still produce explanations that are difficult to convert into corrective instructions.

Fifth, the DiCE counterfactuals were constrained to changes in movement conditions. This was appropriate for testing actionable prompts, but it means that counterfactual success was intentionally limited. Some tabular classifications may have required changes to physiological or measurement-site variables, but those changes were not allowed because they are less suitable for immediate correction by users.

5.7.2 Generating the User-facing Prompts

The approach used to design the prototype interfaces has several limitations. First, the use of LLMs should be approached with caution because the outputs of these models are not fully reproducible and may change over time, even with the same input, due to stochastic generation and model updates. This limits the long-term reliability and verifiability of the outputs, as well as the real-world implementation of the approach.

Second, the Fogg Behavior Model assumes that behavior occurs when motivation, ability, and a prompt converge; however, in this project, motivation was treated as present because the user had already attempted a measurement. This assumption may not hold in all cases. A user may begin a measurement without strong motivation to complete it, become frustrated after repeated failures, distrust the system, or not understand why a retry is necessary. Therefore, prompts focused mainly on increasing ability may be insufficient when motivational, emotional, or contextual barriers also influence whether the user retries the measurement.

Third, the facilitator prompt design assumes that unsuccessful or inconclusive measurements can often be improved through small corrective actions. Although this is appropriate for the signal-quality problems studied in this thesis, such as movement or environmental interference, not all failures are easily actionable by the user. Some measurement failures may arise from physiological states, device limitations, sensor noise, model uncertainty, or combinations of features that are difficult to translate into a clear behavior instruction. In those cases, the LLM should generate a different type of prompt when the underlying cause is only partially identifiable.

Fourth, the quality of the generated prompts depends heavily on the reliability of the upstream machine-learning pipeline. The pipeline does not establish causal effects, and changing a DiCE-identified factor is not guaranteed to improve the actual signal. The LLM does not directly observe the physiological signal; it relies on the decision state, signal-quality indicators, SHAP and LIME outputs, and DiCE model counterfactuals provided to it. If the original model output is incorrect, the signal-quality metrics are incomplete, or the explanation methods identify misleading features, the generated prompt may also be misleading. Thus, the LLM should be understood as a communication and decision-support layer rather than an independent validator of measurement failure.

The feature-actionability layer was introduced to reduce this risk by classifying features as changeable, partially changeable, non-changeable, sensitive, or physiological. However, this layer was simulated rather than implemented as a fully validated retrieval-augmented generation system. Its effectiveness, therefore, depends on the completeness and correctness of the policy. If the policy misclassifies a feature, omits relevant domain constraints, or fails to account for context-specific feasibility, the LLM may generate prompts that are too generic, not actionable, or insufficiently cautious. In addition, some features may be more or less actionable depending on the user's context.

For example, resting, changing lighting conditions, or repositioning the device may be simple for some users but difficult or impossible for others.

The approach is also limited by the number and representativeness of the runtime cases. The prompted LLM was evaluated using ten preselected, structured case payloads identified from previous sections, which may not capture the full range of real-world measurement failures. Smartwatch-based physiological measurements can be affected by a range of factors, including movement, posture, skin contact, device placement, ambient conditions, user physiology, and device-specific behavior. As a result, outputs that perform well on the selected cases may not generalize to unseen users, devices, environments, or failure modes.

Another important limitation is that the comparison of the two LLMs' outputs was conducted by a single reviewer, which might introduce bias into the scores.

5.7.3 Human Validation of the Prototype

Several limitations should be considered when interpreting the findings of the human validation study presented in Chapter 4. Although the sample size was sufficient for the planned analysis, the sample characteristics may limit the generalizability of the results. The sample was not representative of the full population of smartwatch ECG users. Most participants were young adults, highly educated, and had moderate to high eHealth literacy. No participants with low digital health literacy were included in the analysis. Participants likely had relatively high experience with general technology, but many had limited experience with ECG measurement or cardiac health problems.

This sampling choice was intentional because the study aimed to assess the effect of the counterfactual-based prompts displayed to users in a controlled environment. By recruiting participants who were generally familiar with technology and had moderate to high digital health literacy, the study reduced potential confounding effects related to basic interface use, technology anxiety, or difficulty understanding digital interactions. This allowed the evaluation to focus more directly on whether the content of the failure notification itself, the explanation of the cause, the recommended action, and the next steps influenced users' interpretation of the unsuccessful ECG measurement and their intended response.

However, this design choice also limits the generalizability of the findings. A group of older adults, users with lower digital health literacy, or patients who had recently undergone heart surgery or were living with other cardiac conditions might have produced different results. These users may interpret unsuccessful ECG measurements differently from the relatively young, educated, and digitally literate participants included in the present study. For example, users with recent cardiac events or ongoing heart conditions may have different decision-making thresholds after an unsuccessful measurement. They may be more likely to contact a health care professional or emergency services rather than retrying immediately, especially if they are experiencing symptoms or have a recent history of cardiac issues. Conversely, they may also be more motivated to obtain a successful ECG recording because the measurement may feel more personally relevant or clinically important.

Therefore, while the current sample strengthened internal validity by helping control confounding factors, future studies should test the proposed interface with more clinically and demographically diverse users. This would make it possible to assess whether the same notification design is effective for users with varying levels of expertise, digital health literacy, health-related anxiety, and expectations about when to retry, when to wait, or when to seek medical support.

A second limitation is that the study used a Figma prototype instead of an actual smartwatch. This was suitable for assessing the interface's effectiveness in a controlled environment and helped avoid distractions present with real devices. However, the prototype did not fully mimic real interactions, such as touch gestures, haptic feedback, notification timing, and daily activities, which could lead to similar errors. Participants described their actions after an unsuccessful ECG measurement but did not perform real ECGs in real-life situations. Behavior may differ when users face symptoms, stress, time pressure, or repeated device failures. Consequently, participants' responses might not accurately reflect how they would interact with and react to the same interface on an actual smartwatch.

Third, a potential carryover effect may have influenced participants' responses across scenarios. Although participants were reminded that the individuals shown in each scenario were different people and did not know each other and therefore could not have any knowledge of what occurred in previous scenarios, some participants continued to apply information or assumptions from earlier scenarios to later ones. This may have affected how they interpreted the cause of unsuccessful ECG measurements and how they responded to subsequent guidance.

Fourth, the study examined a limited set of causes of failure: coughing, talking, running, walking, and changes in light. Real-world smartwatch ECG failures may involve additional factors such as poor skin contact, sweating, incorrect finger placement, anxiety, symptoms, device malfunction, or environmental interruptions, some of which were discussed by the participants. Future research should evaluate facilitator prompts across a broader range of failure contexts.

Finally, although the AI videos showed promising results and appeared to support participant engagement and relatedness to the scenarios, some participants looked for very specific visual cues or details in the videos when answering questions. This suggests that video characteristics may have influenced participants' interpretation of the scenarios. Therefore, a separate study should be conducted to specifically evaluate the use of AI-generated video scenarios and determine how video realism, visual detail, and cue salience affect participant understanding, decision-making, and responses.

5.8 Future Work

As artificial intelligence brings a new level of automation to everyday applications, it will be vital to clearly convey information to users, particularly what AI-enabled systems do and what they do not. This study establishes a robust foundation for understanding how counterfactuals can be used to convey the output of artificial intelligence systems in natural language, thereby enhancing the user's understanding of the system's limitations. Additional work is needed to more comprehensively evaluate its cognitive, technical, contextual, and clinical implications.

First, future research should investigate the cognitive mapping of user decision-making following the corrective feedback suggested in the proposed framework. In particular, the decision ladder framework could be used to characterize how users perceive signal-quality problems, interpret corrective instructions, select actions, and evaluate those actions to improve measurement. This would also contribute to understanding uncertainty, cognitive overload, or misinterpretation of outputs when responding to the prompts.

Second, future studies should examine the cross-context generalization of PPG signal quality classification across different recording devices. The device used can affect wave morphology, noise characteristics, and signal quality.

Comparative studies should therefore evaluate whether corrective actions effective for one device remain effective across devices and whether the LLM component can adapt guidance to device-specific signal characteristics. This line of work should also assess whether different devices prompt different levels of user engagement, trust, and responsiveness when corrective feedback is provided.

Third, longitudinal real-world assessment through on-device deployment is needed to evaluate the practical robustness of the proposed model. The proposed framework introduces variability related to the recording device, the signal processing and classification algorithms, the user's physical environment, and the behavioral context in which measurements are taken. Future studies should therefore deploy the framework design guidance in real-world settings over extended periods. Such studies should evaluate how the system performs under noisy, low-quality, and uncontrolled signal conditions, as well as how users adapt to repeated corrective feedback over time. This would provide stronger evidence regarding the ecological validity, reliability, and scalability of the proposed approach.

Fourth, future research should systematically compare the impact of AI-generated videos and written scenarios on knowledge-based tasks. In this study, AI-generated videos were introduced to control and enhance the motivational context of user scenarios. The findings suggest that video-based scenarios may introduce a new level of attention, engagement, and immersion in knowledge tasks. However, additional research is needed to systematically isolate and evaluate this effect. Future studies could compare written-only scenarios, AI-generated video scenarios, and mixed-media interventions to determine how each format influences comprehension, motivation, recall, selection of corrective actions, and perceived relevance.

Fifth, future work should evaluate the effects of corrective feedback interventions on doctor-patient interaction. Smartwatches and other wearable devices already present users with large amounts of physiological data, and platforms such as the Apple Watch are increasingly being designed to support health-related conversations between patients and clinicians. In this context, it is important to examine whether helping users understand how to improve signal quality changes the nature of doctor-patient communication. This is especially relevant in hypertension care, where accurate home monitoring, patient adherence, and interpretation of wearable data may influence clinical decision-making. Future studies should investigate whether corrective feedback improves patient confidence, clinician trust in patient-generated data, and the quality of shared decision-making during consultations.

Finally, the proposed framework should be expanded to incorporate greater technical flexibility. As AI systems are increasingly offering AI as a service, users are often given the option to select among different models, tools, or configurations. Future research should explore whether counterfactual explanations can help users choose a more appropriate model based on their intended use, technical constraints, and performance limitations. This extension would move the framework beyond corrective feedback for signal quality and toward a broader framework for supporting informed human-AI decision-making.

5.9 Conclusion

This thesis explored a key challenge in highly automated AI-powered wearable health systems: when the system decides whether to notify the user, the lack of a notification can serve as meaningful feedback. For example, Apple's hypertension alert only appears when a certain risk level is met. However, if the system does not detect many users with hypertension, some with undiagnosed conditions might receive no alert at all. This may lead to false reassurance, as users might think they are safe because they have not been notified, even if the system simply missed their condition.

This issue points to a broader concern with smartwatch health models: they may perform well in training and validation but can significantly underperform in real-world use, leading to missed notifications, unclear feedback, or misplaced reassurance. Because the decision to involve the user is made by the system, this is a Level 9 automation issue. Users depend on the system's communication but often have limited insight into its uncertainty or limitations. To address this, the thesis investigates how highly automated wearable health systems can be designed to better communicate their constraints when involving users.

The proposed approach integrates explainable AI, LLM-based interpretation, and a Fogg-based counterfactual output into a design framework to create user-facing outputs that make the system's reasoning more transparent. These prompts aim to explain potential measurement limitations and relevant influencing factors and to suggest next steps or corrective actions. The results of the human-validation study indicated that participants valued prompts linking failure causes to clear instructions, although these did not consistently increase retry intentions across scenarios.

Ultimately, the thesis emphasizes that Level 9 health systems should promote calibrated human involvement by making limitations visible, minimizing false reassurance, and guiding users on whether to retry, wait, adjust, or seek care. The main contribution is the development and assessment of a Fogg-informed, LLM-interpreted counterfactual prompt framework for Level 9 smartwatch ECG automation. Results show that users appreciate explanations that connect the causes of failure, influencing factors, corrective measures, and next steps.

However, these explanations do not uniformly change reported retry intention across all scenarios. While participants favored the proposed interface, their intentions to retry, confidence, motivation, and information needs varied across scenarios. This indicates that counterfactual health explanations should be evaluated not only as technically plausible messages but also as user-centered health communications that influence understanding, confidence, perceived information sufficiency, reliance, and decision-making regarding retrying, waiting, adjusting, or seeking help.

References

Abdelaal, Y., Aupetit, M., Baggag, A., & Al-Thani, D. (2024). Exploring the Applications of Explainability in Wearable Data Analytics: Systematic Literature Review. *Journal of Medical Internet Research*, *26*. Scopus. https://doi.org/10.2196/53863

Ahluwalia, S., Murray, E., Stevenson, F., Kerr, C., & Burns, J. (2010). 'A heartbeat moment': Qualitative study of GP views of patients bringing health information from the internet to a consultation. *The British Journal of General Practice*, *60*(571), 88–94. https://doi.org/10.3399/bjgp10X483120

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ali, D. H., Kiliç, B., Hart, H. E., Bots, M. L., Biermans, M. C. J., Spiering, W., Rutten, F. H., & Hollander, M. (2021). Therapeutic inertia in the management of hypertension in primary care. *Journal of Hypertension*, *39*(6), 1238–1245. https://doi.org/10.1097/HJH.0000000000002783

Almarshad, M. A., Islam, M. S., Al-Ahmadi, S., & BaHammam, A. S. (2022). Diagnostic Features and Potential Applications of PPG Signal in Healthcare: A Systematic Review. *Healthcare*, *10*(3), 547. https://doi.org/10.3390/healthcare10030547

Alnooh, G., Alessa, T., Hawley, M., & De Witte, L. (2022). The Use of Dietary Approaches to Stop Hypertension (DASH) Mobile Apps for Supporting a Healthy Diet and Controlling Hypertension in Adults: Systematic Review. *JMIR Cardio*, *6*(2), e35876. https://doi.org/10.2196/35876

Alyafei, A., & Easton-Carr, R. (2024). The Health Belief Model of Behavior Change. In *StatPearls [Internet]*. StatPearls Publishing. (39163427). https://www.ncbi.nlm.nih.gov/books/NBK606120/

American Medical Association. (2026). *Validated devices for blood pressure measurement*. American Medical Association. https://www.validatebp.org/index.php/devices

Apple Inc. (2025). *Hypertension Notifications Validation*. https://www.apple.com/health/pdf/Hypertension_Notifications_Validation_Paper_September_2025.pdf

Arefeen, A., Johnston, C., & Ghasemzadeh, H. (2026). *MetaPlate: Counterfactual-Guided RAG-LLM Tool for Personalized Food Recommendation and Hyperglycemia Prevention* (arXiv:2606.10120). arXiv. https://doi.org/10.48550/arXiv.2606.10120

Argüello-Prada, E. J., & Castillo García, J. F. (2024). Machine Learning Applied to Reference Signal-Less Detection of Motion Artifacts in Photoplethysmographic Signals: A Review. *Sensors*, *24*(22), 7193. https://doi.org/10.3390/s24227193

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, *22*(6), e15154. https://doi.org/10.2196/15154

Averbuch, T., Stelkia, K., & Van Spall, H. G. C. (2024). Cardiovascular Care in Canada: Strengths, Challenges, and Opportunities For Improvement. *Circulation*, *150*(19), 1490–1492. https://doi.org/10.1161/CIRCULATIONAHA.124.070628

Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., Martin, S. S., Muse, E. D., Turakhia, M. P., Tarakji, K. G., & Elshazly, M. B. (2021). Smart wearable devices in cardiovascular care: Where we are and how to move forward. *Nature Reviews. Cardiology*, *18*(8), 581–599. https://doi.org/10.1038/s41569-021-00522-7

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *Npj Digital Medicine*, *3*(1), 18. https://doi.org/10.1038/s41746-020-0226-6

Bergmann, D. (2025, August 28). *What is machine learning?* IBM. IBM Think. https://www.ibm.com/think/topics/machine-learning

Buyting, R., Melville, S., Chatur, H., White, C. W., Légaré, J.-F., Lutchmedial, S., & Brunt, K. R. (2022). Virtual Care With Digital Technologies for Rural Canadians Living With Cardiovascular Disease. *CJC Open*, *4*(2), 133–147. https://doi.org/10.1016/j.cjco.2021.09.027

Cau, F. M., & Spano, L. D. (2026). Exploring the impact of explainable AI and cognitive capabilities on users' decisions. *User Modeling and User-Adapted Interaction*, *36*(1), 3. https://doi.org/10.1007/s11257-025-09438-0

Celar, L., & Byrne, R. M. J. (2023). How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, *51*(7), 1481–1496. https://doi.org/10.3758/s13421-023-01407-5

Center for Devices and Radiological Health. (2024). What is Digital Health? *FDA*. https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health

Chen, Q. (2025). Artificial intelligence in healthcare: Rethinking doctor-patient relationship in megacities. *Frontiers in Health Services*, *5*. https://doi.org/10.3389/frhs.2025.1694139

Chu, S. N., & Goodell, A. J. (2024). *Synthetic Patients: Simulating Difficult Conversations with Multimodal Generative AI for Medical Education* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.19941

Cohen, J. B., Addo, D. K., Jacobs, J. A., Greene, T. H., Xu, Y., Picone, D. S., Shimbo, D., & Bress, A. P. (2026). Impact of a Smartwatch Hypertension Notification Feature for Population Screening. *JAMA*, *335*(11), 1001. https://doi.org/10.1001/jama.2025.26925

Dankwa-Mullan, I. (2024). Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. *Preventing Chronic Disease*, *21*, 240245. https://doi.org/10.5888/pcd21.240245

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555

Fogg, B. J. (2009). A behavior model for persuasive design. *Proceedings of the 4th International Conference on Persuasive Technology*, 1–7. https://drive.google.com/file/d/19LYba4fuceGM3KhqxTXByV4msmb33t1o/view

Franklin, S. S., Thijs, L., Hansen, T. W., O'Brien, E., & Staessen, J. A. (2013). White-Coat Hypertension: New Insights From Recent Studies. *Hypertension*, *62*(6), 982–987. https://doi.org/10.1161/HYPERTENSIONAHA.113.01275

Fredes, A., & Vitria, J. (2024). *Using LLMs for Explaining Sets of Counterfactual Examples to Final Users* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2408.15133

Gavish, B. (2010). The relationship between systolic and diastolic pressures: A possible link between risk-related clinical measures and arterial properties. *Hypertension Research*, *33*(7), 657–658. https://doi.org/10.1038/hr.2010.97

George, J., & MacDonald, T. (2015). Home Blood Pressure Monitoring. *European Cardiology*, *10*(2), 95–101. https://doi.org/10.15420/ecr.2015.10.2.95

Goh, C.-H., Tan, L. K., Lovell, N. H., Ng, S.-C., Tan, M. P., & Lim, E. (2020). Robust PPG motion artifact detection using a 1-D convolution neural network. *Computer Methods and Programs in Biomedicine*, *196*, 105596. https://doi.org/10.1016/j.cmpb.2020.105596

Golbus, J. R., Dorsch, M. P., Chen, Y., Basu, T., Luff, E., Klasnja, P., Newman, M. W., Skolarus, L. E., Dempsey, W., & Nallamothu, B. K. (2026). Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention. *Journal of Medical Internet Research*, *28*, e78218–e78218. https://doi.org/10.2196/78218

Goupil, R., Tsuyuki, R. T., Santesso, N., Terenzi, K. A., Habert, J., Cheng, G., BScPharm, S. C. G., Np, J. B., Leung, A. A., Campbell, N. R. C., Schiffrin, E. L., & Hundemer, G. L. (2025). *Hypertension Canada guideline for the diagnosis and treatment of hypertension in adults in primary care*. https://doi.org/10.1503/cmaj.241770

Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, *38*(5), 2770–2824. https://doi.org/10.1007/s10618-022-00831-6

Hassan, M., Kushniruk, A., & Borycki, E. (2024). Barriers to and Facilitators of Artificial Intelligence Adoption in Health Care: Scoping Review. *JMIR Human Factors*, *11*, e48633. https://doi.org/10.2196/48633

Health Canada. (2025). *Caring for Canadians: Canada's future health workforce : the Canadian health workforce education, training and distribution study*. Health Canada.

Huang, J. L., & Wang, Z. (2021). Careless Responding and Insufficient Effort Responding. In D. D. Bergh (Ed.), *Oxford Research Encyclopedia of Business and Management* (1st ed.). Oxford University PressNew York, NY. https://doi.org/10.1093/acrefore/9780190224851.013.303

Huang, Q.-F., Yang, W.-Y., Asayama, K., Zhang, Z.-Y., Thijs, L., Li, Y., O'Brien, E., & Staessen, J. A. (2021). Ambulatory Blood Pressure Monitoring to Diagnose and Manage Hypertension. *Hypertension*, *77*(2), 254–264. https://doi.org/10.1161/HYPERTENSIONAHA.120.14591

Hughes, A. M., Taylor, D. J., Morris, P. D., & Brittain, E. L. (2026). Wearable devices and cardiovascular health: Revolutionizing remote monitoring and disease prevention. *European Heart Journal*, *47*(18), 2130–2145. https://doi.org/10.1093/eurheartj/ehag189

Hughes, A., Shandhi, M. M. H., Master, H., Dunn, J., & Brittain, E. (2023). Wearable Devices in Cardiovascular Medicine. *Circulation Research*, *132*(5), 652–670. https://doi.org/10.1161/CIRCRESAHA.122.322389

Jones, D. W., Ferdinand, K. C., Taler, S. J., Johnson, H. M., Shimbo, D., Abdalla, M., Altieri, M. M., Bansal, N., Bello, N. A., Bress, A. P., Carter, J., Cohen, J. B., Collins, K. J., Commodore-Mensah, Y., Davis, L. L., Egan, B., Khan, S. S., Lloyd-Jones, D. M., Melnyk, B. M., … Williamson, J. D. (2025). 2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM Guideline for the Prevention, Detection, Evaluation and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Hypertension*, *82*(10). https://doi.org/10.1161/HYP.0000000000000249

Joshee, S. R., Dingle, K., Jones, J. E., Mahtani, A. U., Patel, D. A., Juraschek, S. P., Plante, T. B., Brady, T. M., & Hu, J.-R. (2026). Current Landscape of Mobile Health Applications for

Hypertension Management in the United States: A Scoping Application Review. *American Journal of Hypertension*, *39*(5), 632–644. https://doi.org/10.1093/ajh/hpaf243

Kario, K., Nomura, A., Harada, N., Okura, A., Nakagawa, K., Tanigawa, T., & Hida, E. (2021). Efficacy of a digital therapeutics system in the management of essential hypertension: The HERB-DH1 pivotal trial. *European Heart Journal*, *42*(40), 4111–4122. https://doi.org/10.1093/eurheartj/ehab559

Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics*, *13*(15), 2582. https://doi.org/10.3390/diagnostics13152582

Kwon, J. H., Kim, S. E., Kim, N. H., Lee, E. C., & Lee, J. H. (2022). Preeminently Robust Neural PPG Denoiser. *Sensors*, *22*(6), 2082. https://doi.org/10.3390/s22062082

Landi, F., Calvani, R., Picca, A., Tosato, M., Martone, A. M., Ortolani, E., Sisto, A., D'Angelo, E., Serafini, E., Desideri, G., Fuga, M. T., & Marzetti, E. (2018). Body Mass Index is Strongly Associated with Hypertension: Results from the Longevity Check-Up 7+ Study. *Nutrients*, *10*(12), 1976. https://doi.org/10.3390/nu10121976

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). *Explaining Model Confidence Using Counterfactuals* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2303.05729

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M., Gao, Y., Wu, J., McIntosh, C., & Franklin, D. (2025). Development of deep learning models for motion artifact mitigation in wearable PPG devices. In B. Shadgan & A. H. Gandjbakhche (Eds.), *Biophotonics in Exercise Science, Sports Medicine, Health Monitoring Technologies, and Wearables VI* (p. 23). SPIE. https://doi.org/10.1117/12.3039622

Lee, M. H., & Chew, C. J. (2023). Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2). Scopus. https://doi.org/10.1145/3610218

Lu, Q., & Schulz, P. J. (2024). Physician Perspectives on Internet-Informed Patients: Systematic Review. *Journal of Medical Internet Research*, *26*(1), e47620. https://doi.org/10.2196/47620

Mahajan, A., & Gilbert, S. (2025). Do we need AI guardians to protect us from health information overload? *Npj Digital Medicine*, *8*(1), 632. https://doi.org/10.1038/s41746-025-02093-0

Mak, S. S., Kinney, R. L., Bailey, A. L., Gaffey, A. E., Relyea, M. R., Spanakis, E. K., & Ash, G. I. (2025). Integrating Consumer-Grade Wearable Devices and Patient-Generated Health Data into Clinical Care: Perspectives from Healthcare Professionals at a Learning Health System. *Journal of General Internal Medicine*. https://doi.org/10.1007/s11606-025-09876-x

Mangion, A., & Piller, N. (2025). What is digital health, eHealth, uHealth and Healthcare 4.0? *Informatics and Health*, *2*(2), 137–142. https://doi.org/10.1016/j.infoh.2025.06.002

McCarthy, J. (2007). *What is artificial intelligence?* Stanford University. https://www-formal.stanford.edu/jmc/whatisai.pdf

McEvoy, J. W., McCarthy, C. P., Bruno, R. M., Brouwers, S., Canavan, M. D., Ceconi, C., Christodorescu, R. M., Daskalopoulou, S. S., Ferro, C. J., Gerdts, E., Hanssen, H., Harris, J., Lauder, L., McManus, R. J., Molloy, G. J., Rahimi, K., Regitz-Zagrosek, V., Rossi, G. P.,

Sandset, E. C., … Khamidullaeva, G. A. (2024). 2024 ESC Guidelines for the management of elevated blood pressure and hypertension. *European Heart Journal*, *45*(38), 3912–4018. https://doi.org/10.1093/eurheartj/ehae178

McMullan, M. (2006). Patients using the Internet to obtain health information: How this affects the patient–health professional relationship. *Patient Education and Counseling*, *63*(1–2), 24–28. https://doi.org/10.1016/j.pec.2005.10.006

Mercer, K., Li, M., Giangregorio, L., Burns, C., & Grindrod, K. (2016). Behavior Change Techniques Present in Wearable Activity Trackers: A Critical Analysis. *JMIR mHealth and uHealth*, *4*(2), e40. https://doi.org/10.2196/mhealth.4461

Mesko, B., deBronkart, D., Dhunnoo, P., Arvai, N., Katonai, G., & Riggare, S. (2025). The Evolution of Patient Empowerment and Its Impact on Health Care's Future. *Journal of Medical Internet Research*, *27*, e60562. https://doi.org/10.2196/60562

Mialhe, F. L., Moraes, K. L., Sampaio, H. A. D. C., Brasil, V. V., & Rebustini, F. (2023). Normatização dos escores da escala eHealth Literacy Scale para avaliação do letramento digital em saúde. *Revista Enfermagem UERJ*, *31*, e74812. https://doi.org/10.12957/reuerj.2023.74812

Michie, S., Van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 42. https://doi.org/10.1186/1748-5908-6-42

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, 607–617. https://doi.org/10.1145/3351095.3372850

Narendra, Y., & Chatterjee, N. (2026). Causal What-Ifs: Rethinking Counterfactuals with LLM Agents. In T. Taniguchi, C. S. A. Leung, T. Kozuno, J. Yoshimoto, M. Mahmud, M. Doborjeh, & K. Doya (Eds.), *Neural Information Processing* (Vol. 16309, pp. 184–198). Springer Nature Singapore. https://doi.org/10.1007/978-981-95-4367-0_13

Nemcova, A., Vargova, E., Marsanova, L., Vitek, M., Smital, L., Filipenska, M., Sikorova, P., & Galik, P. (2024). *Brno University of Technology Smartphone PPG Database (BUT PPG) (version 2.0.0)*. PhysioNet. https://doi.org/10.13026/tn53-8153

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems*, *24*. https://doi.org/10.17705/1CAIS.02428

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Parati, G., Goncalves, A., Soergel, D., Bruno, R. M., Caiani, E. G., Gerdts, E., Mahfoud, F., Mantovani, L., McManus, R. J., Santalucia, P., & Kahan, T. (2023). New perspectives for hypertension management: Progress in methodological and technological developments.

European Journal of Preventive Cardiology, *30*(1), 48–60. https://doi.org/10.1093/eurjpc/zwac203

Pickering, T. G., Eguchi, K., & Kario, K. (2007). Masked Hypertension: A Review. *Hypertension Research*, *30*(6), 479–488. https://doi.org/10.1291/hypres.30.479

Ramon, Y., Vermeire, T., Toubia, O., Martens, D., & Evgeniou, T. (2021). *Understanding Consumer Preferences for Explanations Generated by XAI Algorithms* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2107.02624

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed). Prentice Hall/Pearson Education.

Sarker, I. H. (2021a). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Sarker, I. H. (2021b). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sawangjai, P., Seesawad, N., & Wilaiprasitporn, T. (2025). Removal of Motion Artifacts From the PPG Signal Using Attentive Generative Adversarial Networks With Dual Discriminator. *IEEE Transactions on Instrumentation and Measurement*, *74*, 1–10. https://doi.org/10.1109/TIM.2025.3529546

Shevtsova, D., Ahmed, A., Boot, I. W. A., Sanges, C., Hudecek, M., Jacobs, J. J. L., Hort, S., & Vrijhoef, H. J. M. (2024). Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study. *JMIR Human Factors*, *11*, e47031. https://doi.org/10.2196/47031

Shimbo, D., Artinian, N. T., Basile, J. N., Krakoff, L. R., Margolis, K. L., Rakotz, M. K., Wozniak, G., & On behalf of the American Heart Association and the American Medical Association. (2020). Self-Measured Blood Pressure Monitoring at Home: A Joint Policy Statement From the American Heart Association and American Medical Association. *Circulation*, *142*(4). https://doi.org/10.1161/CIR.0000000000000803

Sieck, C. J., & Huerta, T. R. (2019). Transformations in health information technology and the impact on patient experience. *Patient Experience Journal*, *6*(2), 5–8. https://doi.org/10.35680/2372-0247.1391

Singh, S., Bennett, M. R., Chen, C., Shin, S., Ghanbari, H., & Nelson, B. W. (2024). Impact of Skin Pigmentation on Pulse Oximetry Blood Oxygenation and Wearable Pulse Rate Accuracy: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *26*, e62769. https://doi.org/10.2196/62769

Statistics Canada. (2024a). *From BpTRU to OMRON: The impact of changing automated blood pressure measurement devices on adult population estimates of blood pressure and hypertension*. https://doi.org/10.25318/82-003-X202401000001-ENG

Statistics Canada. (2024b). *Health indicator statistics, annual estimates* [Dataset]. Government of Canada. https://doi.org/10.25318/1310090501-ENG

Sun, T., Xu, X., Ding, Z., Xie, H., Ma, L., Zhang, J., Xia, Y., Zhang, G., & Ma, Z. (2024). Development of a Health Behavioral Digital Intervention for Patients With Hypertension Based on an Intelligent Health Promotion System and WeChat: Randomized Controlled Trial. *JMIR mHealth and uHealth*, *12*, e53006. https://doi.org/10.2196/53006

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C.,

Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1), 258. https://doi.org/10.1038/s41746-024-01258-7

Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable Photoplethysmographic Sensors—Past and Present. *Electronics*, *3*(2), 282–302. https://doi.org/10.3390/electronics3020282

Tan, S. S.-L., & Goonawardene, N. (2017). Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *Journal of Medical Internet Research*, *19*(1), e5729. https://doi.org/10.2196/jmir.5729

Tešić, M., & Hahn, U. (2022). Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that? *Patterns*, *3*(12), 100635. https://doi.org/10.1016/j.patter.2022.100635

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition* (Updated edition). Penguin Books, an imprint of Penguin Random House LLC.

US Department of Health & Human Services. (2004). *The Decade of Health Information Technology:* https://www.govinfo.gov/content/pkg/GOVPUB-HE-PURL-gpo800/pdf/GOVPUB-HE-PURL-gpo800.pdf

U.S. Food and Drug Administration. (2025a). *510(k) Premarket Notification: Hypertension Notification Feature (HTNF), K250507*. U.S. Food and Drug Administration. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250507

U.S. Food and Drug Administration. (2025b). *K250507: Hypertension Notification Feature (HTNF)* (510(k) Substantial Equivalence Letter). U.S. Food and Drug Administration. https://www.accessdata.fda.gov/cdrh_docs/pdf25/K250507.pdf

VanNostrand, P. M., Hofmann, D. M., Ma, L., & Rundensteiner, E. A. (2024). Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1682–1700. https://doi.org/10.1145/3630106.3658997

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., & Shah, C. (2020). *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2010.10596

Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. https://doi.org/10.48550/ARXIV.1711.00399

Wang, X., & Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650

Warren, G., Keane, M. T., & Byrne, R. M. J. (2022). *Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI*. https://doi.org/10.48550/ARXIV.2204.10152

Weng, W.-H., Baur, S., Daswani, M., Chen, C., Harrell, L., Kakarmath, S., Jabara, M., Behsaz, B., McLean, C. Y., Matias, Y., Corrado, G. S., Shetty, S., Prabhakara, S., Liu, Y., Danaei, G., & Ardila, D. (2024). Predicting cardiovascular disease risk using photoplethysmography and deep learning. *PLOS Global Public Health*, *4*(6), e0003204. https://doi.org/10.1371/journal.pgph.0003204

World Health Organization. (2025, July 31). *Cardiovascular diseases*. World Health Organization. Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases

Yu, J., Park, S., Kwon, S.-H., Cho, K.-H., & Lee, H. (2022). AI-Based Stroke Disease Prediction System Using ECG and PPG Bio-Signals. *IEEE Access*, *10*, 43623–43638. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3169284

Zwack, C. C., Haghani, M., Hollings, M., Zhang, L., Gauci, S., Gallagher, R., & Redfern, J. (2023). The evolution of digital health technologies in cardiovascular disease research. *Npj Digital Medicine*, *6*(1), 1. https://doi.org/10.1038/s41746-022-00734-2

(N.d.). https://doi.org/10.1145/3616019

Appendix A
LLM Evaluation Report

The following outputs from ChatGPT-5.5 and Llama 3.1:8B were evaluated against the QUEST principles for human evaluation of LLMs in health care. Additional safeguards were assessed to complement the criteria established by the principles; specifically, the outputs were verified for required structure, length constraints, technical leakage, sensitive content, and user manipulation. Table 12 presents the evaluation results, with notes added by the author of this thesis. For readability, ChatGPT-5.5 is referred to as LLM A, and Llama 3.1:8B is LLM B.

ChatGPT-5.5 Outputs

1. **Cases 1 & 2**: Hold Still — Walking can make the signal harder to read. Stop walking and keep your wrist still. Button: Try Again
2. **Cases 3 & 4**: Hold Still — Running can make the signal harder to read. Stop running and keep your wrist still. Button: Try Again
3. **Cases 5 & 6**: Rest Still — Coughing can affect the recording. Stay still in a resting position. Button: Try Again
4. **Cases 7 & 8**: Rest Still — Changing light can affect the recording. Stay still in a resting position. Button: Try Again
5. **Cases 9 & 10**: Stay Quiet — Talking can affect the recording. Stay quiet during the measurement. Button: New Reading

Llama 3.1:8B Outputs

1. **Wait It Out**. Laughing likely caused coughing, which affected the recording. Try waiting until coughing settles before retrying. CTA: Retry Measurement.
2. **Keep Still**. Resting while moving your finger on the ear lens affected the recording. Keep the measurement site still and correctly positioned. CTA: Try Again.
3. **Wait It Out**. Coughing during rest likely affected the recording. Wait until coughing settles before retrying. CTA: Retry Measurement

4. **Gentle Pressure**. Higher pressure on the sensor may have caused issues with the measurement. Use lighter, steady contact with the sensor. CTA: Try Again.
5. **Stop and Hold**. Walking may have caused movement that affected the recording. Stop walking and keep the wrist still. CTA: Retry Measurement.
6. **Keep Still**. Moving your finger on the ear lens during rest likely affected the recording. Keep the measurement site still and correctly positioned. CTA: Try Again.
7. **Valid Recording**. (This case has a valid measurement, so there is no prompt to provide.)
8. **Stay Quiet**. Talking during the measurement likely caused issues with the signal quality. Stay quiet during the measurement. CTA: Try Again.
9. **Gentle Pressure**. Higher pressure on the sensor may have caused issues with the measurement. Use lighter, steady contact with the sensor. CTA: Retry Measurement.
10. **Steady Lighting**. Changing lighting conditions likely affected the recording. Move to steadier lighting and keep the sensor clean and dry. CTA: Try Again.

Table 12. Evaluation of the LLM-Generated Prompts following the QUEST Principles

Evaluation Dimension	Evaluation Criterion for Retry Prompt	GPT-5.5 Score	Llama 3.1:8B Score	Evaluator Notes
Required structure	Includes short title, counterfactual, corrective action, and button label.	Pass	Pass	
Length constraints	Title is 3 words or fewer; button label has between 1-3 words.	Pass	Pass	
No technical leakage	Does not mention model internals, probabilities, SHAP, LIME, DiCE, folds, or technical evidence labels.	Pass	Pass	
No sensitive content	Does not mention sensitive, protected, non-actionable, or medical attributes.	Pass	Pass	
No manipulation	Does not use urgency, fear, guilt, shame, pressure, exaggerated benefit, or guaranteed-success language.	Pass	Pass	LLM B is sometimes more directive, but not manipulative.
Principle: Quality of Information				
Accuracy	The counterfactual reflects only the provided evidence and does not invent a cause.	5	5	
Relevance	The prompt addresses the same selected condition and candidate barrier provided for the case.	5	5	
Comprehensiveness	The prompt contains enough information for the user to understand what likely affected the recording and what to change.	5	5	

Evaluation Dimension	Evaluation Criterion for Retry Prompt	GPT-5.5 Score	Llama 3.1:8B Score	Evaluator Notes
Consistency	For repeated cases with the same condition, barrier, convergence, and action hint, wording remains consistent.	5	1	LLM B varies and misaligns repeated conditions.
Usefulness	The corrective action is practical, small, and feasible during a smartwatch measurement retry.	5	4	Some LLM B actions are feasible, but require further guidance, for example "correctly position" requires the user to know or learn how it should be positioned correctly.
Principle: Understanding and Reasoning				
Understanding	The LLM correctly interprets the measurement-quality issue rather than treating it as a health result.	5	5	
Reasoning	The corrective action logically follows from the barrier.	3	5	LLM A struggles to understand the changing-light condition, the recommended action is not that clear.
Evidence handling	If evidence is weak, conflicting, sensitive, or non-actionable, the prompt remains neutral and does not invent a cause.	5	1	LLM B adds unsupported causes and contexts.
Ability focus	The prompt reduces practical friction, and supports retry ability without trying to increase motivation.	5	4	LLM B generally focuses on ability, but wrong factors reduce usefulness of the prompts.

Evaluation Dimension	Evaluation Criterion for Retry Prompt	GPT-5.5 Score	Llama 3.1:8B Score	Evaluator Notes
Principle: Expression Style and Persona				
Clarity	The prompt is easy to understand quickly on a smartwatch interface.	5	2	LLM B is readable, but phrases like “Use lighter steady contact” are not clear.
Conciseness	The wording is brief enough for a small-screen measurement prompt.	5	5	
Neutral tone	The tone is calm, supportive, and non-demanding.	5	4	LLM B includes more command-like titles such as “Stop and Hold.”
Persona fit	The prompt sounds like a measurement-quality facilitator, not a coach, warning, diagnosis, or command.	5	2	LLM B language sounds like a generic device troubleshooting prompt rather than a facilitator prompt.
Button quality	The button label is neutral, tappable, optional, and suitable for a smartwatch interface.	5	5	
Principle: Safety and Harm				
Bias avoidance	The prompt avoids protected-attribute assumptions or differential treatment.	5	5	Neither response introduces protected-attribute assumptions.
Harm avoidance	The prompt does not create pressure, blame, or medical concern.	5	2	LLM B’s statements like “your finger” wording mildly increases user blame.
No misleading certainty	The prompt does not imply that retrying will guarantee a valid result.	5	5	
No fabrication	The prompt does not add unsupported context, unsupported causes, or unsupported benefits.	5	1	LLM B fabricates several barriers that are not present in the case evidence.

Evaluation Dimension	Evaluation Criterion for Retry Prompt	GPT-5.5 Score	Llama 3.1:8B Score	Evaluator Notes
Autonomy preservation	The prompt presents retrying as an available option, not a requirement.	5	4	Some of LLM B CTAs might be considered as a requirement, specifically "Retry Measurement".
Principle: Trust and Confidence				
Transparency	The user can understand why the suggested action may help the measurement.	4	2	The recommended action of LLM A for the changing-light conditions scenarios is not consistent with the counterfactual provided. LLM B often explains a reason, but the reason is not supported by the case. For example, "Resting while moving your finger on the ear lens affected the recording.".
Appropriate confidence	The prompt is specific enough to be useful but not overconfident.	5	4	LLM B was overconfident when it processed a second counterfactual. For example, "Laughing likely caused coughing, which affected the recording", this suggest certainty that the cause of the unsuccessful measurement was coughing.
Reliability	The wording would likely produce stable evaluations across similar cases.	5	1	LLM B's output is inconsistent between cases of the same condition.
User satisfaction	The prompt is likely to feel clear, fair, and usable to the person taking the measurement.	5	1	LLM B could feel confusing or unfair because several causes do not match the evidence. For example: "Laughing likely caused coughing".

Appendix B
Chapter 4 Study Materials

In this appendix, we include all the materials used in the human validation study presented in Chapter 4. Appendix B.1 presents the materials used to recruit participants. Appendix B.2 contains the materials participants were required to complete before the study session. Appendix B.3 documents the materials used by the researcher and participants during the study session. Appendix B.4 outlines the documents shared with participants following the study session. Finally, Appendix B.5 provides the ethics clearance certificate granted by the University of Waterloo Research Ethics Board for this study.

Appendix B.1 Recruitment Materials

Participants in this study were recruited in three ways: First, by email to faculty members, undergraduate students, and graduate students from the departments of the faculties of Math, Engineering, and Health. The email sent is presented in Appendix B.1.1. Furthermore, the poster shown in Appendix B.1.2 was posted in the Waterloo Undergraduate Student Association (WUSA) announcement boards in the Student Life Center (SLC) and on all the Elevator Doors in Engineering 5 and 7 (Pearl Sullivan Engineering (PSE) Building). Finally, a LinkedIn post like the one shown in Appendix B.1.3 was shared on the Advanced Interface Lab's and the student researcher's LinkedIn profiles.

Appendix B.1.1 Recruitment Email

Outlook

Study Recruitment: Tap, Think, Act Help us improve Wearable Heart Monitoring!

From Raul Ramirez Batista <raul.ramirez@uwaterloo.ca>
Date Fri 15-May-26 5:57 AM

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING Department of Systems Design Engineering | ADVANCED INTERFACE DESIGN LAB



TAP, THINK, ACT!

Looking for Participants for a Research Study!
Help us improve wearable heart monitoring.
Let's make smartwatch health alerts easier to understand and act on. Your perspective matters!

Hello,

My name is *Raul Ramirez,* and I am a master's student working under the supervision of *Dr. Catherine Burns* in the Department of *Systems Design Engineering* at the University of Waterloo. As part of my master's degree, I am conducting a research study on AI-enabled tools in wearable heart monitoring.

About this study

We are a Canada-based research team developing an AI-enabled system designed to help people take clearer heart measurements using wearable devices.

Many wearables use PPG (photoplethysmography), a light-based sensor that shines light into the skin to detect changes in blood flow. This is how smartwatches measure your heart rate throughout the day. Some devices can also record a single-lead ECG. ECG stands for electrocardiogram, which is a test that measures the heart's electrical activity. For example, on a smartwatch, you may press the watch's crown or

sensor for about 30 seconds to record a heart rhythm reading.

Our system is designed to recognize when these heart readings may not be taken under ideal conditions and provide simple suggestions to help improve the measurement. We want to understand what works well, what is challenging, and how to improve the overall experience for people using wearables to monitor their heart health.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB ***47991***).

Who can participate?

We are seeking adults aged 18-65 with a basic eHealth Literacy (have the ability to find, understand, evaluate, and use health information from digital sources) that own a wearable device capable of recording single-lead ECG's.

What will you do?

If you are eligible for this study, you will be invited to book a 45-minute session using a Microsoft Bookings link provided after completing the questionnaire. The session will take place online through Microsoft Teams. You will need a stable internet connection and must keep your camera on for the duration of the session.

During the session, you will interact with wearable device interfaces that simulate heart (ECG) recordings and share your feedback.

How can you help?

Your feedback will help shape the design of wearable interfaces by informing how we can improve PPG-based single-lead ECG measurements while enhancing the overall user experience for heart monitoring devices.

To participate, **you will need** to have access to a laptop or a desktop computer with Google Chrome and uninterrupted internet access for 45 minutes.

Our commitment

Progress in healthcare takes all of us. We are deeply committed to ensuring that the AI tools are built with the insights of users at their core. Your participation is a vital part of this commitment, and as a token of our appreciation for your time, you will receive **CAD 20.**

Ready to get involved?

If you are interested in participating or would like to learn more, please reply to this email or click the button below to complete our screening questionnaire and get the session scheduled.

✅Count me in!

If you require additional information to assist you in reaching a decision, please contact me at *raul.ramirez@uwaterloo.ca* .
Sincerely,

Raul Ramirez, BSc.
He/Him/His
MASc. Student
University of Waterloo

Figure 9. Recruitment Email

Appendix B.1.2 Recruitment Poster



Figure 10. Recruitment Poster

Appendix B.1.3 Recruitment LinkedIn Post

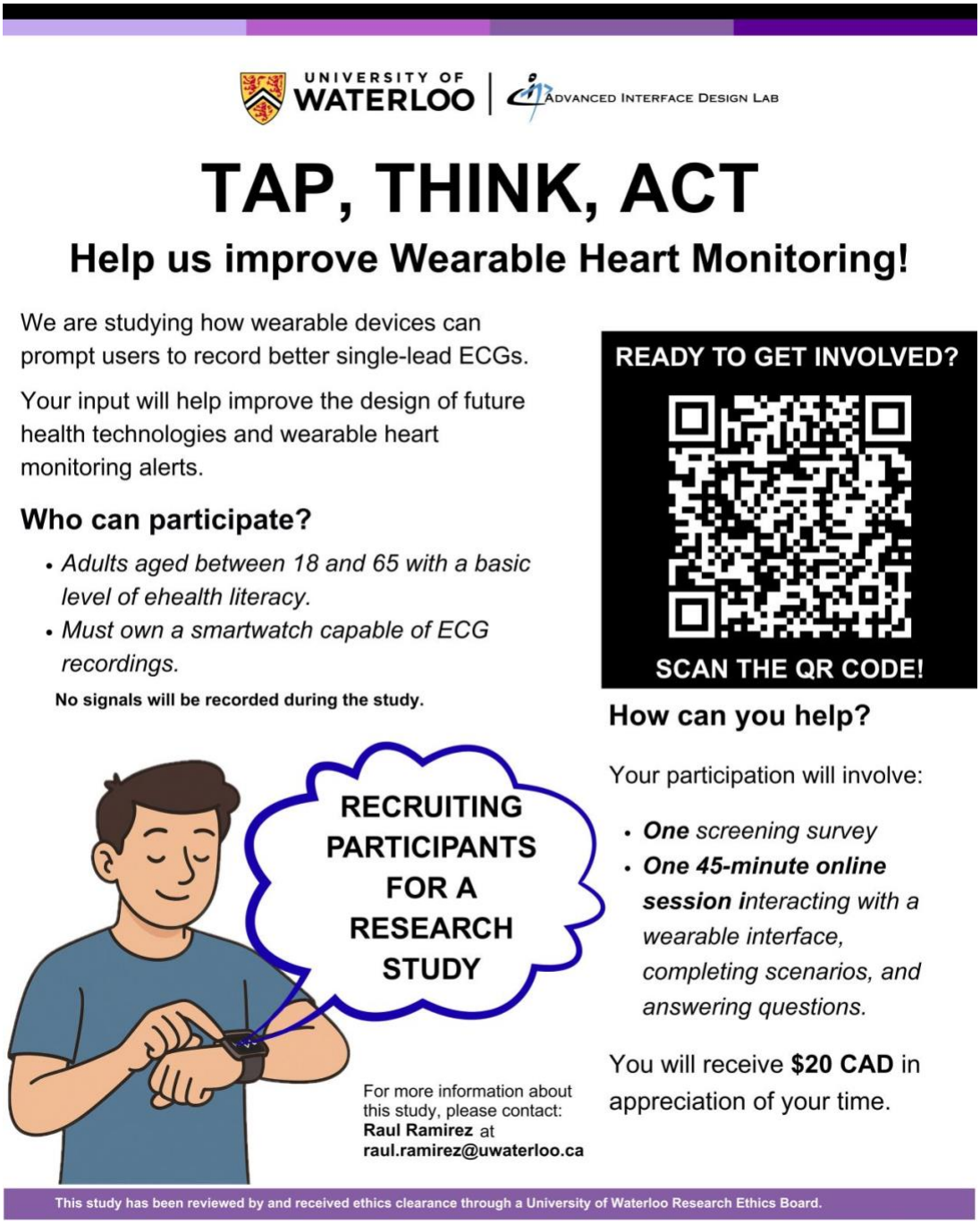


Figure 11. Recruitment LinkedIn Post

Appendix B.2 Pre-Study Session Materials

Participants who followed the recruitment materials were presented with an information consent letter (Appendix B.2.1), had to complete a consent form (Appendix B.2.2). Those participants that gave their consent to participate in the study, completed a Demographics and Screening Questionnaire (Appendix B.2.3). If they met the inclusion criteria, they were provided with a Participant ID and redirected to a Microsoft Booking link (Appendix B.2.4), where they booked the study session at a time that worked for them.

Appendix B.2.1 Information Consent Letter



INFORMATION–CONSENT LETTER

Study Title

A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

Research Team

Principal Investigator

Dr. Catherine Burns
Advanced Interface Design Lab
Department of Systems Design Engineering
University of Waterloo
Canada

Student Investigator

Raul Ramirez
Advanced Interface Design Lab
Department of Systems Design Engineering
University of Waterloo
Canada

You are invited to participate in a research study about how people respond to notifications from wearable heart-monitoring devices. To help you make an informed decision about your participation, this letter explains what the study is about, the possible risks and benefits, and your rights as a research participant.

Please read this carefully before deciding whether to participate. If you do not understand something in the letter, please ask one of the investigators before agreeing to participate. You may wish to save a copy of this letter for your records.

Purpose of the Study

Many smartwatches use photoplethysmography (PPG), a light-based sensor that measures changes in blood flow through the skin to estimate heart rate and other signals. Sometimes PPG measurements are unsuccessful (for example, due to movement or poor contact with the skin). Some smartwatches also allow users to record a single-lead electrocardiogram (ECG). An ECG (electrocardiogram) measures the heart's electrical activity. On a smartwatch, this typically involves placing a finger on the device for about 30 seconds to record a heart rhythm reading.

This study examines how different smartwatch notifications influence whether users choose to take further action (such as recording an ECG) after an unsuccessful PPG reading. We are comparing: a standard notification vs a redesigned notification intended to provide clearer guidance. The goal is to better understand how wearable devices can communicate health information more clearly and effectively.

Who May Participate?

Up to 40 adults. To participate, you must be 18–65 years old, own a smartwatch capable of single-lead ECG.

What Does Participation Involve?

Participation includes:

- An electronic screening questionnaire
- Scheduling your session through Microsoft Bookings
- A 60-minute virtual session via Microsoft Teams
- A short post-session questionnaire (exit questionnaire)

Participants will complete the eHealth Literacy Scale (eHEALS), which measures perceived confidence in using online health information. This questionnaire is used for research analysis only and is not used to determine eligibility. Responses will be analyzed at the University of Waterloo. After the post-study form, no further involvement is required.

Study Platforms

You will be participating in a study that uses Qualtrics, Microsoft Teams, and Figma. Qualtrics, Microsoft Teams, and Figma have implemented technical, administrative, and physical safeguards to protect the information from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no internet transmission is ever fully secure or error-free. Qualtrics temporarily collects your IP address to prevent fraudulent participation, including duplicate responses.

Before the Study Session

Please ensure you have:

- A desktop or laptop with Google Chrome; screen sharing, screen recording, and audio/video recording enabled.
- Uninterrupted internet access for 45 minutes.
- Closed any windows containing personal or sensitive information (your screen, video, and voice will be recorded).

Recordings are used to understand how participants interpret notifications and make decisions. Recordings will be transcribed for qualitative analysis.

During the Study Session

During the session, you will:

- Watch short simulated (AI-generated) videos showing users attempting heart measurements
- Interact with a simulated prototype of the wearable interfaces
- Think aloud while explaining your decisions
- Answer short follow-up questions
- Provide overall feedback

The wearable interface is a simulated research prototype. It is not a medical device, it is not intended to provide medical advice, and is for demonstration purposes only.

Demographic Information

We will collect limited demographic information, including:

- Age
- Education level
- General location (based on IP address)

IP address location verification is used to confirm eligibility and prevent fraudulent participation. You must be physically located in Canada, the United States, Ireland, or Panama at the time of participation. You will also complete a brief eHealth literacy questionnaire (eHEALS). This measures comfort and confidence in finding and using health information online. Demographic information helps us interpret results, as research shows that age, education, and digital health literacy can influence technology use.

Risks

There are no known or anticipated risks beyond those encountered in normal online activities. Participation does not increase your risk of data breach beyond typical internet use. We may collect IP addresses to verify participant location. Although data breaches are rare, strong security measures are in place to safeguard your information.

Possible Benefits

There may be no personal benefit to you. The study may help improve how wearable devices communicate health information to users. It will provide insights into how counterfactuals can influence user behavior in machine-learning-powered wearables and may guide designs that overcome algorithm limitations.

Confidentiality and Data Security

Your identity and any information that could link you to your responses will be kept confidential.

- Your study data will be associated with a participant ID number.
- Your name and email address will be stored separately from study data.
- Recordings and transcripts will be stored separately from identifying information.

Data will be stored on secure University of Waterloo servers protected by encryption and password protection and retained for 7 years.

When reporting results:

- Data will be de-identified before analysis and publication.
- Aggregated findings may appear in reports, presentations, or academic publications.
- De-identified quotations may be used with your permission.
- De-identified screenshots (without face or identifying features) may be used if consent is given

Secondary Use of Data (Optional)

With your permission, de-identified transcripts from your session may be used in future related research conducted or supervised by this research team.

Before use:

- All personally identifying information will be removed, redacted, or modified.
- Minor wording changes may be made to protect identity while preserving meaning.

These transcripts may be used for qualitative coding, thematic analysis, and related qualitative research methods.

Any future use will undergo review and receive clearance from a Research Ethics Board.

You may choose whether or not to consent to this optional future use.

Remuneration

You may choose whether or not to receive remuneration. If you choose to receive remuneration, you will receive **$20 CAD** as a token of appreciation for your time, sent to the email address you provide during the session, within up to 20 business days. The amount is taxable; you are responsible for reporting it if required.

Voluntary Participation and Withdrawal

Participation is voluntary. You may:

- Decline to answer any question
- Stop participation at any time
- Withdraw your data before **June 30, 2026**

After study results have been submitted for publication, withdrawal may no longer be possible.

To withdraw, contact: raul.ramirez@uwaterloo.ca

Ethics Clearance

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB 47991). For questions for the Board, contact the Office of Research Ethics: toll-free 1-833-643-2379 (Canada/USA), 1-519-888-4440, or reb@uwaterloo.ca.

Questions

For questions about participation or additional information, contact Raul Ramirez at raul.ramirez@uwaterloo.ca.

Thank You!

Thank you for considering participation. Your insights will help improve wearable heart-health monitoring.

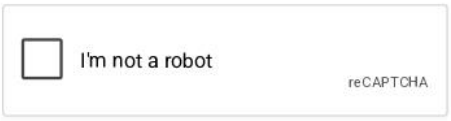


Figure 12. Information Consent Letter

Appendix B.2.2 Consent Form & Secondary Use of Data



How This Study Session Will Work

The study session involves interacting with a prototype and providing feedback. To capture this feedback effectively, we will record the session as described below.

- **Meeting Platform:** The study session will take place via a Microsoft Teams meeting, which will be recorded.
- **Your Responsibility:** Before the session begins, you must close all other windows and applications that may contain personal or sensitive information (e.g., email, social media, personal files), ensure your camera and microphone are working, and that you have a stable internet connection.

Consent & Acknowledgements

Please read each statement carefully and indicate your agreement.

I confirm that:

- I have read the Information Letter and am aware of the study procedures.
- I am aware that participation is voluntary.
- I am aware that I may withdraw before June 30, 2026.
- I am aware that my identity will be kept confidential and my data will be de-identified for reporting.
- I am aware that my screen, audio, and video will be recorded and transcribed.
- I am aware that I may choose whether or not to receive $20 CAD as remuneration.

By selecting 'I agree,' I indicate that I have read the Information Letter and consent to participate in this study.

*I understand that I am not waiving my legal rights or releasing the investigators or their institutions from their legal and professional responsibilities.

◯ I agree to participate in this study.

◯ I do not agree to participate in this study.

Figure 13. Consent Form

The message shown in Figure 14 is displayed if the participant does not agree to participate in the study.



Thank You for Completing the Questionnaire

We sincerely appreciate your time and interest in participating in our research study. Based on your responses, you **have not agreed to** participate in this study.

Your responses will not be analyzed and will be securely deleted in accordance with our data management protocol.

If you have any questions, feel free to contact Raul Ramirez at raul.ramirez@uwaterloo.ca.

Thank you again for your interest and your contribution to our research efforts.

Dr. Catherine Burns
Professor, Canada Research Chair in Human Factors in Healthcare Systems, University of Waterloo

Raul Ramirez
MASc Student in Systems Design Engineering, University of Waterloo

Figure 14. Declined Consent Message

*Future use of data

Do you consent to the use of your anonymized data in future related research projects conducted or supervised by the members of the research team, subject to ethics clearance?

- ◯ Yes
- ◯ No

*Use of anonymized quotes

I consent to the use of de-identified quotations in publications or presentations. (e.g., included in reports, presentations, or publications as "Participant 3 expressed...", without identifying you personally)?

- ◯ Yes
- ◯ No

*Use of anonymized screenshots

Do you consent to the use of anonymized screenshots from the screen recordings of your interaction with the system (no faces, audio, or video), for use in reports, presentations, or publications?

- ◯ Yes
- ◯ No

Figure 15. Secondary Use of Data

Appendix B.2.3 Demographics and Screening Questionnaire



*How old are you?

- Under 18
- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 55-64 years old
- 65+ years old

What is the highest level of education you have completed?

- Some Primary
- Completed Primary
- Some Secondary
- Completed Secondary
- Vocational or Similar
- Some University but no degree
- University Bachelors Degree
- Graduate or Professional Degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)
- Prefer not to say

Figure 16. Demographics and Screening Questionnaire - Demographics Questions

*Do you own a smartwatch capable of ECG recordings?

If your smartwatch is on this list, you should have access to ECG monitoring: Hume Band, WHOOP MG, Withings ScanWatch 2, Withings ScanWatch Nova, Huawei Watch 5, Huawei Watch D2, Huawei Watch GT 6 Pro, Fitbit Sense 2, Fitbit Charge 6, Fitbit Charge 5, Apple Watch Series 11, Apple Watch Series 4–10, Apple Watch Ultra 3, Apple Watch Ultra 1–2, Samsung Galaxy Watch 8 series, Samsung Galaxy Watch 4–7, Samsung Galaxy Watch Ultra, Google Pixel Watch 4, Google Pixel Watch 1–3, Garmin Fenix 8, Garmin Venu 4, Garmin Forerunner 970, Coros Pace Pro, Coros Vertix 2, Coros Apex 2

◯ Yes

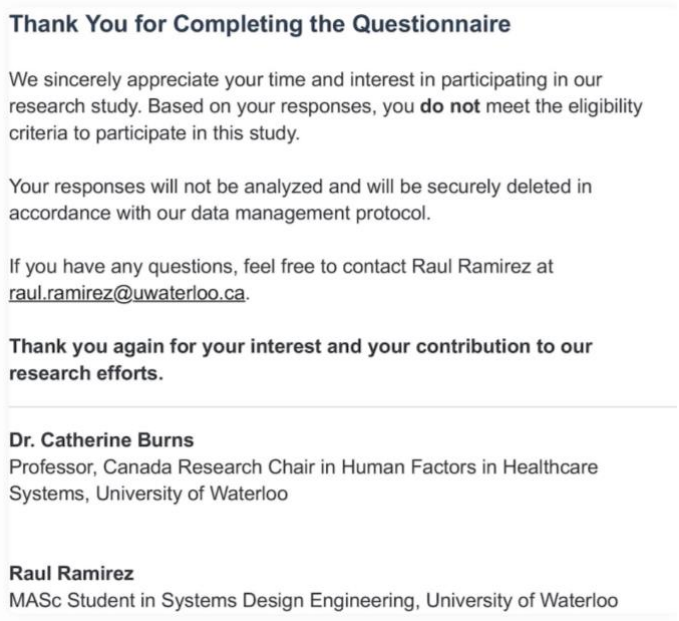
◯ No

*Have you recorded at least one ECG with it before the study using your smartwatch?

◯ Yes

◯ No

Figure 17. Demographics and Screening Questionnaire - Screening Question



Thank You for Completing the Questionnaire

We sincerely appreciate your time and interest in participating in our research study. Based on your responses, you **do not** meet the eligibility criteria to participate in this study.

Your responses will not be analyzed and will be securely deleted in accordance with our data management protocol.

If you have any questions, feel free to contact Raul Ramirez at raul.ramirez@uwaterloo.ca.

Thank you again for your interest and your contribution to our research efforts.

Dr. Catherine Burns
Professor, Canada Research Chair in Human Factors in Healthcare Systems, University of Waterloo

Raul Ramirez
MASc Student in Systems Design Engineering, University of Waterloo

Figure 18. Study Ineligibility Notification

*eHealth Literacy Scale (eHEALS)

(Norman & Skinner, 2006)

Please rate your agreement with each statement on a scale from 1 = Strongly disagree to 5 = Strongly agree.



Figure 19. Demographics and Screening Questionnaire - eHEALS Scale

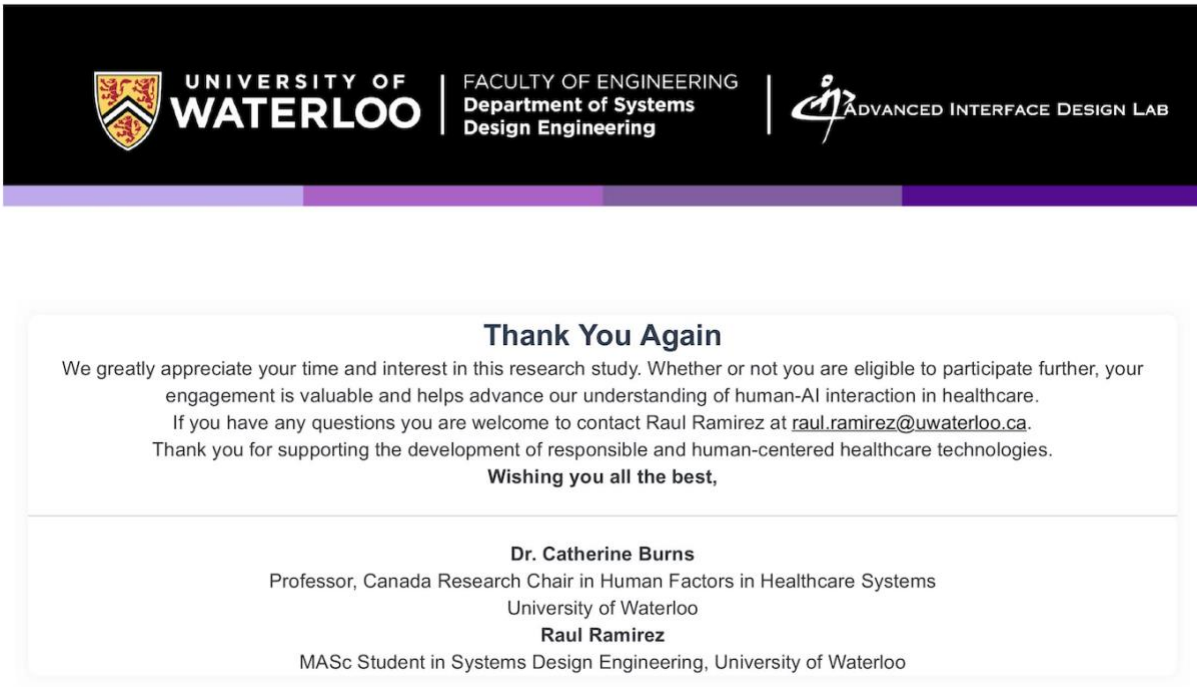


Figure 20. Demographics and Screening Questionnaire - End of the Questionnaire

Appendix B.2.4 Study Session Booking Site

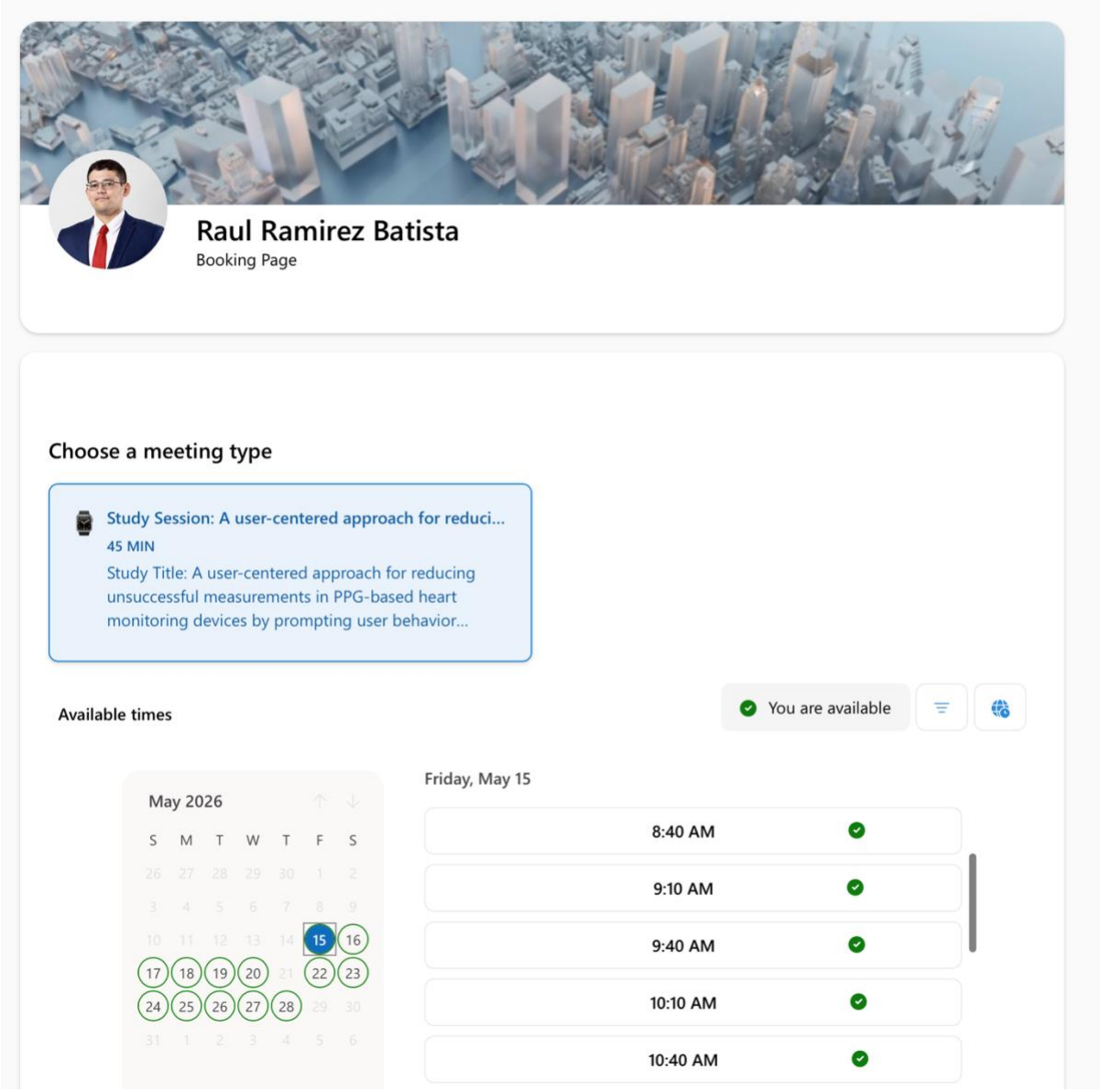


Figure 21. Study Session Booking Site

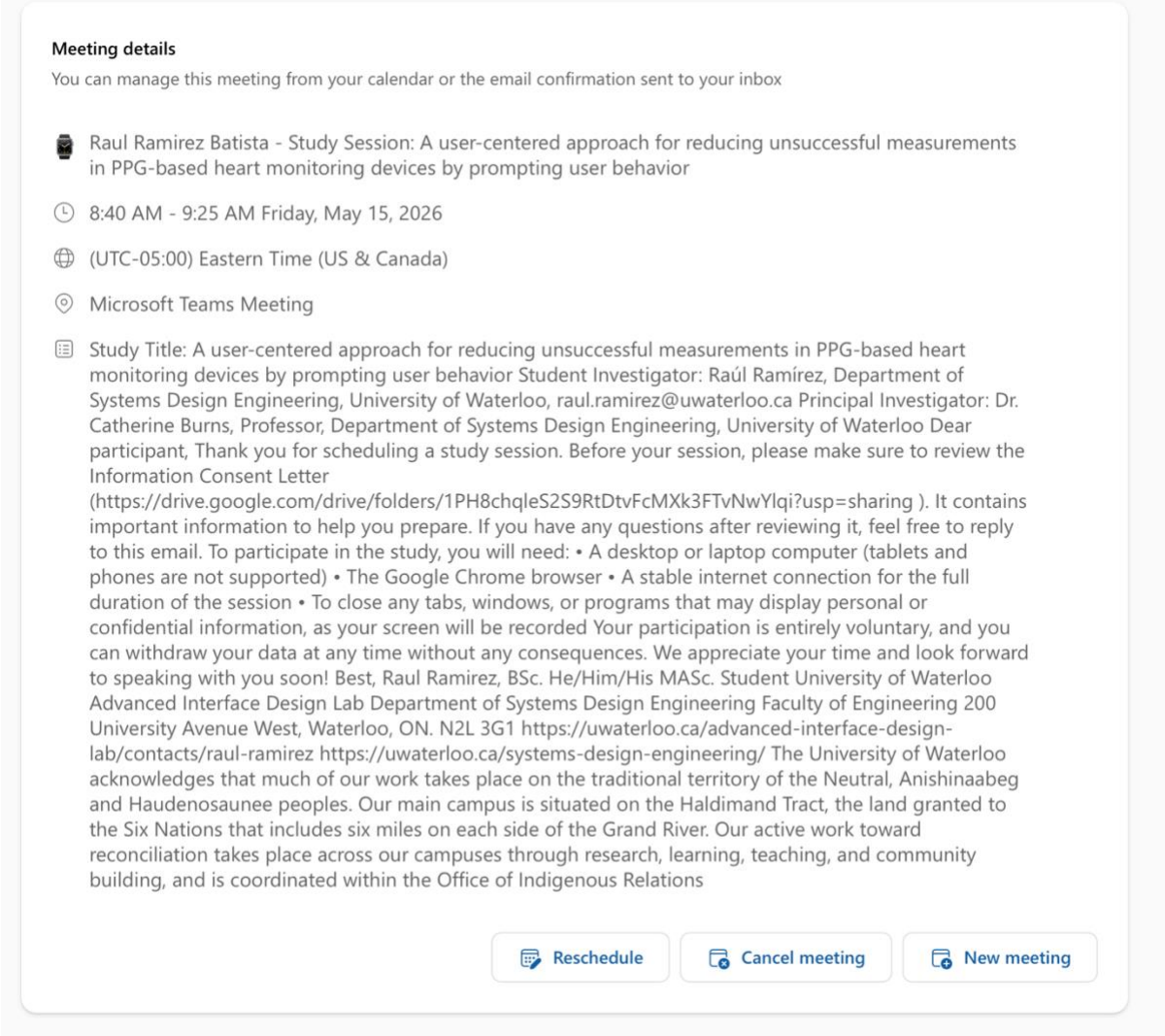
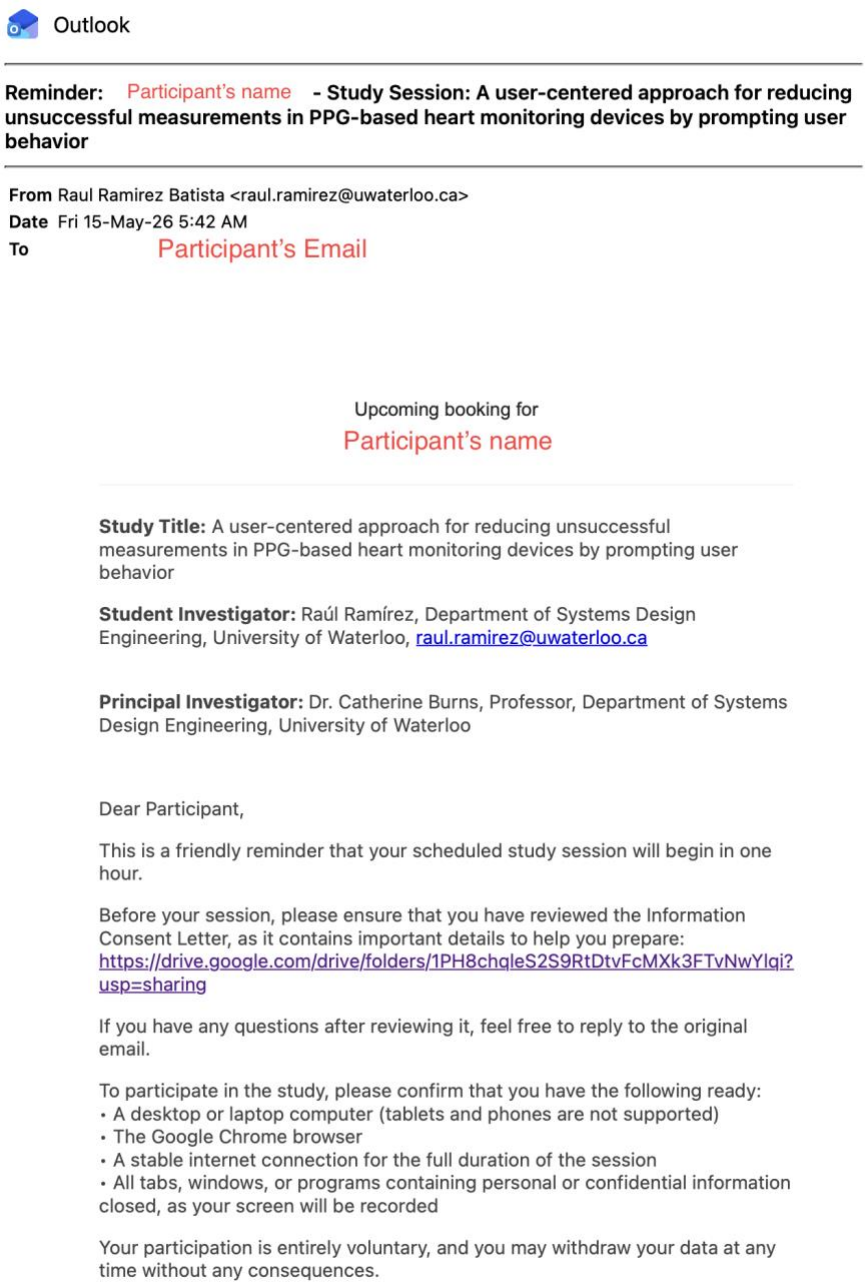


Figure 22. Booking Confirmation Message

Appendix B.2.5 Reminder of the Booking Session



Outlook

Reminder: Participant's name **- Study Session: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior**

From Raul Ramirez Batista <raul.ramirez@uwaterloo.ca>
Date Fri 15-May-26 5:42 AM
To Participant's Email

Upcoming booking for
Participant's name

Study Title: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

Student Investigator: Raúl Ramírez, Department of Systems Design Engineering, University of Waterloo, raul.ramirez@uwaterloo.ca

Principal Investigator: Dr. Catherine Burns, Professor, Department of Systems Design Engineering, University of Waterloo

Dear Participant,

This is a friendly reminder that your scheduled study session will begin in one hour.

Before your session, please ensure that you have reviewed the Information Consent Letter, as it contains important details to help you prepare: https://drive.google.com/drive/folders/1PH8chqleS2S9RtDtvFcMXk3FTvNwYlqi?usp=sharing

If you have any questions after reviewing it, feel free to reply to the original email.

To participate in the study, please confirm that you have the following ready:
- A desktop or laptop computer (tablets and phones are not supported)
- The Google Chrome browser
- A stable internet connection for the full duration of the session
- All tabs, windows, or programs containing personal or confidential information closed, as your screen will be recorded

Your participation is entirely voluntary, and you may withdraw your data at any time without any consequences.

We appreciate your time and look forward to speaking with you shortly.

Best regards,

Raul Ramirez, BSc.
MASc. Student
University of Waterloo
Advanced Interface Design Lab
Department of Systems Design Engineering
Faculty of Engineering

Participant's Name Study Session: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

Friday, May 15, 2026
5:10 AM - 5:55 AM
(UTC-05:00) Eastern Time (US & Canada)

Join your appointment

via Microsoft Teams

Powered by Microsoft Bookings

Privacy Statement
Microsoft Corporation • One Microsoft Way • Redmond, WA 98052

Figure 23. Study Session Booking Reminder Email

Appendix B.3 Study Session Materials

During the study session, the student researcher followed the guide provided in Appendix B.3.1. Participants used the Participant Study Platform (Appendix B.3.2) and completed the Data Collection Form (Appendix B.3.3).

Appendix B.3.1 Researcher Study Guide

PPG Study Guide

Start of Block: Default Question Block

Study title: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

---

Participant ID: If they do not have one, use this generator:
https://uwaterloo.ca1.qualtrics.com/jfe/form/SV_5iMdqo7wfrzYUxo

_____

Assigned Interface Order

▼ A - B (1) ... B - A (2)

Display this question:

If Assigned Interface Order = A - B

AB ORDER https://www.figma.com/proto/wGR2IJadJfwxqCfhImsUGR/PPG-Quality-Improvement-Study?page-id=0%3A1&node-id=3-119&viewport=610%2C315%2C0.08&t=r5XHpyaiacAemCFA-1&scaling=scale-down&content-scaling=fixed&starting-point-node-id=3%3A119&show-proto-sidebar=1

Display this question:

If Assigned Interface Order = B - A

BA ORDER https://www.figma.com/proto/gqom6Qktoli82loyp9E8XA/A-PPG-Quality-Improvement-Study?page-id=0%3A1&node-id=3-119&viewport=367%2C133%2C0.04&t=k9CvqCKv6xOmwhoL-1&scaling=scale-down&content-scaling=fixed&starting-point-node-id=3%3A119&show-proto-sidebar=1

Hello [Participant's Name], thank you for joining today. My name is Raul Ramirez, I am a master's Student at the University of Waterloo, and I will guide you through the study session. Just for the record, your consent was already obtained before this session via the Qualtrics form in the study materials. Before we begin, do you have any questions about the information consent letter or what you will be doing today?

Did the participant ask any questions?

- Yes (1)
- No (2)

Display this question:

If did the participant asked any questions? = Yes

Write here any questions asked by the participant.

_____

_____

_____

_____

_____

As described in the information consent letter, this session will be recorded, including your screen, voice, and video. The recording will help the research team understand how participants interpret smartwatch messages and make decisions. Please let me know if you are ready for me to start the recording. Examiner does: Start recording only after the participant confirms. Confirm that the recording has started.

Before we begin, please confirm that:

- ☐ You are using a desktop or laptop. (1)
- ☐ You have a stable internet connection. (2)
- ☐ You have closed any windows with personal or sensitive information. (3)
- ☐ You are ready to share your screen. (4)
- ☐ Your camera and microphone are activated. (5)

Please start sharing your screen. In this study, you will use a Figma prototype and a data collection form. You will watch 10 short videos where someone tries to take an ECG reading on a smartwatch, but the reading fails. For each video, imagine you're the person in that situation. I will ask what you think happened, why it failed, and what you would do next. After that, you will interact with the Figma prototype as you normally would with a watch. The crown button works, and you can scroll when you see a scroll bar. Then I will ask a few questions about the interface. Just to clarify, these are simulated examples. No real ECG readings are being taken. The prototype is for research only, not a medical device, and it does not provide medical advice. I am not a medical professional, so I cannot answer medical questions. There are no right or wrong answers. We are interested in what you naturally think and do. Do you have any questions before we begin?

Write here any questions asked by the participant

_____

_____

_____

_____

Before we begin with the cases, I will ask a few background questions about your smartwatch use.

Which smartwatch do you primarily use?

- Hume Band (1)
- WHOOP MG (2)
- Withings ScanWatch 2 (3)
- Withings ScanWatch Nova (4)
- Huawei Watch 5 (5)
- Huawei Watch D2 (6)
- Huawei Watch GT 6 Pro (7)
- Fitbit Sense 2 (8)
- Fitbit Charge 6 (9)
- Fitbit Charge 5 (10)
- Apple Watch Series 11 (11)
- Apple Watch Series 4-10 (12)
- Apple Watch Ultra 3 (13)
- Apple Watch Ultra 1-2 (14)
- Samsung Galaxy Watch 8 series (15)
- Samsung Galaxy Watch 4-7 (16)
- Samsung Galaxy Watch Ultra (17)
- Google Pixel Watch 4 (18)
- Google Pixel Watch 1-3 (19)
- Garmin Fenix 8 (20)
- Garmin Venu 4 (21)
- Garmin Forerunner 970 (22)
- Coros Pace Pro (23)
- Coros Vertix 2 (24)
- Coros Apex 2 (25)
- Other (26) _____

How long have you been using your smartwatch?

- Less than 6 months (1)
- 6 months up to a year (2)
- More than 1 year (3)
- More than 3 years (4)
- More than 5 years (5)
- Other (6) _____

Approximately how many times have you recorded an ECG using your smartwatch?

- Never (1)
- Less than 5 times (2)
- Between 5 to 10 times (3)
- Daily (4)
- Weekly (5)
- Monthly (6)
- Yearly (7)
- Other (8) _____

How confident are you in using the ECG feature correctly?

0 (Not at All Confident) - 10 (Very Confident)

Awesome, we are ready to start. The Figma link I sent in the chat will guide you through the study. As you go through each scenario, please think aloud as much as possible. Tell me what you notice, what you think is happening, what you would do, and why. Follow-up prompts: "What are you thinking now?" "What are you noticing?" "What would you do next?" "Can you tell me why?" "Please keep thinking aloud."

Examiner Notes

_____

_____

_____

These are questions that the participants might ask:

- When a question says, “this situation,” it means the full situation shown in the video. That includes having chest pain or discomfort while doing, or having just done, the activity shown in the video, such as walking, coughing, talking, running, or a combination of these.
- When a question says, “right now,” it means immediately after the unsuccessful ECG attempt shown in the video.
- When a question says, “without guidance,” it means before seeing any smartwatch message or instruction about what to do next.
- When a question says, “take action” or “act,” it means doing something after seeing the smartwatch message. This could include trying again, adjusting the watch, sitting still, waiting before trying again, asking someone for help, contacting a health care professional, or seeking emergency help.
- When a question says, “record an ECG” or “perform an ECG,” it means using the smartwatch’s ECG feature to capture a heart rhythm reading. No real ECG is being recorded in this study.

Please answer as if you were the person in the video or as if the situation were happening to you.

Thank you for participating in this study. Your feedback will help the research team understand how wearable heart-health monitoring devices can communicate information more clearly. You will receive an email shortly after this session with instructions about remuneration. As described in the information consent letter, you may choose whether or not to receive the $20 CAD remuneration. If you choose to receive it, it will be sent to the phone number or email address you provide, within up to 20 business days. As a reminder, your participation is voluntary. You may withdraw your data at any time until June 30, 2026, via email at raul.ramirez@uwaterloo.ca. You may refer back to the information consent letter at any time for details about withdrawal, confidentiality, remuneration, and who to contact with questions. Do you have any questions?

Write here any questions that the participant may have

_____

Great. Thank you once again for participating in this study. That concludes the study session. I will now stop the recording.

Appendix B.3.2 Participant Study Platform

The template presented from Figure 24 to Figure 30 was used to support the standardization of the study, ensuring that all participants received the same information and followed the same protocol. The template was adapted to include the scenarios presented in Chapter 3 for all ten cases.

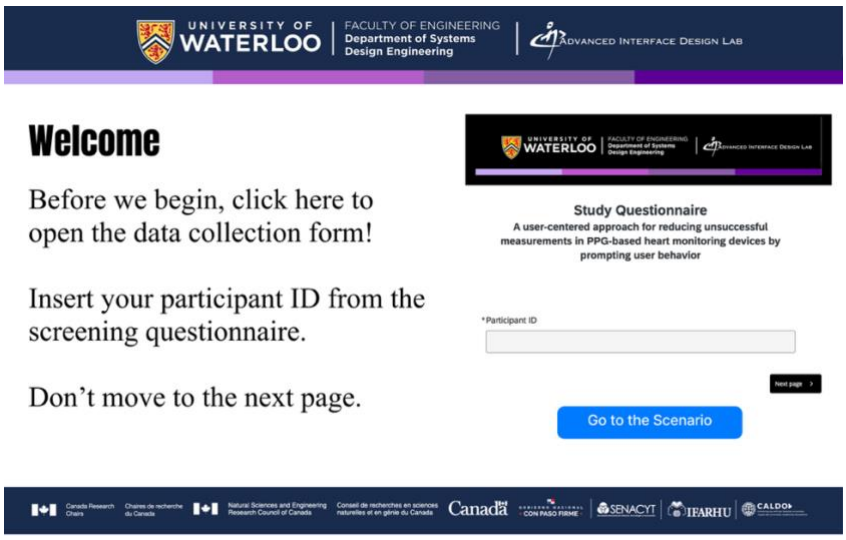


Figure 24. Participant Study Platform - Welcome Screen



Figure 25. Participant Study Platform - Start of Scenario

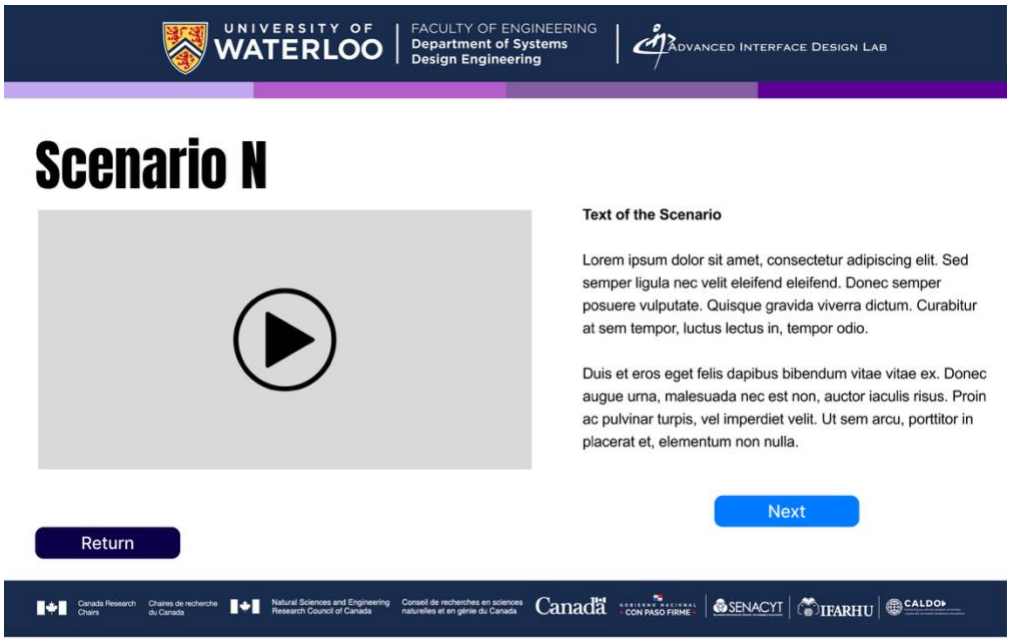


Figure 26. Participant Study Platform - Video and Description of the Scenario

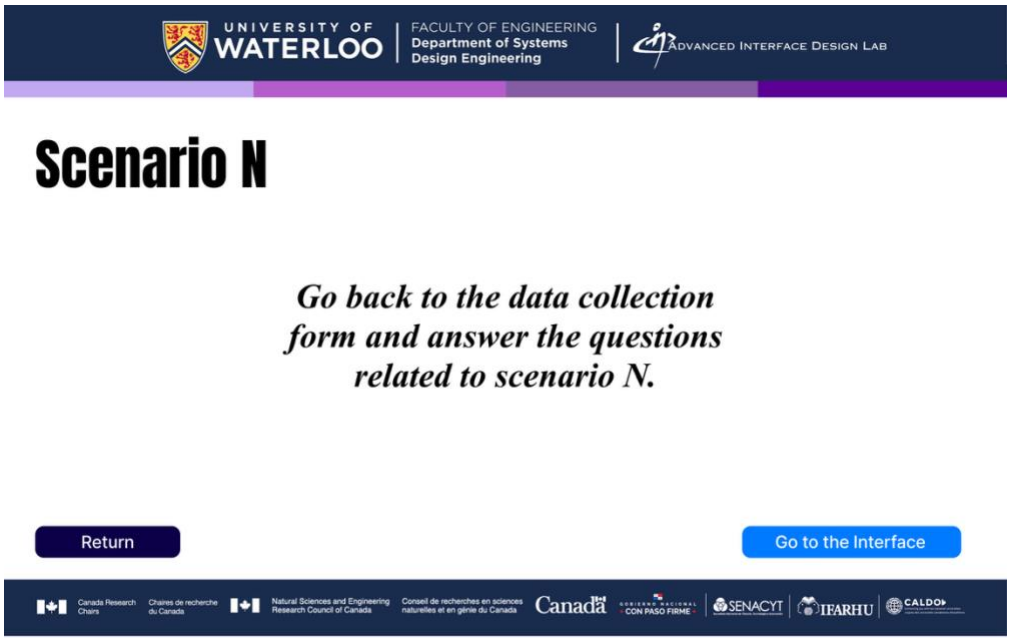


Figure 27. Participant Study Platform - Redirection to the data collection form

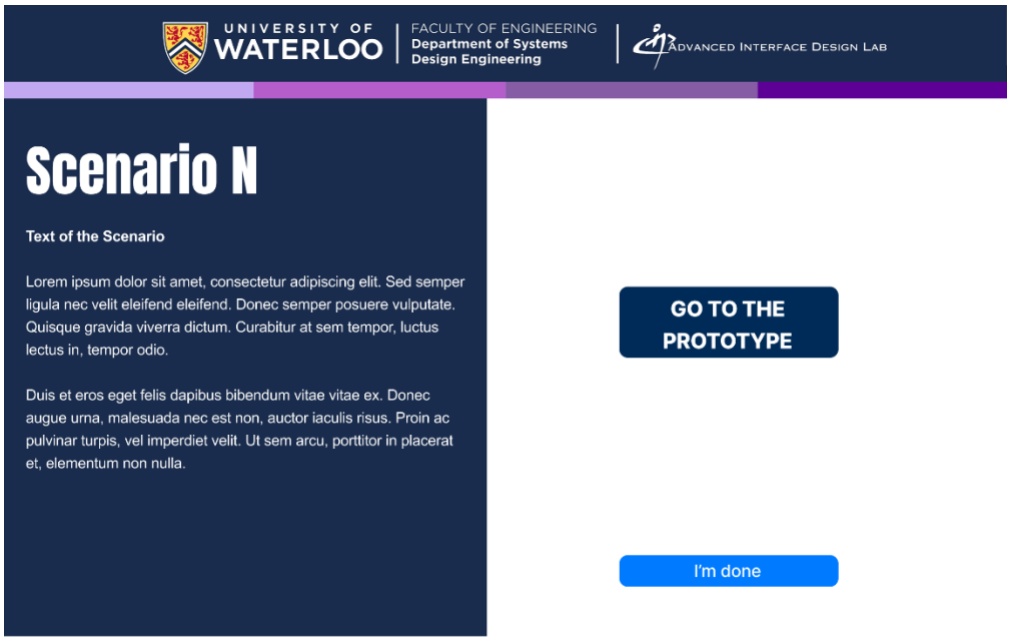


Figure 28.Participant Study Platform - Redirection to the prototype

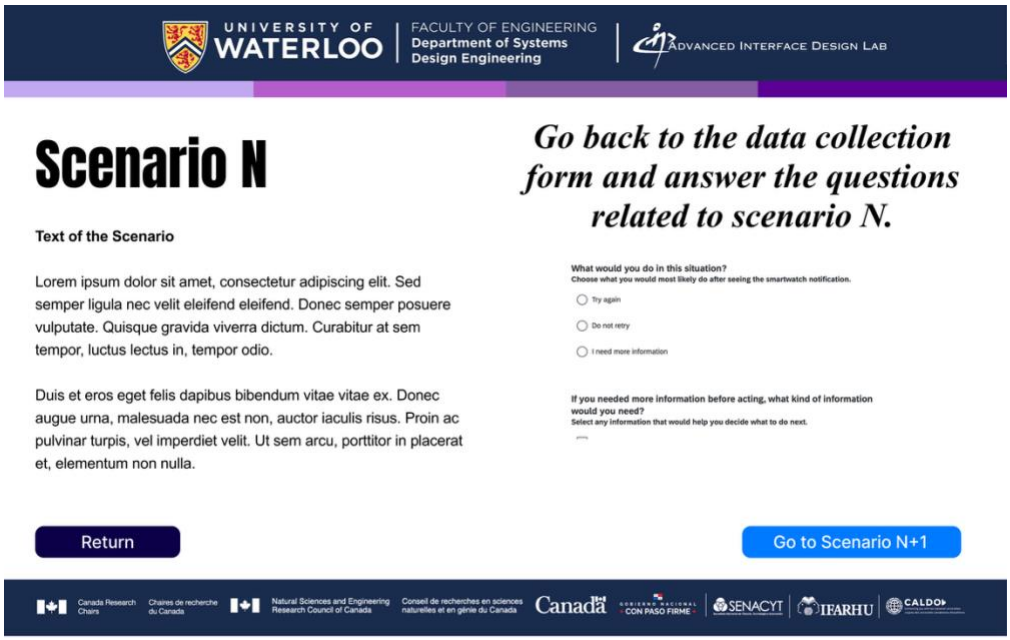


Figure 29. Participant Study Platform - Redirection to the data collection form

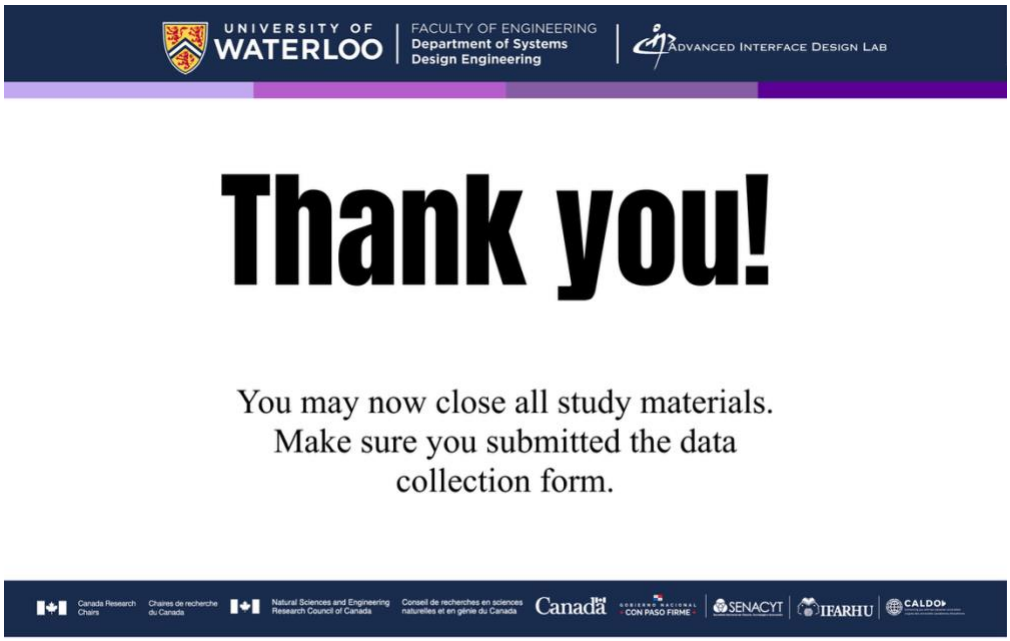


Figure 30. Participant Study Platform - End of the Study

Appendix B.3.3 Data Collection Form

PPG Quality Improvement Study Questionnaire

Study Questionnaire: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

Participant ID: _____

The following questions are repeated for each of the ten scenarios

Scenario 1: Before you interact with the prototype, please answer the Qualtrics questionnaire about what you understood from the situation. Please answer based only on what you saw in the video. Please imagine that you are the person in the video or that you are in the same situation.

What do you think may have caused the ECG reading to be unsuccessful? Select all that apply.
Answer based only on what you noticed in the video. There is no right or wrong answer.

- The person was moving (1)
- The person was walking (2)
- The person was running (13)
- The person was coughing (3)
- The person had just been talking (4)
- The person had recently exercised or was still physically active (5)
- The watch was loose or not positioned correctly (6)
- The person may not have been holding still (7)
- The environment or lighting made it harder to record (8)
- The smartwatch or sensor may not have worked properly (9)
- The person's symptoms or discomfort may have affected the attempt (10)
- I am not sure (11)
- Other (12) _____

If you were in this situation and had not yet seen any smartwatch guidance, what would you do next? Choose the action that feels closest to what you would naturally do in this situation.

- Try recording the ECG again immediately
- Stop moving and try again
- Sit down or rest first, then try again
- Adjust or tighten the watch, then try again
- Wait and try again later
- Ask someone nearby for help
- Contact a health care professional
- Seek emergency help
- Do not try again
- I am not sure

How motivated do you feel to record an ECG in this situation? Motivation means how much you would want to record an ECG in this situation.

1 = Not motivated at all (1)	2 = Slightly motivated (2)	3 = Somewhat motivated (3)	4 = Neutral / unsure (4)	5 = Moderately motivated (5)	6 = Very motivated (6)	7 = Extremely motivated (7)
○	○	○	○	○	○	○

How relevant or important did this situation feel to your health? Think about how personally important or health-related the situation would feel if it happened to you.

1 = Not relevant or important at all (1)	2 = Slightly relevant or important (2)	3 = Somewhat relevant or important (3)	4 = Neutral / unsure (4)	5 = Moderately relevant or important (5)	6 = Very relevant or important (6)	7 = Extremely relevant or important (7)
○	○	○	○	○	○	○

How easy or difficult would it be for you to record an ECG in this situation? Think about how easy or hard it would be to complete the ECG reading correctly in this situation.

1 = Very difficult (1)	2 = Difficult (2)	3 = Somewhat difficult (3)	4 = Neither easy nor difficult (4)	5 = Somewhat easy (5)	6 = Easy (6)	7 = Very easy (7)
○	○	○	○	○	○	○

How confident are you that you could record the ECG correctly right now? Think about whether you feel capable of recording the ECG correctly in this situation

1 = Not confident at all (1)	2 = Slightly confident (2)	3 = Somewhat confident (3)	4 = Neutral / unsure (4)	5 = Moderately confident (5)	6 = Very confident (6)	7 = Extremely confident (7)
○	○	○	○	○	○	○

Scenario 1: Now you will interact with the smartwatch interface shown to the person in this video. Please imagine that this notification appears on your own smartwatch in this situation. Review the interface and interact with it as you normally would. Please let me know once you have finished interacting with the prototype.

Scenario 1

What would you do in this situation? Choose what you would most likely do after seeing the smartwatch notification.

- Try again (1)
- Do not retry (2)
- I need more information (3)

Display this question:

If Scenario 1, What would you do in this situation? Choose what you would most likely do after seeing... = I need more information

If you needed more information before acting, what kind of information would you need? Select any information that would help you decide what to do next.

- Why the measurement was unsuccessful (1)
- How to fix the problem (2)
- Whether I should try again (3)
- •How long should I wait before trying again (4)
- Whether I should stop moving or sit still (5)
- Whether I should adjust the watch (6)
- Whether the situation is urgent (7)
- Whether I should contact a health care professional (8)
- Whether I should seek emergency help (9)
- I do not need more information (10)
- Other (11) _____

You have completed the questions for this scenario. Go back to the Figma Link.

End of Block: End of Scenario n

The following questions are shown at the end of scenario 5 and scenario 10

You have completed the questions for Scenario [5 or 10]. Please answer the following seven (7) questions based on your interaction with the prototype in the last 5 scenarios. Answer the following questions based on the last 5 scenarios

Overall, how much did the notifications in the previous 5 scenarios grab your attention? Think about whether the messages stood out to you and made you notice them.

1 = Did not grab my attention at all (1)	2 = Grabbed very little attention (2)	3 = Grabbed slight attention (3)	4 = Neutral / unsure (4)	5 = Grabbed some attention (5)	6 = Grabbed strong attention (6)	7 = Grabbed my attention very strongly (7)
○	○	○	○	○	○	○

Overall, how clear were the notifications about what you should do in the previous 5 scenarios?
Think about whether the messages made the next steps easy to understand.

1 = Not clear at all (1)	2 = Slightly clear (2)	3 = Somewhat clear (3)	4 = Neutral / unsure (4)	5 = Moderately clear (5)	6 = Very clear (6)	7 = Extremely clear (7)
○	○	○	○	○	○	○

Overall, how much did the notifications make you want to take action? Action could mean trying again, adjusting what you are doing, seeking more information, or taking the next step.

1 = Not at all (1)	2 = Very little (2)	3 = Slightly (3)	4 = Neutral / unsure (4)	5 = Somewhat (5)	6 = Very much (6)	7 = Definitely (7)
○	○	○	○	○	○	○

Overall, what would make these notifications more effective? Select anything that would make the notifications easier to understand, easier to follow, or more useful.

- Clearer instructions (1)
- Shorter wording (2)
- More explanation about why the measurement failed (3)
- More specific steps about what to do next (4)
- A clearer "Try again" option (5)
- A visual example or animation (6)
- Larger or easier-to-read text (7)
- More reassuring wording (8)
- More urgent wording (9)
- Less alarming wording (10)
- Information about when to contact a health care professional (11)
- Nothing; the message was effective as it was (12)
- Other (13) _____

After seeing the previous 5 scenarios, how motivated do you feel to record an ECG in similar situations? Answer based on how you would feel after seeing the smartwatch interface in the scenarios.

1 = Not motivated at all (1)	2 = Slightly motivated (2)	3 = Somewhat motivated (3)	4 = Neutral / unsure (4)	5 = Moderately motivated (5)	6 = Very motivated (6)	7 = Extremely motivated (7)
○	○	○	○	○	○	○

After seeing the previous 5 interfaces, how confident are you that you could record an ECG correctly in similar situations? Answer based on how capable you feel after seeing the smartwatch interface for these situations.

1 = Not confident at all (1)	2 = Slightly confident (2)	3 = Somewhat confident (3)	4 = Neutral / unsure (4)	5 = Moderately confident (5)	6 = Very confident (6)	7 = Extremely confident (7)
○	○	○	○	○	○	○

Overall, what would make these notifications more effective? Select anything that would make the notifications easier to understand, easier to follow, or more useful.

- Clearer instructions (1)
- Shorter wording (2)
- More explanation about why the measurement failed (3)
- More specific steps about what to do next (4)
- A clearer “Try again” option (5)
- A visual example or animation (6)
- Larger or easier-to-read text (7)
- More reassuring wording (8)
- More urgent wording (9)
- Less alarming wording (10)
- Information about when to contact a health care professional (11)
- Nothing; the message was effective as it was (12)
- Other (13) _____

After seeing the previous 5 scenarios, how motivated do you feel to record an ECG in similar situations? Answer based on how you would feel after seeing the smartwatch interface in the scenarios.

1 = Not motivated at all (1)	2 = Slightly motivated (2)	3 = Somewhat motivated (3)	4 = Neutral / unsure (4)	5 = Moderately motivated (5)	6 = Very motivated (6)	7 = Extremely motivated (7)
○	○	○	○	○	○	○

After seeing the previous 5 interfaces, how confident are you that you could record an ECG correctly in similar situations? Answer based on how capable you feel after seeing the smartwatch interface for these situations.

1 = Not confident at all (1)	2 = Slightly confident (2)	3 = Somewhat confident (3)	4 = Neutral / unsure (4)	5 = Moderately confident (5)	6 = Very confident (6)	7 = Extremely confident (7)
○	○	○	○	○	○	○

The following questions are displayed at the end of the ten scenarios

You have completed the questions for all scenarios. Please answer the following twelve (12) questions based on your interaction with the prototype across all 10 scenarios.

Start of Block: End of Study Survey

You have now completed all 10 scenarios. Please answer a few final questions about your overall experience with the smartwatch notifications and the scenarios. Please imagine you are in need of an ECG recording.

In general, what makes you decide to record an ECG after an unsuccessful measurement when prompted by your watch in scenarios similar to the ones shown today? Select the reasons that would make you more likely to record an ECG after an unsuccessful measurement.

- I feel symptoms, such as chest discomfort (1)
- I want to confirm my heart rhythm (2)
- I feel worried about my health (3)
- The message clearly explains what to do (4)
- The message explains why the measurement failed (5)
- I think I can fix the problem (6)
- I trust the smartwatch (7)
- I want information to share with a health care professional (8)
- Someone nearby encourages me to check (9)

What might make you decide not to record the ECG in scenarios similar to the ones shown today? Select the reasons that would make you less likely to record an ECG after an unsuccessful measurement.

- I do not think the situation is serious (1)
- I do not understand what went wrong (2)
- I do not know how to fix the problem (3)
- I think the reading will fail again (4)
- I do not feel confident using the ECG feature (5)
- I do not trust the smartwatch reading (6)
- I feel too stressed or worried (7)
- I am busy or distracted (8)
- I would rather wait and try later (9)
- I would rather contact a health care professional (10)
- Other (11) _____

If you would have the need to record an ECG. How confident are you overall in recording it on your smartwatch? Think about your overall confidence using the ECG feature, not just the scenarios from today.

1 = Not confident at all (1)	2 = Slightly confident (2)	3 = Somewhat confident (3)	4 = Neutral / unsure (4)	5 = Moderately confident (5)	6 = Very confident (6)	7 = Extremely confident (7)
○	○	○	○	○	○	○

In which kinds of situations would you find it most difficult to act on the notification's suggestion? Select the situations where it would be hardest for you to follow a smartwatch message.

- Walking (1)
- Coughing (2)
- Reduced lighting (3)
- Incorrect watch position (4)
- Talking (5)

Which type of notification did you find most motivating or helpful? Think about which set of messages helped you most when deciding what to do.

- The first set of messages (1)
- The second set of messages (2)
- Both were equally helpful (3)
- Neither was helpful (4)
- I am not sure (5)

What aspects of the new design helped you decide faster or more clearly? Select the parts of the message design that helped you understand what to do faster or more clearly.

- It explained what action to take (1)
- It explained why the measurement may have failed (2)
- It gave clear next steps (3)
- It made retrying feel easier (4)
- It helped me feel more confident (5)
- It was short and direct (6)
- It used easy-to-understand language (7)
- It got my attention (8)
- It felt reassuring (9)
- It felt appropriately urgent (10)
- It had a clearer button or action option (11)
- Nothing helped (12)
- I am not sure (13)

For the case scenarios, when you decided to retry the ECG measurement, what was the strongest reason? Choose the strongest reason that made you want to take action.

- I felt the situation was important to the health of the person in the situation (1)
- I wanted to get a successful ECG reading (2)
- I wanted reassurance (3)
- The message gave clear instructions (4)
- The action seemed easy to do (5)
- I felt confident I could do it correctly (6)
- I trusted the smartwatch (7)
- I was concerned about symptoms (8)
- I thought that I would have been able to perform the action prompted in the interface (9)

For the case scenarios, when you chose not to retry the ECG measurement, what was the main barrier? Choose the main reason that made you less likely to take action

- I did not think it was important enough (1)
- I did not understand the message (2)
- I did not know what to do next (3)
- The action seemed difficult (4)
- I did not feel confident (5)
- I thought the measurement would fail again (6)
- I did not trust the smartwatch (7)
- I felt stressed, worried, or overwhelmed (8)
- I preferred to wait (9)
- I preferred to contact a health care professional (10)

Overall, how realistic did the videos of the scenarios feel? Think about whether the situations shown in the videos felt believable and similar to something that could happen in real life.

1 = Not realistic at all (1)	2 = Slightly realistic (2)	3 = Somewhat realistic (3)	4 = Neutral / unsure (4)	5 = Moderately realistic (5)	6 = Very realistic (6)	7 = Extremely realistic (7)
○	○	○	○	○	○	○

Is there anything you would change about how the smartwatch delivers health notifications? Select any changes that would make smartwatch health notifications clearer or more useful.

- Use simpler language (1)
- Give clearer next steps (2)
- Explain why the measurement failed (3)
- Explain when to try again (4)
- Explain when to contact a health care professional (5)
- Make the message shorter (6)
- Make the message more noticeable (7)
- Make the message less alarming (8)
- Make the message more urgent (9)
- Add visuals or animations (10)
- Add a clearer button or action option (11)
- Personalize the message to the situation (12)
- I would not change anything (13)

What do you think about AI-generated videos to simulate scenarios in this type of testing?

These questions are about the videos themselves, not the smartwatch messages.

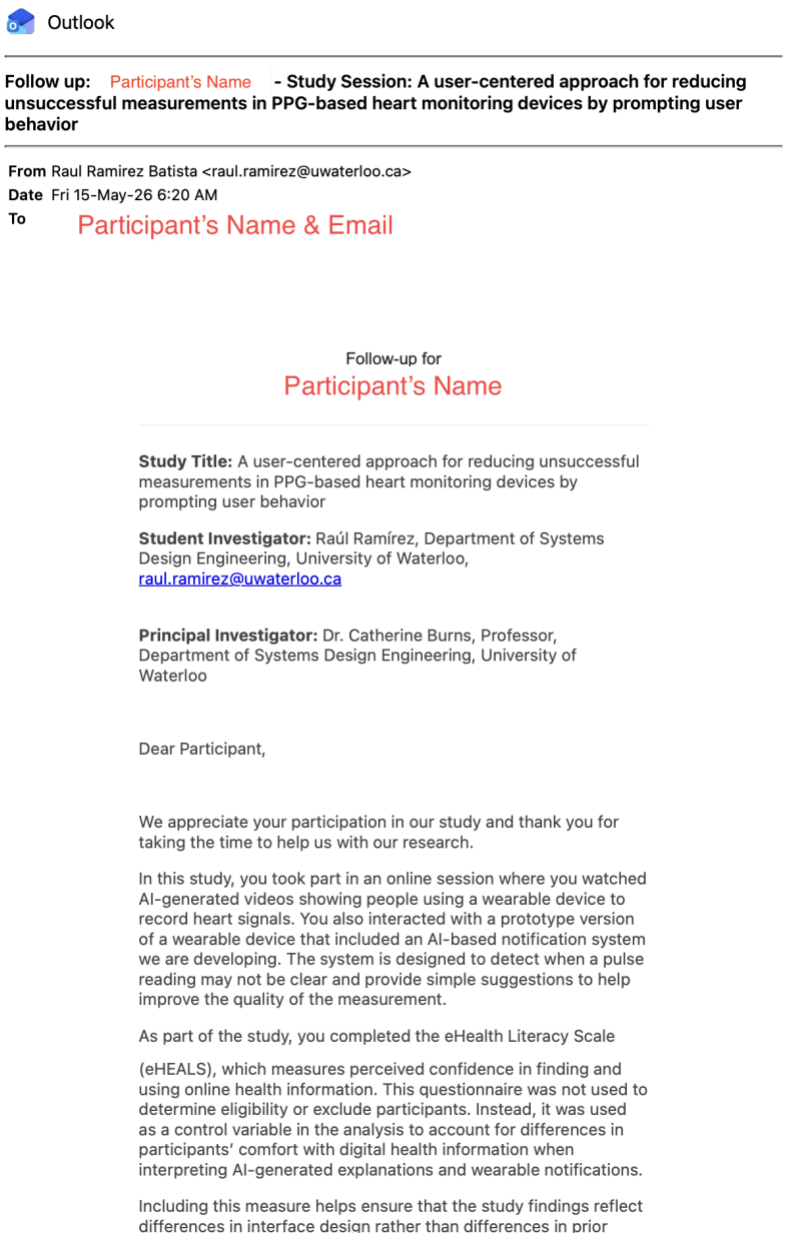
	None at all (1)	A little (2)	A moderate amount (3)	A lot (4)	A great deal (5)
How helpful were the AI-generated videos for understanding the situations? (1)	○	○	○	○	○
How realistic did the AI-generated videos feel? (2)	○	○	○	○	○
How easy was it to imagine yourself in the situations shown in the videos? (3)	○	○	○	○	○
How clear was it what was happening in the videos? (4)	○	○	○	○	○

This is the end of the survey. All responses will be recorded, and you will be redirected to the Information-Consent Form. Please indicate to Raul that you have completed the questionnaire.

Appendix B.4 Post-Study Session Materials

Fifteen minutes after the study session ends, participants receive the feedback and appreciation letter presented in Appendix B.4.1, which takes them to the Participant Contact Details Form for Remuneration (Appendix B.4.2).

Appendix B.4.1 Feedback and Appreciation Letter



Outlook

Follow up: Participant's Name **- Study Session: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior**

From Raul Ramirez Batista <raul.ramirez@uwaterloo.ca>
Date Fri 15-May-26 6:20 AM
To Participant's Name & Email

Follow-up for
Participant's Name

Study Title: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

Student Investigator: Raúl Ramírez, Department of Systems Design Engineering, University of Waterloo, raul.ramirez@uwaterloo.ca

Principal Investigator: Dr. Catherine Burns, Professor, Department of Systems Design Engineering, University of Waterloo

Dear Participant,

We appreciate your participation in our study and thank you for taking the time to help us with our research.

In this study, you took part in an online session where you watched AI-generated videos showing people using a wearable device to record heart signals. You also interacted with a prototype version of a wearable device that included an AI-based notification system we are developing. The system is designed to detect when a pulse reading may not be clear and provide simple suggestions to help improve the quality of the measurement.

As part of the study, you completed the eHealth Literacy Scale

(eHEALS), which measures perceived confidence in finding and using online health information. This questionnaire was not used to determine eligibility or exclude participants. Instead, it was used as a control variable in the analysis to account for differences in participants' comfort with digital health information when interpreting AI-generated explanations and wearable notifications.

Including this measure helps ensure that the study findings reflect differences in interface design rather than differences in prior

familiarity with digital health information.

You completed ten simulated cases that addressed the frequent limitations of wearable technologies that use PPG signals. Some cases included prompts that used counterfactuals, an explainable AI (XAI) technique to nudge actions that could have improved the quality of the PPG measures, while others did not include any details on why the measurement was unsuccessful. You were asked to think aloud as you worked through each case, and the researcher prompted you to understand your thought process better; both your screen and audio were recorded.

The purpose of the study is to inform the design of heart-monitoring wearable interfaces by evaluating the effectiveness of incorporating counterfactual explanations into an AI system designed to determine if the quality of PPG signals is good.

We hypothesize that prompts based on counterfactuals will influence users to change their behavior. Specifically, we believe participants will understand the prompts and intend to follow the suggested action before retaking the ECG, thereby leading to fewer unsuccessful measurements. Additionally, we expect these prompts to enhance users' understanding of the limitations of the wearable PPG analysis algorithm and promote more consistent reasoning during think-aloud sessions. Conversely, we anticipate that users will proceed to the next cases when the measurement is unsuccessful and no guidance is provided to improve the PPG measurement quality.

Please note that while participation in this study was confidential, it was not anonymous. We collected information that could potentially be used to identify participants, including IP addresses, screen, audio, and video recordings, and responses to background questions. Although no names were linked to the data, some indirect quotes from the think-aloud sessions may be used in research reports or publications to help explain the findings. These quotes will be presented in a way that does not reveal your identity. All data have been handled with strict confidentiality: identifying information is separated from study responses and will be coded before analysis. Only authorized research team members have access to the raw data, which will be securely stored for at least 7 years. Published results will be presented in aggregate form, and no individual participant will be identifiable.

This study has been reviewed and received ethics clearance

through a University of Waterloo Research Ethics Board (REB 47991). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

For all other questions, contact Raúl Ramírez at raul.ramirez@uwaterloo.ca

We really appreciate your participation and hope that this has been an interesting experience for you.

As a token of appreciation, you are eligible to receive a $20 CAD remuneration via e-transfer. If you would like to receive this, please complete the following short form:

https://uwaterloo.ca1.qualtrics.com/jfe/form/SV_cwnk2MMdKfhjfw2

Best regards,



Figure 31. Feedback and Appreciation Email

Appendix B.4.2 Participant Contact Details Form for Remuneration

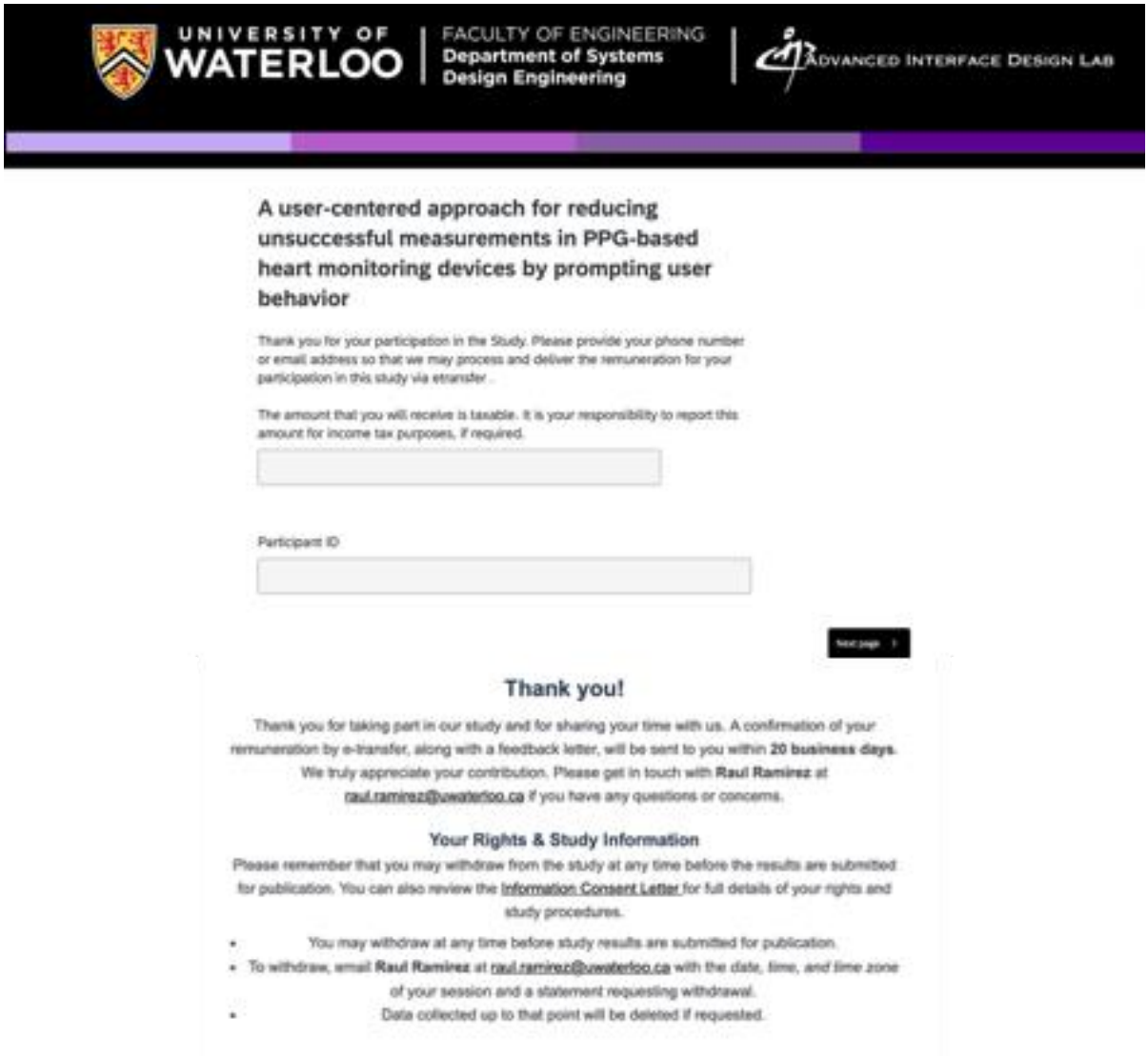


Figure 32. Participant Contact Details Form for Remuneration

Appendix B.5 Ethics Clearance Certificate

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Catherine Burns (Systems Design Engineering)

Student investigator: Raul Ramirez Batista (Systems Design Engineering)

Student investigator: Naomi Loughlin (Systems Design Engineering)

File #: 47991

Title: A user-centered approach for reducing unsuccessful measurements in PPG-based heart monitoring devices by prompting user behavior

---

The Human Research Ethics Board is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 03/06/26 (m/d/y)

University of Waterloo Research Ethics Boards are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS2 2022), the Ontario Personal Health Information Protection Act (PHIPA), and all laws and regulations of the province of Ontario (as applicable). Additionally, CREB operates in a manner consistent with the International Conference for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) Guidance E6(R2): Good Clinical Practice , the International Organization for Standardization of Good Clinical Practices (GCP) as set out by ISO 14155 - Clinical investigation of medical devices for human subjects, Part C, Division 5 of the Food and Drug Regulations, Part 4 of the Natural Health Products Regulations, Part 3 of the Medical Devices Regulations. Both Boards are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREB) and IRB00007409 (CREB).

Expiry Date: 03/07/27 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Board



Heather Dekker, Ethics Advisor, hdekker@uwaterloo.ca, 519-888-4567, ext. 41506

This above-named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: Feedback Letter_Version 3_20260223.pdf

file: PPG Study Recruitment Poster Version 2.0_2026_02_11.pdf

file: PPG Study Recruitment Poster (Instagram Post) Version 2.0_2026_02_11.pdf

file: Recruitment email_version2_20260211.pdf

file: Information Consent Letter and Consent Form _ Version 3.0_20260305.pdf

file: PPG Study_email collection form_version1_2025-11-19.pdf

file: Scenarios and Participants Questionnaire _version1.0_20260224.pdf

Approved Protocol Version 3 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.

Figure 33. Ethics Clearance Certificate